RADBOUD UNIVERSITY NIJMEGEN

MASTER THESIS

Triple Scoring Utilizing Annotated Web Data

Author:
Frank DORSSERS

Supervisor:
Arjen P. DE VRIES

Student number:
s0824704

Second reader:
Eelco HERDER

August 30, 2017

Abstract

Ranking entities is an active field of research, with ties to different applications. We research the use of web data for triple scoring, specifically ranking professions or nationalities in order of relevance for given persons, inspired by the triple scoring task for WSDM Cup 2017. This web data is enhanced with spam rankings indicating the spamminess of documents and annotations indicating where entities occur in documents, both annotations available as a dataset online focusing on precision, as well as those generated for this thesis, focusing on recall. This data is used in two primary ways, using document entity co-occurrence counts to judge relevancy and using machine learning on contextual information around entities extracted from the web data using the annotations, both approaches with several variations. We found there to be no large consistent differences in terms of accuracy between the two types of annotations as well as that machine learning generally outperforms the simple co-occurrence counts.

Contents

1 Introduction

At the tenth ACM international conference on Web Search and Data Mining a challenge was posted for triple scoring, where the goal was to determine, for a selection of professions and nationalities for different persons, a ranking indicating how relevant each of these is for that person.

This triple scoring challenge describes a problem called entity ranking, which often has connections to other fields like entity typing. For example determining the type of a word in text, whether it is a person, a location, or a time. In many of those cases giving a single result is enough, for example plainly stating that it is a person.

For this challenge however, it focusses on ranking candidate selections. Given a person you get a collection of professions or nationalities and the goal is to put these in the right order, from most relevant to least relevant. To achieve this they provided different sets of data, including annotated sentences from Wikipedia.

This thesis will investigate the use of a different source of data for this challenge, specifically the use of raw web data. ClueWeb12 is the result of a large webcrawl, containing over 700 million webpages taking up nearly 30 terabytes. However, much like the provided annotated Wiki sentences, we need to know where the entities of the challenge actually appear in these webpages, so an additional dataset called FACC1 is used, which stands for Freebase Annotations of the ClueWeb Corpora v1, and adds annotations showing which entities occur in which documents, allowing one to easily pick a document or context for training. Unfortunately, there does not appear to be a public description of the method used to generate these annotations, which is not ideal as the primary research goal in this thesis focuses on web data, and a single specific webscrape.

The secondary research question pertains to the use of annotations in webdata: How well do the results based on annotations from a naive annotation generator compare to the results based on FACC1 annotations, namely how much does a high quality and tuned entity tagger add to the process of relationship detection, as it is only a small part of the pipeline.

Last but not least a third data source is introduced: Spam rankings. These spam rankings indicate how spammy a document is, where the assumption is that less spammy documents should provide better webpages and contexts.

We start by introducing some related work on entity ranking and other relevant areas of research like entity typing and multilabel classification. The following section takes a closer look at the actual WSDM challenge, for example how the scoring system works and what metrics they utilize. The fourth section focuses on the data, both the provided data, for example all possible entities, different knowledge bases, and the wiki sentences, but also on the previously described data from different sources, like ClueWeb12, FACC1, and also how the new annotations are generated.

To get a better understanding of what these additional datasets look like the following section provides an analysis of this data, for example what are the distributions of entities without entity groups like professions or nationalities, how often do professions occur, how often do persons occur, and how are entities distributed over different spam rankings.

Different approaches which are employed for actual scoring are described in the next section, describing variations of approaches which solely use the annotations, one which uses annotations combined with

some Wikipedia information and the most important approach, which uses actual ClueWeb12 textual information.

Results from these different approaches are described in the second to last section, showing how some of these approaches perform on the train as well as the test data provided for the challenge, including two mock submissions which can be compared to other submissions to the challenge.

2 Related work

A varied number of research areas are of relevance for entity ranking, several of which are described in this section.

Entity ranking Entity ranking is, given an entity or a small piece of text, ranking these these in a specified way, depending on the task at hand. Some focus on using structured data like Freebase and DBPedia, and the relationships that are present within these databases, to rank entities [21]. Others use semi-structured data like Wikipedia for entity ranking, basing the rank on the number of incoming connections on potential types, category similarity and full text relevance estimates [22]. More textual features on unstructured data like ClueWeb12 also seems to result in promising results [20].

The last approach is also one used in this thesis, using webdata and textual features to perform entity ranking.

Entity typing Much like entity ranking, instead of focusing on labelling or classifying queries, large pieces of text, or something similar, entity typing focuses solely on entities, which may or may not be in a context. Much like entity ranking the type of data used varies, however the number of steps also varies. Entity typing can be used in contexts where it is completely clear what the entity is, but it can also be used as the last step of named entity recognition. A large issue either way is getting training data, which is why distant supervision [7] is occasionally used [14, 25]. This has been applied on Wikipedia entries used for named entity recognition and typing using mostly textual features like tokens, part-of-speech-tags and ReVerb patterns [14]. A very recent approach uses ClueWeb12 and FACC1 to perform entity typing based on word embeddings around the entities present in the ClueWeb12 training data [25].

The approach by Yaghoobzadeh and Schütze in their paper [25] is slightly similar to the research being performed in this thesis in terms of data and the usage of contexts, however our focus will be more on character features and additional external, and automatically generated data.

Named-entity recognition using classification methods Named-entity recognition is the task of identifying and classifying words in text into different types, for example locations, persons, expressions of time, et cetera. Often there were only a small number of categories, however more recent research has looked at this issue with an increasing number of tags, up to 112, using textual features among others and the perceptron algorithm [14]. This number brings it closer to the problem described in this paper.

A variant of named-entity recognition allows for multiple types to be chosen for entities, which is often described as multilabel classification and typing. Some approaches view the multilabel problem as a

collection of multiple binary classifications which are independent of each other. Other approaches on the other assume that there are dependencies between labels and exploit these [13, 27].

The approaches so far are limited to a few hundred types at most, however there is an area of research which focusses on situations where the number of types is significantly more, aptly referred to as extreme classification. Extreme classification is a workshop on multi-class and multi-label learning in extremely large label spaces given at NIPS, most recently in 2016.[1] The number of labels for datasets in the extreme classification repository ranges from 101 all the way through 8,838,461 [2]. Extreme cases like these can be tackled in different ways, some of which use partitioning, either on the labels [24] or on the features using decision trees [19, 1]. The reason partitioning is chosen is that many other approaches solve these problems on a label by label basis, which in turn makes the approaches scale linearly in time, making them unsuitable for extreme classification. The reason some approaches use feature partitioning is that generally only a limited number of labels is relevant for regions in the feature space. For the challenge described in this thesis, persons will be assigned ranked labels, which are taken from a list of 100 nationalities and 200 professions. Due the number of possible labels being on the lower end, more traditional approaches can be used.

3 WSDM Triple Scoring challenge

The WSDM Triple Scoring challenge is part of WSDM Cup 2017, organized for the tenth ACM International Conference on Web Search and Data Mining. The challenge consists of two tracks: vandalism detection in Wikidata and triple scoring, which are described in detail by Heindorf et al. [12].

The goal of the first task, vandalism detection, focuses on computing a score denoting the likelihood that revisions of information on Wikidata were vandalism or otherwise damaging edits. The second task, triple scoring, is the task being tackled in this thesis and will be described in more detail in this section.

Given a triple from a type-like relationship the goal is to determine a score which measures the relevance expressed by the triple, compared to other triples [4]. These triples have the following elements: Two entities and the type relation. For this challenge, the space of possible triples is severely restricted as there are only two types of relations that are being considered, both with a finite number of entities: Persons having a profession and persons having a nationality. Several examples for the profession triples are seen below.

```
Barack Obama has-profession Politician
Barack Obama has-profession Author
Billy Joel has-profession Pianist
Billy Joel has-profession Businessperson
```

Several other examples for the nationality triples can be seen below.

```
Arnold Schwarzenegger has-nationality Austria
Arnold Schwarzenegger has-nationality United States of America
```

[1] See http://manikvarma.org/events/XC16/schedule.html, last accessed June 28th, 2017.
[2] See http://manikvarma.org/downloads/XC/XMLRepository.html, last accessed June 28th, 2017.

```
Albert Einstein has-nationality Germany
Albert Einstein has-nationality Switzerland
```

The goal is to give these triples a relevance score indicating how relevant that specific combination is. For example `Billy Joel has-profession Pianist` would have the highest score out of Billy Joel's triples, as it is primarily relevant, while `Billy Joel has-profession Businessperson` would have the lowest score, as this is secondarily relevant. These scores are in the interval $[0, 7]$, 0 indicating the lowest possible relevance and 7 indicating the highest possible relevance. An important thing to note here is that completely irrelevant combinations are not present in the data. All combinations that appear are at least somewhat relevant, hence why the score indicates a triple being primarily or secondarily relevant.

Scores were generated by providing judges with a person and either its relevant professions or nationalities. They were then ordered to move each of these professions to either 'primarily relevant' (1) or 'secondarily relevant' (0). In total the professions or nationalities for each person were judged by seven judges, these were then summed to form the final scores, hence why the scoring for this challenge is in the previously mentioned interval of $[0, 7]$.

Three different metrics are used for scoring submissions, two of which are score based and one which is rank based:

Accuracy The percentage of triples for which the predicted relevance score is within 2 points of the actual relevance score (e.g. scores 3 through 7 are correct if the actual score is 5),

Average Score Difference The sum of the absolute differences between the predicted and true scores, divided by the total number of triples,

Kendall's Tau Groups the predicted and true scores per person in a task, computes the Kendall Tau as defined in the paper by Fagin et al. per group and divides the resulting sum of these groups by the number of unique persons [10].

The challenge organizers describe several other metrics in their paper [4], however the metrics above were used for the actual challenge. The implementations of these metrics as they were used are also available online.[3]

The organization of the challenge provided several different datasets:

- Three lists containing all possible persons, professions and nationalities,
- two knowledge bases containing all possible person-profession and person-nationality combinations,
- two train sets with relevance scores, one for person-profession and the other for person-nationality combinations,
- a large collection of wiki sentences where all persons occur.

These datasets will be described in more detail in the following section.

[3] See http://broccoli.cs.uni-freiburg.de/wsdm-cup-2017/evaluator.py, last accessed May 22nd, 2017.

4 Data

The data used for this project can be divided into two general categories: Data provided by the WSDM Cup 2017 organization and additional data taken from other sources. Both of these categories are described in the following sections.

4.1 Provided data

As part of the challenge several datasets were provided, which are available at the WSDM Cup 2017 website.[4] These datasets can be divided into several types, providing different parts or different types of data.

4.1.1 Entity list

Three different entity types are considered: Persons, professions and nationalities. The number of possibilities in each of these categories is finite and is given in the corresponding tab-separated values files.

The persons file contains a total of 385,426 unique persons, associated with their corresponding Freebase ID. Professions and nationalities both have respectively 200 and 100 entities, however these do not contain the corresponding Freebase IDs.

Two attributes were missing for the approach described in this paper, and the following adjustments were made. Firstly, the provided nationalities (which currently listed countries), have been extended with their actual nationalities (for example, adding 'Dutch' to 'Netherlands' and 'American' to 'United States of America'), using an online dictionary.[5]

Secondly, Freebase IDs were added to nationalities and professions where possible, as also described in previous work [9], but explained in more detail here. The first step is checking DBpedia using SPARQL for possible known Freebase IDs, these are stored under the relation `owl:sameAs` on resource pages. This relationship often occurs multiple times, however the one relevant here contains `freebase.com` in the value. Listing 1 shows the specific query, which works for basically all countries.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
SELECT ?label
WHERE <http://dbpedia.org/resource/{Entity}> owl:sameAs ?label
```

Listing 1: Querying resource page for a given entity

In the case of some professions the first resource page does not contain the expected information, but redirects to a different page which hopefully does using `dbo:wikiPageRedirects`. This problem can also be solved with SPARQL as seen in Listing 2, which automatically follows any wiki page redirect that is present and tries to find any `owl:sameAs` relationships on that page.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX dbo: <http://dbpedia.org/ontology/>
```

[4] See http://www.wsdm-cup-2017.org/triple-scoring.html, last accessed May 22nd, 2017.
[5] See http://www.esldesk.com/vocabulary/countries, last accessed January 24th, 2017.

```
SELECT ?label WHERE <http://dbpedia.org/resource/{Entity}> dbo:wikiPageRedirects ?
    redirect . ?redirect owl:sameAs ?label
```

Listing 2: Querying a redirected resource page for a given entity

The next step in the lookup is manually verifying the automatically queried Freebase IDs, which is another two-step process. Wikidata is used for the first verification step, by doing a reverse lookup using the Freebase ID to get the entity name from Wikidata which is then compared to the expected entity. This is done using the SPARQL query seen in Listing 3.

```
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
SELECT ?item ?itemLabel WHERE {{
?item wdt:P646 "{FreebaseID}" .
SERVICE wikibase:label {{
bd:serviceParam wikibase:language "en" .
}}
}}
```

Listing 3: Querying Wikidata with Freebase ID for the entity name

The final step utilizes the Freebase Annotations for the ClueWeb Corpora v1, these are annotations of a dataset of crawled webpages, indicating which entities appear in what documents and where exactly, this data is explained in more detail further ahead.

FACC1 is used in several ways for verification of the Freebase IDs that have been found so far. First of all, annotations are retrieved based on the Freebase IDs and these are used to verify whether the entities they represent match the nationality or profession that is expected. If no annotations can be retrieved based on the Freebase ID then this ID is removed. Secondly, annotations are retrieved based on the entity name, for example "carpenter", and the corresponding Freebase IDs are then similarly compared to what has already been found in the previous steps. This is also used to add any Freebase IDs that have not yet been discovered in the previous steps.

4.1.2 Knowledge base

A total of two knowledge bases were provided as challenge data, one containing all possible person-profession combinations and the other one containing all possible person-nationality combinations. Any possible triple provided during the challenge is taken from these knowledge bases, so the number of possible combinations is also finite, and it allows for precomputation for quick answering during the challenge. The data is stored in a tab-separated format, with the first column containing the name of the person and the second column containing the target entity, either the profession or the nationality depending on the knowledge base.

Professions The profession knowledge base contains a total of 499,244 unique person-profession combinations, all of which are judged to be at least somewhat relevant. This data contains all 200 possible professions from the previously described profession entity list, however it only contains $343,329$ out of $385,426$ unique persons. Table 1a shows the most and least common professions that appear in the knowledge base. The top is not really surprising, as these are common topics of interest on the web. Other professions which could have appeared here would be ones that have to do with sports.

After counting the number of professions per person and adding the missing persons with a value of 0, a person has on average 1.295 relevant professions, with a standard deviation of 0.985.

Figure 1: Number of relevant professions that persons have in the knowledge base

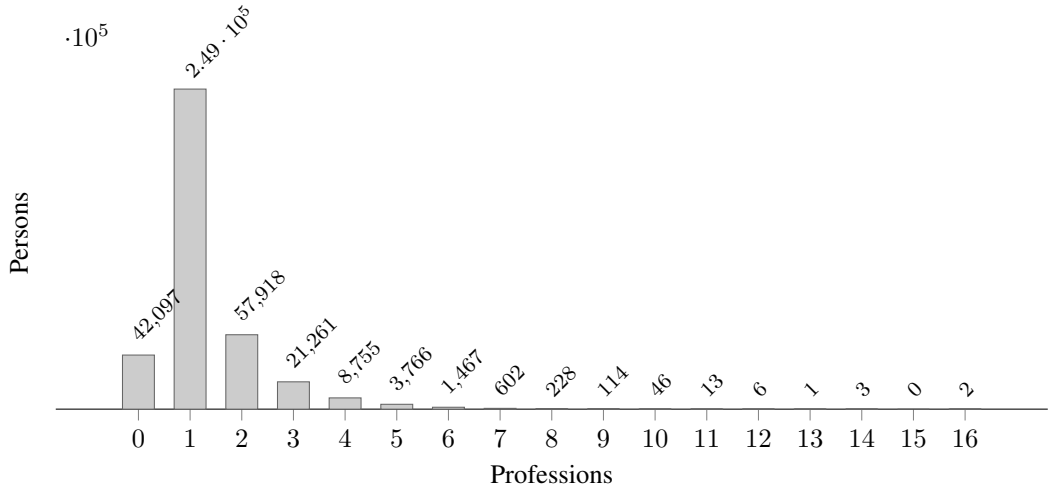
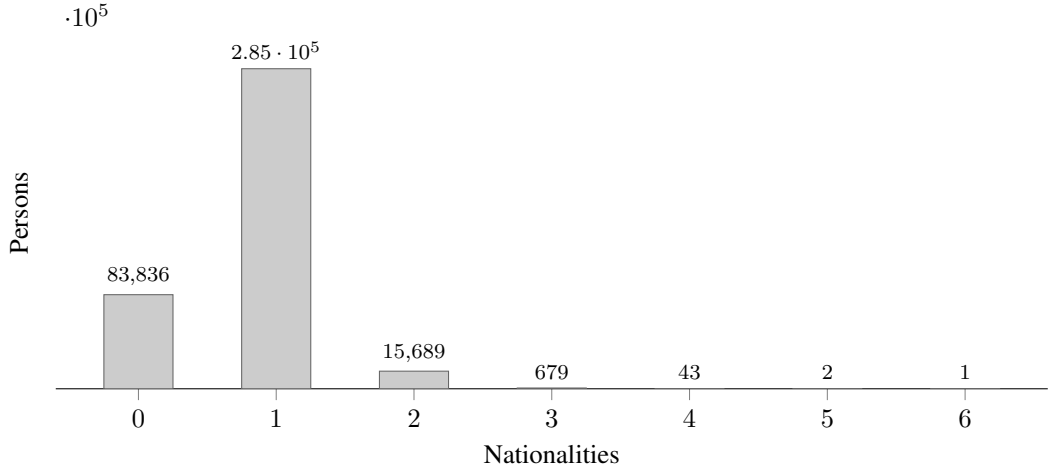


Figure 1 shows the distribution of the number of people against the number of professions with log scaling on the y axis. This shows there is a sharp decline for the number of people as the number of professions increases. The two busiest people in this knowledge base are Paul McCartney and Sakis Rouvas, both with 16 professions. There is an overlap of six out of the sixteen professions between the two, these are: businessperson, entrepreneur, film producer, multi-instrumentalist, record producer and songwriter.

Nationality The nationality knowledge base contains about $180,000$ fewer person-nationality combinations than the person-profession data, for a total of $318,779$ rows. The data still contains all 100 possible nationalities, however neither does this set contain all unique persons, as it only contains information about $301,590$ people. An interesting thing to note here is that the intersection of all persons in the profession as well as the nationality knowledge base results in $376,214$ persons. So there are 9212 persons that will never appear during the triple ranking challenge, as the knowledge base contains all possible combinations that the organization is able to ask. Fortunately, the methods applied in this paper assume that persons are completely independent of each other, so persons that are present in the person list, but not in the knowledge base, have no effect on the results. Table 1b shows the most and least common nationalities in the knowledge base. Unsurprisingly several English-speaking countries are at the top, as ClueWeb12 also mostly contains English webpages. The average number of nationalities per person is, as expected given the missing persons, lower than the average number of professions, with a mean of 0.827 and standard deviation of 0.486.

Figure 2 shows the distribution of the number of nationalities per person. As opposed to the profession knowledge base the variation is a lot more limited here, with only two persons, Marc Rich and Christian Rakovsky, having five nationalities, and a single person who has six. This was Anny Ondra, a Czech film

Figure 2: Number of relevant nationalities that persons have in the knowledge base



actress for whom, according to the knowledge base, the following six nationalities are at least somewhat relevant: Austria, Czechoslovakia, Czech Republic, France, Germany and Poland.

Table 1: The most and least common professions and nationalities in the knowledge bases

(a)

Profession	Count
Actor	61450
Politician	43234
Singer	20640
Screenwriter	18747
Writer	17383
..	..
Alchemist	4
Cantor	3
Rodeo clown	3
Tentmaker	1
Sound Sculptor	1

(b)

Nationality	Count
United States of America	122759
United Kingdom	25229
Canada	15345
England	12370
Germany	11750
...	...
Cameroon	163
Ecuador	161
Paraguay	158
Ivory Coast	126
Bahrain	64

4.1.3 Scored data

A small amount of training data was provided, a set consisting of 515 ranked person-profession and 162 ranked person-nationality combinations. Below is a small snippet from the profession training data showing the ranked results for Barack Obama, indicating that being a politician has the highest relevance, while being a law professor is significantly lower.

```
Barack Obama Politician 7
Barack Obama Lawyer 0
```

```
Barack Obama Law professor 1
Barack Obama Author 0
```

The nationality data is formatted in the same way, tab separated values between the person, target entity and rank. Below is the data for Albert Einstein, showing that his primary nationality is German, where he was born 1879, while Sweden and the USA are also relevant, as he gained citizenship in both respectively in 1901 and 1940.

```
Albert Einstein Germany 7
Albert Einstein Switzerland 4
Albert Einstein United States of America 4
```

The train sets contain information for a total of 209 unique persons, 134 in the profession set and 77 in the nationality set. Paul McCartney is, again, the most common person in the training set, as the profession data contains all his sixteen relevant professions. A total of 137 out of 200 professions is present in the profession data, of which actor is the most common, occurring 48 times. This is followed by film producer (22), record producer (20) and singer-songwriter (19). All 0 through 7 scores are also present, though 6 and 7 are the most common.

For the nationality data 36 out of the 100 different nationalities are present, with the USA being the most frequent with 40 out of 162 rankings being about America, followed by the UK (27), France (13) and Canada (10). The scores are leaning a lot more to the higher values, with 7 representing 67 out of 162 data points, and both 0 and 2 only occurring 7 times.

4.1.4 Wiki sentences

The last dataset provided by the organization of the triple scoring challenge is the wiki sentences dataset which consists of a total of 33,159,353 sentences, with each person appearing in at least three sentences and the most common person in 68,662 sentences. On average, each person has approximately 86 sentences.

The wiki sentence below (number 37482), shows what these sentences look like. It shows a few interesting things, for example that a sentence is not limited to a single person. A single person can occur multiple times, but it can also occur together with different persons. The snippet also shows that co-reference resolution was used for this task, as 'his' is also tagged as Barack Obama.

Occurrences of persons are substituted for an alternative representation, which contains a generalized form of the name of the person it refers to, together with the text snippet. This makes it easy to find occurrences of people one is looking for and retrieving these sentences.

```
[Barack_Obama|President Obama] asked [Barack_Obama|his] Yemen
counterpart [Ali_Abdullah_Saleh|Ali Abdullah Saleh] to ensure
closer cooperation with the US in the struggle against the
growing activity of al-Qaeda in Yemen , and promised to send
additional aid .
```

One thing to note is that while this information is provided by the organization and described in this section, it is not actually used for the approaches in this thesis. The data that is used though is similar to the wiki sentences, albeit from a different source.

4.2 Additional data

The challenge described in this thesis did not restrict participants to the provided data, instead participants were allowed to use any data however they saw fit, barring of course manually labelled answers to knowledge base triples.

Most, if not all of the additional data used can be seen as webdata, data pertaining to crawled websites and related information. All these sets are described in the sections below.

4.2.1 ClueWeb12

ClueWeb is part of the Lemur Project, an initiative by the University of Massachusetts, Amherst, and the Carnegie Mellon University, and is named after the U.S. National Science Foundation's Cluster Exploratory (CluE) program.[6] ClueWeb is a text based webcrawl of mostly English webpages, collected in the span of a few months. Currently there exist two ClueWeb versions, ClueWeb09, which was crawled in January and February of 2009 and Clueweb12, which was crawled between February 10 and May 10 2012. This thesis employs the most recent version for its research, ClueWeb12. While this dataset is already a few years old, and much may have changed over the years, we assume that this data is still relevant to the questions asked in the challenge. This is also one of the more recent datasets with Freebase annotations for the entire scrape, indicating in what document and what offset certain entities appear. This annotation data is described in more detail in the next section.

The initial seed list for the scrape of ClueWeb12 consisted of almost 3 million unique URLs, taken from different sources, e.g. ClueWeb09 pages with the highest PageRank [18] score that were not categorized as spam, the 262 most popular English sites as ranked by Alexa and a few thousand travel sites. There were also several crawlers that were focusing on specific regions of the web, for example URLs that appeared in tweets, and pages from Wikitravel. The crawled data was postprocessed in a general cleanup phase, which performed several steps, e.g. remove robots.txt files, webpages that are too large, non-English pages and blacklisted pages. The resulting dataset contains a total of 733,019,372 documents over 33,447 files taking up 27.3 TB of disk space when uncompressed.

4.2.2 FACC1

ClueWeb12 contains terabytes of data, providing valuable information. This amount of data does make it difficult to find exactly the right piece of information one is looking for. The Freebase Annotations of the ClueWeb Corpora, v1 dataset [11] aims to solve this problem and is available on the Lemur Project webpage.[7] These annotations were generated at Google, however they only provide a very brief description of the process on their blog.[8] They note that the annotation process was fully automated, so the chances are high that there are mistakes. However, they optimized for precision as opposed to recall, making the process skip sentences or even entire documents if the tagger was not confident enough to assign Freebase IDs to entities, or could not find any relevant Freebase entities. On a small sample of the dataset they report a precision of 80-85%, and a recall of 70-85%.

[6] See http://lemurproject.org/clueweb12/FAQ.php, last accessed July 13rd, 2017.
[7] See http://lemurproject.org/clueweb12/FACC1/, last accessed May 3rd, 2017.
[8] See https://research.googleblog.com/2013/07/11-billion-clues-in-800-million.html, last accessed May 3rd, 2017.

Below is a small sample of the FACC1 dataset:

```
clueweb12-0500tw-00-09073 UTF-8 Make It Big 46770 46781 0.967847
0.000640 /m/05j8sq
clueweb12-0500tw-00-09073 UTF-8 Irene Cara 49452 49462 1.000000
0.000009 /m/01r7pq
clueweb12-0500tw-00-09073 UTF-8 Billy Ocean 49464 49475 0.992633
0.000007 /m/0178d6
```

These annotations contain several fields:

- ClueWeb12 document id
- Encoding of the document
- Entity string which was tagged
- Byte offset where the entity string starts in the page
- Byte offset where the entity string ends in the page
- Probability that this entity is tagged correctly based on the entire sentence
- Probability that this entity is tagged correctly based on the entire sentence minus the actual entity string
- Predicted Freebase ID of the tagged entity

Over 647 million documents were analyzed to generate the FACC1 dataset, 456 million of which have at least one annotation. An annotated document has 13 annotations on average, resulting in approximately six billion annotations for the entire ClueWeb12 dataset.

These annotations provide a connection between the provided data by WSDM and ClueWeb12, opening up a multitude of interesting options. Some possibilities are extracting context about entities mentions and using this for predictions, finding co-occurrences between entities, weighing combinations based on the distance between them, and so forth. Some of these approaches are described in this thesis.

4.2.3 Spam rankings

An additional dataset for ClueWeb12 has been created by Cormack et al.. While their paper describes the method used to generate spam rankings for ClueWeb09 [6], they applied the same methodology to ClueWeb12 and made this dataset available online.[9]

Their paper describes three separate models for spam ranking, based on three different training sets. They also provide an additional model by combining the initial three models, effectively creating a fourth model, predicting fusion scores. Scores produced by all four methods are available for ClueWeb09, however only the fusion scores are available for ClueWeb12.

[9] See http://www.mansci.uwaterloo.ca/~msmucker/cw12spam/, last accessed May 3rd, 2017.

An important aspect of the data, which should be clearly mentioned here, is that a lower score represents a spammier page. Thus, the most spammy pages have a score of 0, while the least spammy pages have a score of 99.

4.2.4 FRANK1

Currently the approaches in this thesis are mostly limited to ClueWeb due to the presence of the annotated data FACC1. This is unfortunate as there are more recent webcrawls available which do not have these type of annotations, making them unsuitable for the methods described in this thesis. This section will introduce an efficient, albeit a very naive way, of generating a custom annotation set, in this case called FRANK1, or FReebase Annotations Naively $\mathcal{K}$omputed v1. The generated data for FRANK1 will have the exact same format as FACC1 with one exception, making it almost completely interchangeable with FRANK1. Results will show whether or not a naive implementation like this is a possible replacement for annotations which were created with a high precision in mind.

Algorithm 1 shows a pseudo description of the steps required to generate the FRANK1 data. These steps, and their results, will be described and analyzed in more detail further ahead.

Data: ClueWeb12 data & list of all entities
Result: FRANK1 data
remove descriptions in parenthesis from persons;
add one space padding around each entity;
lowercase entities;
initialize Aho-Corasick double array trie using entity data;
while *not annotated all documents* **do**
    get encoding from HTML header (UTF-8 default);
    convert ClueWeb12 document to string using the encoding;
    lowercase the document;
    replace the characters ".", ",", "!" and "?" with spaces;
    run the Aho-Corasick algorithm over the document;
    for each hit from Aho-Corasick generate an annotation;
end

Algorithm 1: Creating the FRANK1 dataset

Two data sources are required for this approach: ClueWeb12 and the entities one is interested in; in this case all entities from the provided data, which includes the names of all persons, the names of professions, both singular and plural and the nationalities, both the country as well as the actual nationality. An additional point that comes into play in this case is the descriptions that several persons have in the entity list, below are some examples:

- Aaron Williams (American football) /m/0bx_l4r
- Aaron Williams (cartoonist) /m/02x7xls
- Love (footballer) /m/0ddb_m
- Min (singer) /m/0hgprs2

- K (singer) /m/0l00br
- Do (singer) /m/01wb83m

Given that this approach uses string matching, the exact string "Aaron Williams (cartoonist)" is unlikely to appear (often). In this approach, the descriptions between parentheses are removed, significantly increasing recall, however also decreasing precision, as for example both Aaron Williams the American football player and Aaron Williams the cartoonist will now match the same sentences.

One space padding is added around the entities to prevent the Aho-Corasick algorithm from matching unintended entities and focusing on complete words, each entity is also lowercased, preventing inconsistencies. Given the profession 'model' and the plural 'models', if these exact words were matched on the following sentence it would have three hits:

```
The models were remodelling the supermodel's house.
```

Adding the one space padding leaves only a single hit:

```
The models were remodeling the supermodel's house.
```

Using a naive substring search approach is not feasible for this amount of data, as it has a complexity of $\theta(nm)$, where n is the length of the document and m the length of the pattern. This would have to be executed `n_entities x n_documents` times, with $385,726$ entities and $733,019,372$ documents. To improve performance and make the search for this many entities over all documents feasible the Aho-Corasick algorithm by Aho and Corasick was used [2], specifically, the implementation by hankcs using double array trie structure.[10] The data structure is initialized using all entities one wants to find, after which it can be reused for each document.

The ClueWeb12 webpages are initially loaded as byte representations, with the HTTP headers containing the original content type, which often includes the charset which was used to render and store the page. Byte representations of webpages are converted to strings using the charset specified (if available). If for any reason it fails to find a charset, or finds an invalid one, the conversion to string defaults to UTF-8.

Two preprocessing steps happen on each document, one of which is lowercasing it to match the lowercased entities, this helps increase recall, making capital usage in names like 'Seamus McGoon' consistent over all documents and making it possible to find entities that are otherwise capitalized, for example at the start of sentences. The second step is replacing specific characters, the period, comma, question mark and exclamation mark, with spaces. This helps with retrieval of entities that occur next to these characters, for example at the end of a sentence. The example below does not match against either the entity 'models' or 'carpenters'.

```
The models, who were moonlighting as carpenters, started remodeling.
```

However, after replacing the characters it finds both entities.

```
The models  who were moonlighting as carpenters  started remodeling
```

[10] See https://github.com/hankcs/AhoCorasickDoubleArrayTrie, last accessed July 5th, 2017.

After these preprocessing steps are completed the actual Aho-Corasick algorithm can be executed over this webpage. The specific implementation used here returns triples containing the matched entity, the start offset and the end offset. For the output to be consisted with FACC1 these triples are converted to the same format with the following fields:

- The document ID of the current page
- The encoding which was previously extracted from the header
- The matched entity with the one space padding removed
- The start character offset plus one, removing the one space padding from the offset
- The end character offset minus one, same as above
- The probability that this entity is tagged correctly, uses 1 as default
- The probability that this entity is tagged correctly based on the context, uses 1 as default
- The Freebase ID which is looked up based on the entity string

There are two important things to note. First of all is that a single matched string can have multiple Freebase IDs due to the detail removal in the person data as seen in a previous example, resulting in multiple annotations for the same entity in text. Secondly the offsets in these annotations are string offsets with the relevant encoding as opposed to the byte offsets used for FACC1. This results in having to convert the entire document to a string before the entity can be, for example, sliced out. On the other hand, with FACC1 this entity can be extracted while it is still in a byte representation, after which only this tiny bit has to be converted to a string, which is more efficient.[11]

5 Preliminary analysis

The goal of this section is to get a better understanding of what exactly is in both FACC1 and the spam ranking datasets.

5.1 FACC1

Statistics provided by the authors of FACC1 show that 456,498,584 out of 647,222,268 analyzed ClueWeb12 documents contain an entity, resulting in a total of 6,133,750,307 annotations.[12] This section investigates FACC1 a bit closer to see how well it can be used for the challenge.

[11] Given the time available for this thesis project, we unfortunately had to skip the additional coding necessary for the conversion from string offsets to byte offsets.

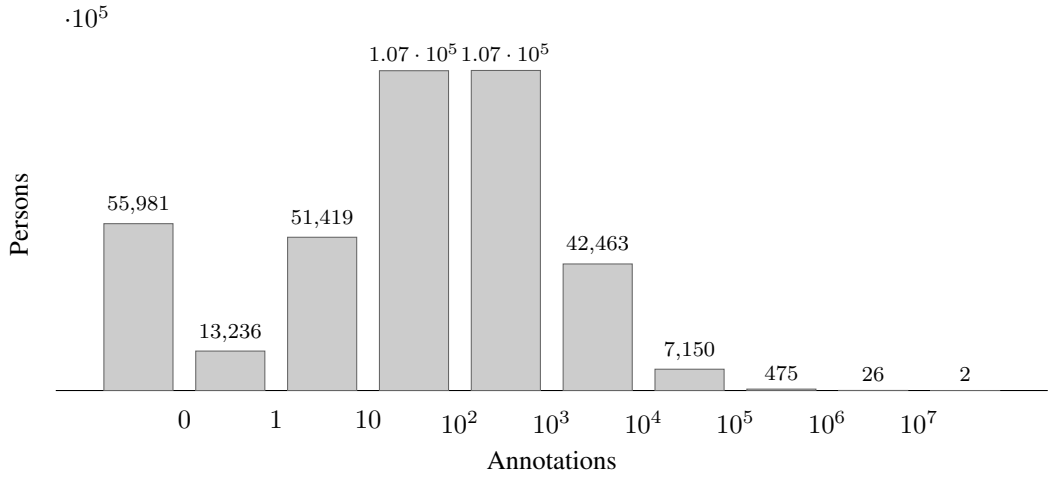
[12] See http://lemurproject.org/clueweb12/FACC1/ClueWeb12_stats.txt, last accessed May 15th, 2017.

5.1.1 Persons

The provided data consists of 385,426 unique persons and 385,426 corresponding Freebase IDs. These Freebase IDs are good for a total number of 548,580,693 annotations in FACC1, which is approximately 9% of all annotations. Assuming an even distribution of annotations over all 385 thousand persons this would come out at an average of 1423 annotations per person.

Unfortunately, and not completely unexpected due to the Zipfian nature of many types of data [16], the annotations are not evenly distributed over the available persons. Almost 66 thousand persons do not have a single annotation, 13 thousand persons only have a single annotation and 51 thousand persons have 10 or less annotations. Figure 3 shows how the annotation counts are divided over all persons.

Figure 3: Distribution of person annotations over different bins



A cursory inspection of the first group, the persons with no annotations, shows that many of their names appear to be non-English. Consider for example, Aad de Bruyn, a Dutch athlete, who was proficient in shot put, discus and hammer throw and won several national championships in the 30s and 40s. This would be in line with ClueWeb12 primarily focusing on English pages.

The Figure also shows two persons with more than ten million annotations, these are Jesus with 22 million annotations and Barack Obama with 15 million annotations, both having significantly more annotations than those ranked below them. Several other names at the top are George W. Bush (6.9M), Celine Dion (4.0M), Mitt Romney (2.7M), John McCain (2.2M), Bill Clinton (2.1M), Steve Jobs (1.9M), Hillary Clinton (1.7M), Tim Burton (1.5M) and Lady Gaga (1.5M).

This analysis raises two important issues, as described before:

1. Not all persons are available when working solely with FACC1
2. A large number of persons only has very few occurrences in FACC1

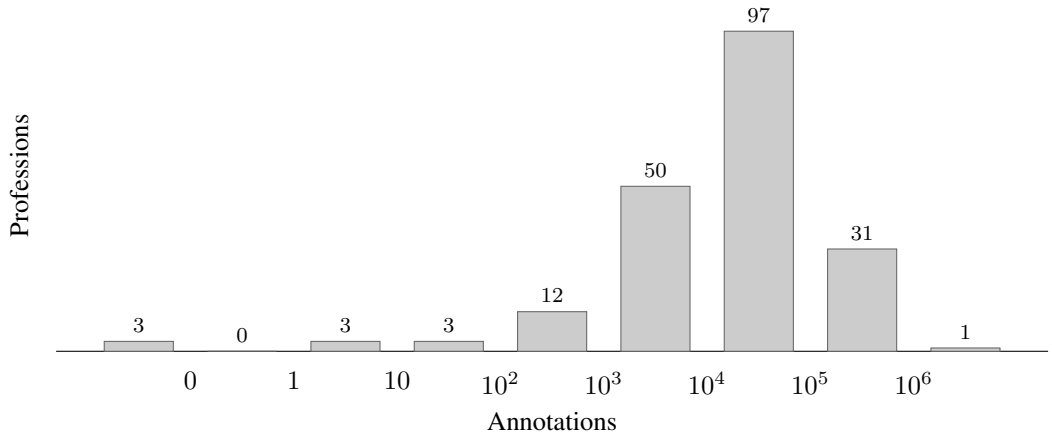
The consequences of these issues are that for issue 1: any program will be unable to accurately predict anything about those persons as long as it solely relies on FACC1. And for issue 2: persons with few occurrences are less likely to appear in relevant and useful context.

5.1.2 Professions

The provided data consists of 200 different professions, however, as opposed to the persons, the professions did not come with Freebase IDs, so these had to be manually added as described in a previous section. Manually adding the Freebase IDs leads to a total of 202 unique Freebase IDs. These 202 Freebase IDs have a total of 10,842,813 annotations in FACC1, which is 0.18% of all available annotations. Assuming an even distribution of annotations over all annotations each profession would on average have 53,677 annotations.

Much like for persons the annotations are not evenly distributed. The known Freebase IDs for talk show host, rodeo performer and ice hockey player have no occurrences in FACC1. Other professions like public speaker, orator and radio producer occur less than ten times. Figure 4 shows how the annotation counts for professions are divided over different bins.

Figure 4: Distribution of profession annotations over different bins



The most occurring profession in FACC1 based on the Freebase IDs is disc jockey, which appears about 1.4 million times. It is likely this is number is an effect of the specific entity tagger used for FACC1, as the word DJ might occur in contexts which do not immediately refer to a profession, but rather to, for example, a name. Several other popular professions are businessperson (764K), educator (391K), bishop (365K), professor (310K), prophet (294K), editor (286K) and TV editor (286K)

The missing professions and professions with a small number of annotations form the same issues as previously described in the section about persons. There is also an additional issue, visible in the previous paragraph which illustrates some popular professions in FACC1. Both editor and TV editor have the same count, which is due to the manually annotated Freebase IDs. These two professions, together with film editor and book editor, all share the same Freebase IDs, as there does not appear to be a unique Freebase IDs, which also occurs in FACC1, for each specific profession.

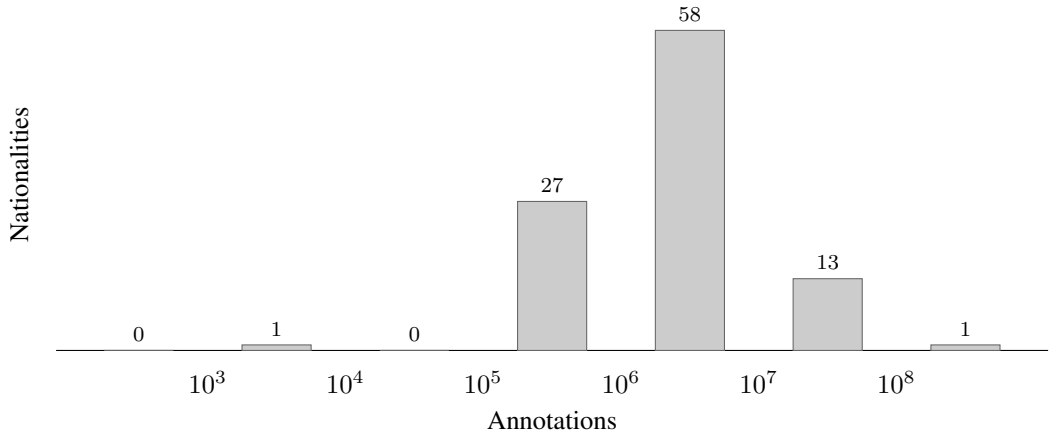
5.1.3 Nationalities

The provided data consists of 100 different nationalities, but just like the professions these did not come with Freebase IDs. A total of 105 different Freebase IDs are known for the 100 nationalities. Nationalities

have approximately half the number of Freebase IDs that professions have, however these nationalities have a total of 596,087,598 annotations, almost 600 times as many as professions, and a little more than the total number of annotations referring to persons. These annotations represent 9.7% of all available annotations in FACC1, and an average of 5,960,876 annotations per country.

As opposed to the other two groups, there are no entities here which do not occur in FACC1. All countries, except for one, occur at least 182,442 times in FACC1. The one exception here is the Freebase ID for Estonia, which appears only 1819 times. Figure 5 shows the distribution of all nationality counts.

Figure 5: Distribution of nationality annotations over different bins



The most common nationality is, unsurprisingly, The United States of America, which appears 144 million times. USA occurs almost four times as much as the second country, which is the United Kingdom and occurs 38 million times. Several other common countries are China (27M), Canada (25M), England (24M), India (20M), France (20M) and Japan (17M).

An important point to note here is that the nationality does not suffer from the same issues as the previously described two datasets. While Estonia does not occur as often as all other countries, all other countries do appear a lot, so the chance they can be found in relevant contexts and documents increases.

5.2 FRANK1

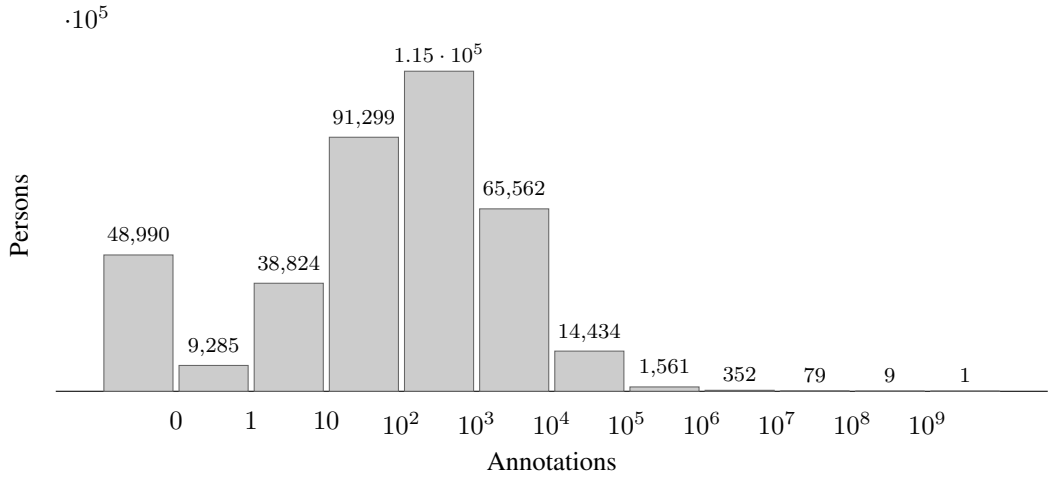
The FRANK1 annotation generator analyzed all documents and created a total of 11,955,817,671 annotations over 674,701,556 out of 733,019,372 ClueWeb12 documents. This is almost double the annotations that are available in FACC1, and FRANK1 only contains annotations for persons, professions and nationalities that occur in the triple ranking challenge.

5.2.1 Person

A total of 8,870,461,483 annotations are available for Freebase IDs belonging to persons, approximately sixteen times as many person annotations as there are in FACC1. Figure 6 shows the distribution of FRANK1 annotations for persons over different bins. This already shows that there are less people which

have none, one, between one and ten and between ten and a hundred annotations. So, in general it manages to tag more people than FACC1 does.

Figure 6: Distribution of FRANK1 person annotations over different bins



The highest number of annotations for a single person is also significantly higher. While FACC1 had two persons with more than ten million and less than a hundred million annotations, FRANK1 has a total of 441 persons who have more than a million annotations, with the most common person having 1,155,560,402 annotations. Unfortunately, however, these numbers are due to the previous choice of removing specifics about a person while generating the FRANK1 annotations. For example, for the person "Aaron Williams (American football)", the last part is removed and the annotation generator looks for "Aaron Williams". The person occurring most often in this data is Willy Maltaite, a Belgian comic creator and comic artist. He is mostly known by his pseudonym Will, and he is in the provided person list as "Will (comics)", meaning that any occurrence of the word "will" is tagged as this person.

Table 2 shows the top 10 most common persons in the FRANK1 data. This top 10 clearly illustrates why bluntly performing string matching might not always be a good approach and why a well performing named entity recognizer can significantly improve the quality of the data.

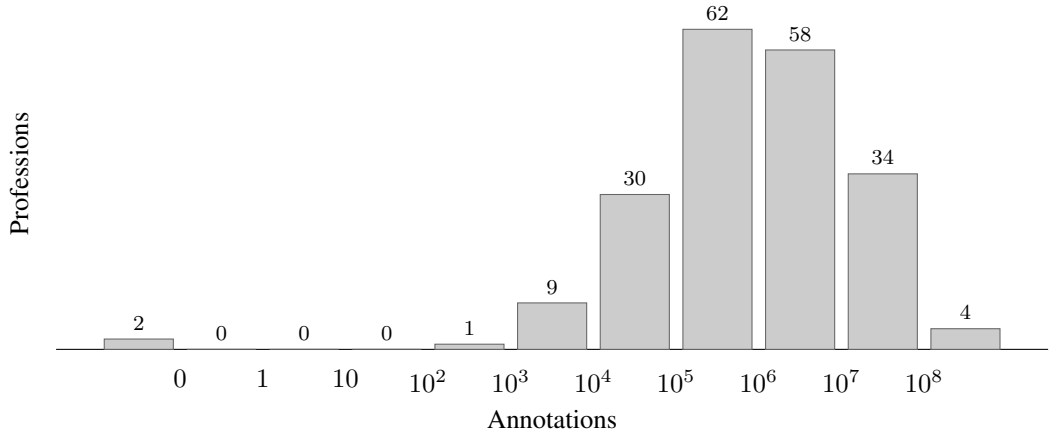
5.2.2 Profession

The FRANK1 annotation generator created 1,437,921,568 annotations for professions, approximately 130 times as many as FACC1. Figure 7 shows that both the average number of annotations per profession, as well number of annotations for the most common professions, are, unsurprisingly, higher than in the FRANK1 data. The five most common professions are artist, fashion model, model, manager and author, ranging from 192,342,212 down to 95,869,614 annotations. One point to note here though is that unlike the person data, the profession and nationality data have entities that share Freebase IDs, so while both model and fashion model appear here, a ‘fashion model’ might also appear as ‘model’ in the documents.

Table 2: Top 10 most common persons in the FRANK1 data

Person	Count
Will (comics)	1,155,560,402
Do (singer)	874,619,114
May (singer)	717,023,549
Min (singer)	401,989,879
Love (footballer)	342,530,103
Jan (comics)	255,364,035
J (Korean singer)	243,783,254
K (singer)	178,820,516
Case (singer)	175,305,015
Nature (rapper)	136,706,285

Figure 7: Distribution of FRANK1 profession annotations over different bins



5.2.3 Nationality

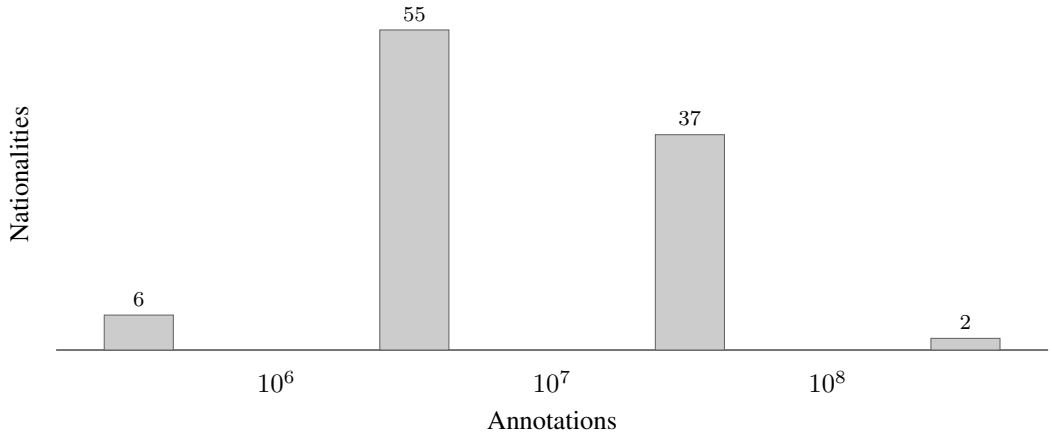
For the nationalities 1,647,434,620 annotations were generated, almost three times as many as in FRANK1. Figure 8 shows the distribution of annotations, indicating that each and every nationality occurs at least a hundred thousand times. Whereas Estonia only occurs 1819 times in FACC1, it now has 1,588,753 annotations. The two most common countries are still the USA and the UK, occurring respectively 132 million and 113 million times.

5.3 Spam rankings

The spam rankings provided by Cormack et al. [6] run from 0 through 99. One assumption that can be easily tested is that spammier documents are likely to contain more keywords due to possible keyword stuffing, and, in turn, contain more entities in FACC1 entities.

Figure 9 shows the average number of FACC1 entities per document. This shows that for most spam scores the average number of entities is between 13 and 15, except for the interesting valley around spam

Figure 8: Distribution of FRANK1 nationality annotations over different bins



scores 22 to 45, where the average number of entities dips to approximately 10 per document. Both person and nationality seem to have the same decline around that value, but the average number of persons also keeps declining after a spam score of 80.

Figure 9: Average number of entities for documents

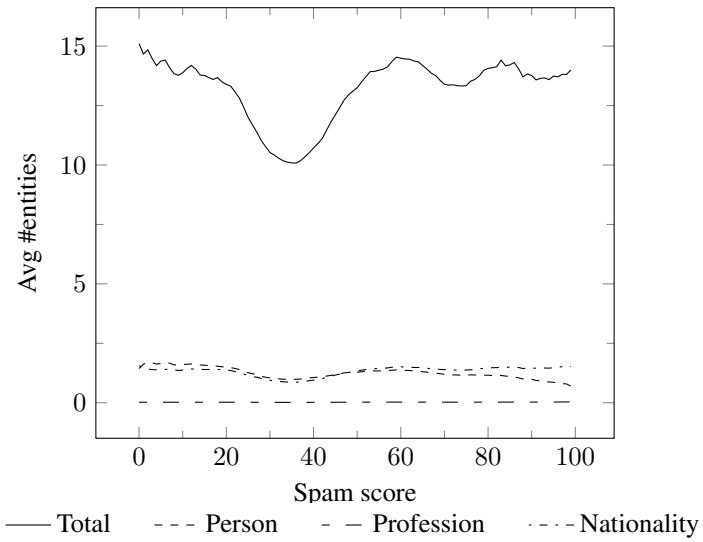
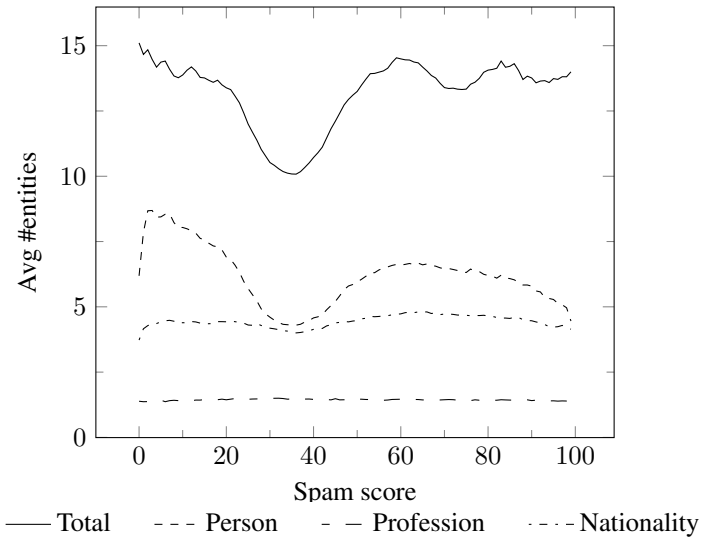


Figure 10 shows the same values for the total average number of entities per spam score. However, the average of person, profession and nationality counts was calculated using a slightly different method. The average is only calculated over the documents that contain at least a single entity for that specific type. The Figure shows the same valley for persons, and a slight dip for nationality. The number of mentions for professions and nationalities are quite stable and independent of the spam score. One interesting piece of

information that this new Figure shows is that both persons and nationality have an increase immediately after the first few lowest spam scores.

Figure 10: Average number of entities for documents where entities are present



6 Approaches

Approaches used for the experiments in this thesis are described in more detail in this section. The first approach is based on co-occurrence counts in annotations, the second approach is a combination of co-occurrence counts and wiki abstract occurrences, while the third approach utilizes snippets from ClueWeb12 for entity ranking.

6.1 Co-occurrence counts

There are many ways to utilize the FACC1 data, be it with or without the associated ClueWeb12 or other datasets. This section describes a number of different approaches, all based on the use of co-occurrence counts between persons and nationalities or professions. The assumption for this approach is that the frequency of entities co-occurring in the same document is indicative of the relevance between these entities. These resulting counts can be influenced by various factors, some of which will be described in the following sections.

For example, co-occurrences of Barack Obama and Politician should be more frequent than Barack Obama and Author, which in turn should be more frequent than Barack Obama and Farmer.

6.1.1 Simple co-occurrence counts

This approach employs the previously mentioned assumption that co-occurrence counts indicate the level of relevance between entities. Given annotations like the ones in Table 3 this results in the co-occurrence counts seen in Table 4.

Table 3: Example annotations showing entity names instead of Freebase IDs

Document	Entity
Doc1	Peter Jackson
Doc1	Tom Cruise
Doc1	Tom Cruise
Doc1	Film Producer
Doc1	Actor
Doc1	Actor

Table 4: Example annotations

Person	Profession	Co-occurrence count
Peter Jackson	Film Producer	1
Peter Jackson	Actor	2
Tom Cruise	Film Producer	2
Tom Cruise	Actor	4

Generating the co-occurrence counts is done in several stages. Taking the previously shown FACC1 sample data into consideration it has to perform the following steps:

1. Group all annotations per document;
2. Extract all persons, professions and nationalities into their own lists;
3. Calculate the Cartesian product between persons and professions and persons and nationalities;
4. Convert the current Freebase IDs to their actual entity names;
5. Sum all the resulting counts for each combination.

The steps described above are used to calculate the co-occurrence counts for combinations of persons and nationalities or professions. The actual implementation of this algorithm has several additional steps. These steps do not affect the effective outcome, but they do improve the speed. One of the actions is filtering out all irrelevant Freebase IDs before they are grouped together. Another additional action is filtering out any combinations in the third step that do not occur in the knowledge base. These steps reduce the information that has to be transferred and handled in the program.

6.1.2 Co-occurrence counts with spam ranking thresholds

All documents available in ClueWeb12, and in turn in FACC1, have been given a spam rating by Cormack et al. [6], as shown in a previous section. There are, again, several ways to apply this dataset to this problem, and this section will describe one of those.

Several assumptions could be made, for example that spammier documents might have a higher entity, and in turn, annotation count. Preliminary analysis in section 5.1 however has shown that this is not immediately the case and that for person entities it might actually be the other way around.

Another assumption is that spammier documents exhibit a larger variety of entities, leading to less relevant combinations. This assumption would lead to spammy documents adding noise to the generated scores.

Below are the steps required for determining the co-occurrence counts with spam ranking thresholding, where documents below a certain spam threshold can be left out:

1. Group all annotations per document;
2. Join all spam scores with the documents;
3. Extract all persons, professions and nationalities into their own lists;
4. Calculate the Cartesian product between persons and professions and persons and nationalities;
5. Combine each combination produced by the previous step with the spam score of the document it is from;
6. Convert the current Freebase IDs to their actual entity names;
7. Sum all resulting counts based on what bin the spam score falls into.

Table 5 shows some example data, where all three documents have a different spam score, but all contain the same co-occurrence. Table 6 shows what the resulting scores would look like. These scores would be grouped into different bins based on their spam score, together accumulating the exact same number of entity co-occurrences as found in the previous section.

Table 5: Example annotations with spam scores

Document	Spam score	Entity
Doc1	5	Tom Cruise
Doc1	5	Actor
Doc2	38	Tom Cruise
Doc2	38	Actor
Doc3	91	Tom Cruise
Doc3	91	Actor

At this point there are several options as to what to do with this information. Scores can be weighted based on their spam score, where a lower spam score reduces the influence of the count. In this case, we opted for simple thresholding based on spam-score, where a minimum spam score will be set and only

Table 6: Co-occurrence counts when applying spam score binning

Person	Profession	$0 \leq s \leq 10$	...	$30 < s \leq 40$	...	$90 < s \leq 100$
Tom Cruise	Actor	1		1		1

counts that occur in documents above this threshold will be considered, as a higher score indicates a less spammy document.

6.1.3 Co-occurrence counts with distance thresholds

This approach uses an additional piece of information from the FACC1 data that the previous approaches do not use: The location of the entity in the document. Each entity mention has associated information of the exact byte offset where it starts and where it ends. This information can be used to verify whether or not the following assumption holds: The distance between entities is indicative of the relevance between these entities.

To test this assumption the usual co-occurrence counts are used, however much like the spam rank threshold a threshold is put on the maximum distance between two entities. The distance is based on the start and end of a word, or end and start depending on which comes first, not on the center of a word. Given the sample annotations in Table 7, the distance between Actor and Tom Cruise is $21 - 15 = 6$, while the distance between Tom Cruise and Carpenter is $50 - 31 = 19$. In this case if the threshold were set on 10 then the profession carpenter would not count anymore due to the distance and the assumption that it would not be relevant.

Table 7: Example annotations with offset values

Offset start	Offset end	Entity
10	15	Actor
21	31	Tom Cruise
50	59	Carpenter

6.1.4 Scaling counts to scores

Table 8 shows all known professions for Tom Cruise according to the provided knowledge base and their co-occurrence counts. This example shows how some combinations can occur significantly more often than others, which poses a challenge. The goal is to generate scores between 0 and 7 to indicate the relevancy, so the co-occurrence counts have to be scaled down to that range.

Two methods are described in this section: A linear and a logarithmic scaling. The linear scaling takes all counts for a single person, divides this by the maximum count for that person and rounds it down to the nearest integer. The logarithmic scaling takes all counts for a single person, adds one, takes the log and performs linear scaling on these values. Both the results from the linear as well as the logarithmic scaling are then multiplied by 7, resulting in the final 0-7 scores. The linear scaling is similar to maplin by Bast et al. and while our logarithmic approach is similar in idea to maplog, but takes a different approach [3], as

Table 8: Co-occurrence counts for Tom Cruise taken from FACC1

Person	Profession	Co-occurrence
Tom Cruise	Actor	1245
Tom Cruise	Film Producer	36
Tom Cruise	Screenwriter	213
Tom Cruise	Television director	0

theirs starts with linear scaling, transforming the values into a 0-1 range, which are then multiplied by 2^7 of which the log2 is taken.

Table 9 shows the computed relevance scores for the counts previously seen in Table 8, this clearly illustrates the issue with large discrepancies between relevance scores when a certain entity might occur significantly more often. Using the logarithmic approach helps dampen this effect and produce more intuitive relevance scores.

Table 9: Co-occurrence counts for Tom Cruise taken from FACC1

Person	Profession	Co-occurrence	Lin	Log
Tom Cruise	Actor	1245	7	7
Tom Cruise	Film Producer	36	0	3
Tom Cruise	Screenwriter	213	1	5
Tom Cruise	Television director	0	0	0

Missing data is currently handled as having the lowest relevance score, the score 0. Changing how this is handled might also change the resulting scores, especially accuracy. As previously mentioned accuracy views a score as correct if it is within two points of the true score, so the 0 predicted for missing is also viewed as correct if the true score is 1 or 2. Changing this default value to 5 would mean that this prediction would be viewed as correct for scores 3 through 7, as seen below.

$$1\,2\,\overbrace{3\,4\,\underbrace{5}_{\text{Predicted}}\,6\,7}^{\text{Scored as correct}}$$

Given that missing data is sometimes caused by persons not having any FACC1 annotations at all, using 5 to replace all 0 scores is likely to increase accuracy, the metric which is used to judge different approaches and choose the winner for the challenge.

6.2 Co-occurrence counts and wiki occurrences

This approach is based on the previous co-occurrence counts, combining it with information extracted from Wikipedia abstracts. Results from this approach were submitted for the final run at the end of 2016, before most of the other research in this thesis was done, the precise implementation can be viewed in the submitted workshop paper [9], however it will also be briefly described here.

As mentioned it uses two types of information, the first one of which is based on the abstracts. Combinations of a user and an entity, be it a nationality or profession, are given a relevance score of 7 if the entity is the first one of its type to appear in the abstract, otherwise the combination is given a relevance score of 0. The second one type is based on the co-occurrence counts as described in the previous sections. The number of occurrences between a person and the entity is calculated, of which the log is taken and then normalized to a 0-7 scale. It should be noted that due to implementation difficulties the scores described in the result section are based on a subset of 51.2 million annotations of FACC1. Score fusion is done by taking the maximum result of both approaches and if this is zero a default score of four is used.

6.3 Machine Learning with ClueWeb12 snippets

The approaches discussed so far all rely on a direct mapping of co-occurrence counts to relevance estimates. We will now consider approaches using machine learning, to allow for less trivial relations between the count observations and the resulting relevance estimates. The assumption for the machine learning approach developed in this thesis is that professions or nationalities can be accurately predicted for snippets belonging to users when the models are trained on snippets belonging to either of the two entity groups. This opens up a large amount of training data as seen in the FACC1 analysis section. Ideally this approach learns context sensitive clues, increasing the recognizability of sentences in which persons occur. While the assumption is intuitively clear and seemingly straightforward to develop, there are several key issues that have to be taken into consideration.

6.3.1 Generating FACC1 snippets

Snippets are generated from ClueWeb12 using FACC1, which tells exactly what entity is in which document at what byte offset. These snippets have a context of 256 bytes both left and right, meaning that, at first, they are all at least 512 bytes. This size provides the snippets with a generous context, though in many cases it will be shorter (as explained later in this section).

In total, there are two types of snippets: With and without the FACC1 entity mentions in the middle, where the entity mention is the name of the profession or nationality. The second version is introduced to investigate how much influence the FACC1 entity mention has during learning if it is still present, and how well it works if it solely learns context.

Algorithm 2 shows how the snippets with entity mentions are extracted. It combines each annotation with their respective document and then uses the info in the annotation to extract the snippet. The last step tries to remove the HTML markup from the extracted snippets to extract the text, given that the snippets often have HTML in there which likely does not add any useful information.

The process is mostly the same for snippets without entity as seen in Algorithm 3. Instead of doing a single slice, it slices the start and end separately and combines this, extracting only the text surrounding the entity mention while leaving out the entity mention itself. The reasoning for leaving out the entity mention is that it might allow machine learning algorithms to better focus on context, instead of maybe overfitting on single features (the entity mentions), which are almost always the same within a class, and acting as a complicated co-occurrence detector. Whether or not this adds any valuable information can be seen in the results.

Data: FACC1 annotations & ClueWeb12 data
Result: ClueWeb12 text snippets
join FACC1 with ClueWeb12 on docId;
while *not processed all annotations* **do**
    slice ClueWeb12 document from `startOffset-256` through `endOffset+256`;
    convert to string using encoding from annotation;
    parse text from HTML;
end

Algorithm 2: Extracting snippets from ClueWeb12 using FACC1 with entity

Data: FACC1 annotations & ClueWeb12 data
Result: ClueWeb12 text snippets
join FACC1 with ClueWeb12 on docId;
while *not processed all annotations* **do**
    slice start of snippet from `startOffset-256` through `startOffset-1`;
    slice end of snippet from `endOffset` through `endOffset+256`;
    combine start and end slices;
    convert to string using encoding from annotation;
    parse text from HTML;
end

Algorithm 3: Extracting snippets from ClueWeb12 using FACC1 without entity

The first part of the snippet ends at `startOffset-1`, because using `startOffset` would in many cases leave slight inconsistencies. For example slicing `models` from the sentence `The models were redecorating` would result in double spacing between `The` and `were`. Ending the slice one character earlier removes these inconsistencies in many cases.

Completely cleaning the snippets up and removing HTML is a challenge with the way that the snippets are currently extracted. Because the snippets are sliced at specific char offsets there is no guarantee that it would not slice in the middle an HTML tag, leaving half of it in and half of it out, which is something that HTML parsers likely cannot handle. An alternative way would be to perform cleanup before extracting snippets, resulting in completely cleaned up text, however this would invalidate the byte offsets provided in FACC1, introducing the need to keep track of any position changes during the cleanup.

Handling HTML markup is however necessary, even if only in a crude way, as it can remove a lot of unnecessary information. Listing 4 shows what a raw snippet looks like, taken straight from ClueWeb12 without any cleanup. This snippet contains a lot of useless information which would likely not add any valuable context for machine learning to learn from.

```
1 na Hospital, Napa Valley, California<br></div>
2
3 <div class="emptyClear"> </div>
4 </div>
5
6 <h3 class="productDescriptionSource">About the Author</h3>
7 <div class="productDescriptionWrapper">
8 <div><b>Jane A. Plant, Ph.D., C.B.E.</b>, is one of Britain's most distinguished
      scientists. She is chief scientist of the British Geological Survey and continues
```

```
to sit on many influential government and international committees. In 1999, she
was awarded Britain's most prestigious honor, the Lord Lloyd of Kilger
```

Listing 4: Snippet for the annotation (`clueweb12-0500tw-00-25246 ISO-8859-1 Britain 163626 163633 0.982396 0.001314 /m/07ssc`) without HTML cleanup

Listing 5 on the other hand shows what this snippet would look like after the HTML parser extracted the text from the snippet. As one can see all the HTML information has been removed from this snippet, leaving only the actual relevant text.

```
1 na Hospital, Napa Valley, California About the Author Jane A. Plant, Ph.D., C.B.E., is
    one of Britain's most distinguished scientists. She is chief scientist of the
    British Geological Survey and continues to sit on many influential government and
    international committees. In 1999, she was awarded Britain's most prestigious
    honor, the Lord Lloyd of Kilger
```

Listing 5: Snippet for the annotation (`clueweb12-0500tw-00-25246 ISO-8859-1 Britain 163626 163633 0.982396 0.001314 /m/07ssc`) with HTML cleanup

It should be noted though that this is an example that works well. Handling HTML in a large crawl like ClueWeb12 is a daunting task, and any approach that is easy to implement will encounter many exceptions to the rule, including, for example, CSS styles in the snippet. Another issue is that for every character the parser removes, it also reduces the length of the snippet. In some cases reducing it to lengths that are not useable anymore, as seen in the few examples below. The way this is handled will be described further ahead in this thesis.

- `March 1, 2012 JT --`
- `« Back to Profile Jesus Had by`
- `"class="topic" title="Beyonce" >BeyonceRihannaKaty PerryLady Gagamore ... Site Links Ask a Question`

6.3.2 Generating FRANK1 snippets

The primary difference between FRANK1 and FACC1 annotations is what the offset represents. While FACC1 offsets are based on the byte representation, FRANK1 offsets are based on the encoded string representations. This in turn means that snippet generation has to happen slightly differently. The modified algorithms are visible in Algorithms 4 and 5.

6.3.3 Sampling snippets

Initial analysis has already shown that profession and nationality entities are not evenly represented in FACC1, this was an issue in previous described approaches, but perhaps even more so for the machine learning approach, where unbalanced training data is a well-known cause to lead to inferior results. Other factors that need to be addressed are the variety in snippet quality, and repetitive information in the crawl. To address these issues, an approach to produce a balanced and representative training set has been developed using a sampling method, which is based on several parameters:

Data: FRANK1 annotations & ClueWeb12 data
Result: ClueWeb12 text snippets
join FRANK1 with ClueWeb12 on docId;
while *not processed all annotations* **do**
    convert to string using encoding from annotation;
    slice ClueWeb12 document from `startOffset-256` through `endOffset+256`;
    parse text from HTML;
end

Algorithm 4: Extracting snippets from ClueWeb12 using FRANK1 with entity

Data: FRANK1 annotations & ClueWeb12 data
Result: ClueWeb12 text snippets
join FRANK1 with ClueWeb12 on docId;
while *not processed all annotations* **do**
    convert to string using encoding from annotation;
    slice start of snippet from `startOffset-256` through `startOffset-1`;
    slice end of snippet from `endOffset` through `endOffset+256`;
    combine start and end slices;
    parse text from HTML;
end

Algorithm 5: Extracting snippets from ClueWeb12 using FRANK1 without entity

- Spam score of the document,
- Length of the snippet,
- Deduplication,
- Balancing classes.

During snippet sampling, which generates the actual datasets used for the machine learning approaches, the parameters above are used to influence the resulting set. The first parameter is used to set a minimum threshold on the spamminess of the documents, so if one wants samples from all snippets this can be set to 0, if one wants samples from only the top half of non-spammy documents, one can set 50 as minimum spam threshold.

The second parameter is the length of the snippet, which is where the previously described HTML parsing comes into play. When parsing the text from the snippets it can happen that the resulting snippets are very short and contain little information. This parameter allows for a threshold to be set for the minimum snippet length, for example that it should still contain at least half of the original length.

Deduplication is performed due to certain websites occurring multiple times with slight variations. An example would be a QA site, where each answer has its own page which contains the original question. If an entity were to fall within the question and the snippet length did not go out of bounds of this question, each page with an answer would contain the exact same snippet. To prevent too much duplicated data, each snippet within the group of snippets for a specific Freebase ID is unique within that group.

An important aspect is keeping a balanced distribution of classes in the sampled snippets. However, the initial analysis of FACC1 already showed that the data by itself is not actually balanced within entity

groups like professions or nationalities. A threshold is added to sampling, setting a maximum number of snippets for each Freebase ID, which should alleviate the issues caused by the unbalanced data when using it for machine learning, albeit only partially. If a Freebase ID has more than the maximum number of snippets then the data is randomly sampled to reduce it. If an ID already has less snippets, then the entire set is taken.

A default sample size of 10,000 snippets per Freebase ID is used, which are randomly sampled from the dataset, this is enforced for all person as well as nationality datasets, be it with FACC1 or FRANK1 data. The primary exception to the default sample size is the profession data from FACC1, which has a large number of classes occurring below the 10,000 snippet limit as also previously illustrated in the analysis section. Sample sizes for this data were picked by taking the average of all profession occurrences after certain spam thresholds and rounding this to the nearest 500, which resulted in sample sizes of 10,500 when using no spam threshold, 7000 for a threshold of 50, 4000 for 75 and 2500 for the strictest spam threshold of 88. FRANK1 profession data has a larger number of occurrences, thus the default sample size of 10,000 was maintained here.

The person data also uses the default sample size as described above, however it should be noted that this is not due to the previous reason of machine learning, as these person snippets are solely used for predictions and are completely independent of each other. This sample size is largely used to limit the amount of output data produced.

6.3.4 Feature generation

Extracting features from text is a widely researched subject and the possibilities are nearly endless. Any text mining book will likely give an extensive list of possibilities, for example basic features like a bag of words model, different versions of n-grams, like character n-grams, word n-grams, or skip-grams. Features like word n-grams and word skip-grams can in turn also be applied to the actual syntax of a sentence, extracting information about how a verb occurs after a noun.

Several of these features have been tried and tested in entity ranking, for example both word embeddings [25] and bag of word models [14, 13] perform well. Due to the way the data is extracted for this project, resulting in noisiness due to possible words that are cut off and remaining web code in the snippets, cleanly extracting words can be a challenge. Thus, a character based feature was chosen: character n-grams. These character N-grams are used to compute the TF-IDF scores for these N-grams, reducing the scores of N-grams that occur in many documents.

Character N-grams are computed over complete sentences, removing the need for any tokenization. These sentences are lowercased before the N-grams are computed, reducing the dimensionality of the resulting feature space. Several different values for N will be used and compared to each other, however research has shown larger character N-grams, e.g. 4 or 5, to work quite well for information retrieval tasks [15].

Computing these N-grams and their corresponding counts can be a computationally expensive task for larger document collections and larger feature spaces, as it requires a dictionary to map terms to feature indices. This method also requires this dictionary to be fitted on the training data, before it can be used to determine the counts of the test data. While it is computationally expensive, it also has several advantages,

for example being able to filter on minimum and maximum document frequencies, and being able to actually understand the features.

Unfortunately, both time as well as memory and storage capacities had to be kept in mind for these experiments. To solve the expensive nature of normal document frequency calculations for large amounts of data an alternative implementation was used, one based on the hashing trick [23]. The idea behind the hashing trick is that one does not need to keep track of a dictionary anymore, which maps terms to their index. Instead of a dictionary, a single array is used for each document. This array has size K, which represents the number features in the feature space. Counting N-grams is done by hashing each one of them, for example using MurmurHash3 (a non-cryptographic hash function), and taking modulo K. The remaining value is used as the index in the K sized array, and this index is incremented by one.

There are several reasons why using the hashing trick improves speed, all of which have to do with the fact that the first approach uses a dictionary. The dictionary for the first approach has to be created using training data, to know which terms map to which feature indices, only after this can the feature vectors be generated for both the train as well as the test data. This means that when performing 5-fold cross-validation the creation of the dictionary and the generation of the train as well as the test features all have to happen five times. The hashing trick however has a fixed mapping of hashes to feature indices, meaning that it can compute the feature space for the entire dataset in one go, which can then simply be split into the respective train and test data for cross-validation.

Generating the entire feature space using the hashing track with 5-grams on the FRANK1 profession data without a spam threshold and with the entity mentions present in the snippets takes approximately 800 seconds on the server used for the experiments. Fitting the model for the traditional method on train data, as well as generating the feature vectors, take both approximately 640 seconds. Generating the feature vector for the test data takes an additional 160 seconds. In the situation of 5-fold cross-validation the feature hashing method takes approximately 800 seconds, while the traditional method would take $(640 + 640 + 160) * 5 = 7200$ seconds.

The additional memory usage of the traditional method depends on the type of features that are used. For 5-grams an additional 208MB would be used, while in the case of 7-grams this would already be 1.1GB.

The hashing trick also reduces the feature dimensionality. For example, using 5-grams for the FRANK1 profession data, without spam threshold and with entity mentions present in the snippets, the traditional method produces a feature space of 1,711,947 by 7,974,504 with a total of 651,078,022 elements, while the method using the hashing trick results in a feature space of 1,711,947 by 1,048,576 with 650,962,860 elements. Unfortunately, this dimensionality reduction also means that hash collisions can occur, where different N-grams may be mapped to the same index, as indicated by the smaller number of columns as well as elements.

Generally spoken 2^{20} distinct feature values are used unless otherwise specified.

6.3.5 Cross-validation and Evaluation

Performance of different parameters, for example different spam thresholds on the data or different n values for n-grams, are measured using five-fold cross-validation over the profession and nationality entity

sets. The predictions on the test set in the folds are compared to the true values using a weighted F1 score, which calculates the F1 score for each label and averages these based of the number of occurrences each label has in the true data. This metric provides a good balance between recall and precision, while also taking class imbalance into account.

An important point to keep into consideration here is that information about persons does not come into play here. Cross-validation will either use profession or nationality data depending on the run, meaning that both the train as well as the test partitions of the data during cross-validation are about either professions or nationalities. The weighted F1 score will only measure how well the model is able to predict the relevant entity, so whether or not a snippet about a profession or nationality is classified as the correct profession or nationality. Actual evaluation of the models on persons is done by converting the probabilities produced by the models to relevance scores, which is described further ahead, and using the official metrics to evaluate these generated relevance scores.

6.3.6 Learning algorithm

As described in the related work section there are several methods to tackle a problem like this. Approaches might use a binary approach, where each label is judged separately, or a non-binary approach, where correlations or relationships between labels are also taken into consideration.

The primary focus in this thesis is not the algorithm used for the machine learning approach, but rather the data and the entire process, which has resulted in the use of a more general-purpose machine learning algorithm, one that approaches the problem in a binary approach, judging each label separately. There are several algorithms that are able to do this, for example random forests and support vector machines, but in this case logistic regression was chosen. This choice was also partly due to the availability of implementations in `spark.ml` (and in a lesser extent `scikit-learn`, due to its more extensive library of possibilities).

Logistic regression is used to estimate the probability of a binary class based on one or more features. This initially forms an issue as the problem in this thesis consists of a minimum of a hundred classes, however one-versus-rest (OVR) is used to tackle this. A total of n (number of classes) logistic regression classifiers are trained, each one on data from one class, against data from all other classes.

Two different types of predictions are used from this OVR logistic regression approach. For cross-validation a straightforward prediction is used; given a snippet, all n logistic regression classifiers produce their score and the highest is taken as the correct class. For applications in entity ranking however, a slightly different prediction format is required. Just like the previously described prediction, given a snippet, all n logistic regression classifiers produce their score, however instead of choosing the highest ones, all scores are normalized.[13] These scores are then used, after further rescaling, for entity ranking.

The actual logistic regression implementation uses an L2 penalty with stochastic average gradient. Other than what is mentioned before, most default values were kept from the implementation which was used, except for the tolerance which was changed from 10^{-4} to 10^{-3} to speed up training while having a minimal effect on the weighted F1 scores.

[13] See https://github.com/scikit-learn/scikit-learn/blob/ab93d657eb4268ac20c4db01c48065b5a1bfe80d/sklearn/linear_model/logistic.py#L1259, last accessed August 2nd, 2017.

6.3.7 Scaling to relevance scores

Scaling from logistic regression probabilities to actual relevancy scores is done using a linear and a logarithmic method, which this time are both taken from the paper by Bast et al. [3], instead of using a different log scaling. Scaling happens per person, so the probabilities for one person will not influence those of another. If the description says that the probabilities are divided by the maximum probability, then it means that the probability that a single person has for entities is divided by the maximum probability that person has for an entity.

The maplin method starts by scaling all probabilities generated by logistic regression to an interval of $[0, 1]$ by dividing probabilities by the maximum probability. These values in the $[0, 1]$ range are then multiplied with seven and rounded to the closest integer to get the final relevancy scores.

The maplog method starts with the same scaling of probabilities to a range of $[0, 1]$, these values are then multiplied by 2^7 and the $\log_2$ is taken. These resulting values are then rounded to the closest positive integer, so probabilities that generate a negative score, for example 0.001, are rounded up to 0.

The start of this section described that scores for a person are based on all probabilities that they have for different entities, however there are two ways to interpret this, and both results are also available in the results section. The first approach uses all probabilities that logistic regression predicts for a person, so nearly all 100 nationalities or 200 professions. This produces a range of probabilities for persons, however a single wrong prediction of an entity which is completely irrelevant can have the effect that it scales all other actually relevant entity scores down. A way to partially avoid this problem is by filtering the predicted probabilities using the provided knowledge base. All probabilities that a person has for a profession or nationality which does not occur with them in the knowledge base is discarded, reducing the chance of irrelevant professions or nationalities influencing all other relevancy scores.

An additional method of improving accuracy can be applied: Reducing the interval of scores from $[0, 7]$ to $[2, 5]$. Any prediction which is below 2 or above 5 is a risk in terms of accuracy, the score 2 will cover all true values from 0 through 4 and the score 5 covers all true values from 3 through 7. This reduction is done by first calculating scores in the normal $[0, 7]$ interval, followed by increasing or decreasing all values that fall outside of the $[2, 5]$ range to the closest valid value.

7 Results

This section will cover the results from the approaches described in the previous section. Not all of these results are directly comparable to each other, as they might have a different meaning or purpose.

The first section covers the co-occurrence counts, comparing different results from different methods, for example the accuracies achieved when using all data, as opposed to thresholding on spam or inner-document distance. Results from co-occurrence counts are generated using the provided metrics from the challenge and are tested on the person data, both train as well as test data.

Actual results submitted to the challenge are shown in the second section, which describes the results when using co-occurrence counts with entity occurrences in Wikipedia abstracts.

The following section is slightly different from the other sections discussed here. This machine learning section shows results as described in section 6.3.5, the weighted F1 scores here are averaged over the

results of performing 5-fold cross-validation on either profession or nationality data. Any person data whatsoever is not used here, as the primary goal for this section is to show how well these models are able to predict profession or nationality classes on actual profession or nationality snippets, the reason being that if the models can not accurately classify these snippets they also would not be able to classify snippets belonging to persons.

The fourth section covers the actual results of the machine learning models on person data, including both train as well as test data. Different approaches here compare results based on maplin and maplog, whether or not to use the entire probability range generated by logistic regression, and whether truncating the scores to a smaller range helps, specifics are available in section 6.3.7. The results generated here are in the same format as those for co-occurrence counts, making these comparable.

The last section takes the best profession and nationality machine learning models based on person training data from both FACC1 as well as FRANK1 and combines these models to generate results as a sort of mock submission, making the results generated in this section comparable to actual other results submitted to the WSDM Triple Scoring challenge, and those seen in section 7.2.

All machine learning sections also discuss the use of entity mentions, whether or not entity mentions being present in the snippets before the n-gram features were generated improves results. Sections 6.3.1 and 6.3.2 show how this pertains to the actual snippets, whether or not the annotated word, or entity, for example 'carpenter', is still present in the snippet. What this means for the machine learning approach is that the actual feature vectors used for learning and predicting are generated using either snippets that do or do not have this entity mention, so these are two completely different feature vectors, they are not manually changed when the snippet should not be present. If it is stated that the data has no entity mentions present this means that the TF-IDF scores used for training and predicting were based on n-gram counts generated using snippets that do not have this entity present, and vice versa with the snippets when entity mentions should be present.

7.1 Co-occurrence counts

Several variations of the co-occurrence counts were described: Using all the data, using spam-thresholding and using inner-document distance-thresholding. Results for all these variations are described in this section.

One of the challenges of the co-occurrence counts is that the data is quite sparse, several person-profession or person-nationality combinations might not even occur. This also plays a part in the two different thresholding approaches, on distance and spam score. A higher threshold results in higher sparsity, which has a negative influence on the challenge metrics. To combat this an additional way of calculating the metrics is introduced. Persons who do not have a single co-occurrence with any of their entities in the profession or nationality kb, depending on what the type of the current task is, are removed before the metrics are calculated. This way a stricter threshold should not automatically worsen the metrics, and the focus is more on how accurate the actual remaining predictions are.

Full results for this section are provided in the appendix, specifically section A for results using FACC1 data and B for results based on FRANK1 data. Both sections cover the results based on spam (A.1, B.1) and inner-document thresholding (A.2, B.2).

Default co-occurrence counts using FACC1 data with linear scaling achieves an accuracy of 0.505 on the profession and 0.581 on the nationality test data. Logarithmic scaling achieves accuracies of 0.497 and 0.631. FRANK1 achieves respective scores of 0.532 and 0.571 on the profession and nationality data with linear scaling and 0.550 and 0.661 with logarithmic scaling. This is using the entirety of the data, without any thresholding

All initial observations below are based on using the default value of 0 if no data is present, unless stated otherwise.

Table 10 shows the best accuracy scores with spam thresholding for each set on FACC1, with the exact threshold in brackets. If multiple thresholds achieve the same accuracy the strictest one is chosen. We see that in nearly all cases, except for linear scaling on test data, using as much data as possible is preferable. The complete results in the appendix do show however, that it does improve on a per person basis, as the accuracy improves in several cases when people without any information are not taken into consideration when calculating the metrics. Thus, the decline in accuracy when using a stricter threshold is likely due to the more common use of the default value 0 as missing data increases. Table 12 shows the same type of information, only for FRANK1 instead of FACC1. Comparison of these two tables shows that when using the default score of 0, FRANK1 achieves higher maximum accuracies six out of eight times, albeit in some cases by margins as small as 0.006. When using a default score of 5, FRANK1 only achieves higher maximum accuracies in three situations, which is likely due to the fact that FRANK1 has more data available as seen in the initial analysis, and in turn FRANK1 has less places where the default score of 5 can be applied when using stricter thresholds.

Table 11 shows the maximum accuracies and their corresponding thresholds for inner-document thresholding with FACC1 data. Results here show that, especially for nationalities, focusing on co-occurrences occurring near each other achieves higher maximum accuracies on the train as well as test data with 0 as default value. Table 13 shows the FRANK1 results in the same context. This shows some interesting results when compared to FACC1. First of all, FRANK1 achieves higher maximum accuracies for all profession instances when using 0 as default value, also using much stricter thresholds. This is likely caused due to the sheer increased number of profession annotations in the FRANK1 data. For nationalities on the other hand FACC1 achieves higher maximum accuracies in all situations, indicating that the annotations by FRANK1 might not be as accurate.

For all observations made above, it should be noted that even though the tables show the highest accuracies achieved by certain thresholds, it is by no means a definitive answer as to which threshold is best. A lot of this depends on the actual data and on the metrics used to judge the results, as in some cases there might only be a small difference of a few hundredths behind the decimal. For example, when using log scaling on inner document thresholds with FACC1 annotations for the profession training data, the 0-MAX threshold has an accuracy of 0.528, while a threshold of 0-500 has an accuracy of 0.522, a difference of 0.004, which given the amount of data can easily be swayed one way or another by adding or removing a few lines in the training data. The same goes for the comparison of the results for FACC1 and FRANK1, for example the maximum accuracy achieved by FACC1 in Table 10 on the nationality train data with linear scaling and 0 as default value achieves a higher maximum accuracy than the same value for FRANK1 data in Table 12, even though it is only by 0.006.

All tables in this section also have one thing in common: They show how something as simple as changing the default value can have a large effect on the resulting accuracies. Whether or not this finding is generalizable is hard to say, as in this case using the value 5 makes sense due to the accuracy metric used for the challenge, which views 5 as a valid score for true scores 3 through 7. It is, however, not intuitive in real-world scenarios, as there being no evidence for a combination of a person and a profession, or nationality, whatsoever is information in and of itself, assuming that the person does appear in the data.

Table 10: Maximum accuracy scores and their threshold for spam thresholding on FACC1 data

		Profession		Nationality	
Scale	Default	Train	Test	Train	Test
Log	0	0.528 (0-100)	0.497 (10-100)	0.700 (0-100)	0.631 (0-100)
	5	0.619 (70-100)	0.674 (90-100)	0.756 (80-100)	0.717 (90-100)
Lin	0	0.522 (10-100)	0.505 (0-100)	0.600 (0-100)	0.596 (40-100)
	5	0.612 (70-100)	0.661 (40-100)	0.731 (90-100)	0.727 (90-100)

Table 11: Maximum accuracy scores and their threshold for inner-document distance thresholding on FACC1 data

		Profession		Nationality	
Scale	Default	Train	Test	Train	Test
Log	0	0.528 (0-10000)	0.515 (0-2500)	0.725 (0-100)	0.657 (0-500)
	5	0.662 (0-250)	0.674 (0-1000)	0.831 (0-10)	0.768 (0-10)
Lin	0	0.513 (0-MAX)	0.505 (0-MAX)	0.706 (0-100)	0.626 (0-1000)
	5	0.631 (0-50)	0.659 (0-2500)	0.806 (0-50)	0.727 (0-10)

Table 12: Maximum accuracy scores and their threshold for spam thresholding on FRANK1 data

		Profession		Nationality	
Scale	Default	Train	Test	Train	Test
Log	0	0.588 (40-100)	0.561 (10-100)	0.713 (20-100)	0.611 (10-100)
	5	0.625 (90-100)	0.657 (90-100)	0.769 (90-100)	0.697 (90-100)
Lin	0	0.550 (40-100)	0.532 (0-100)	0.594 (20-100)	0.601 (40-100)
	5	0.550 (90-100)	0.579 (10-100)	0.762 (90-100)	0.692 (90-100)

7.2 Co-occurrence counts and wiki occurrences

Results for this approach were generated on the test data for both profession as well as nationality data and were submitted to the challenge as the final result, which was before most of the other experiments described in this thesis were run. The approach achieved an accuracy of 0.63, an average score difference of 1.97 and a Kendall's tau of 0.35. Several additional results are available in the original paper [9].

Table 13: Maximum accuracy scores and their threshold for inner-document distance thresholding on FRANK1 data

		Profession		Nationality	
Scale	Default	Train	Test	Train	Test
Log	0	0.627 (0-100)	0.585 (0-50)	0.713 (0-500)	0.641 (0-10000)
	5	0.641 (0-50)	0.673 (0-50)	0.819 (0-10)	0.773 (0-10)
Lin	0	0.575 (0-250)	0.556 (0-50)	0.644 (0-100)	0.586 (0-10000)
	5	0.604 (0-10)	0.612 (0-10)	0.775 (0-50)	0.753 (0-10)

Unfortunately, the scores produced by this method were not very convincing, however what made it special was the way it was implemented. This approach used an implementation based on PRA, probabilistic relational algebra [17]. The implementation provided a way to graphically create strategies, of which two were made, one for the wiki occurrences and one for the FACC1 co-occurrences. These strategies would then be converted into SQL queries which were run on the MonetDB column store. This method provided an intuitive way to create implementations, however, also partly due to inexperience, generating the results was relatively slow, as also indicated by it being the fourth slowest in the challenge results, as results were generated only when they were needed. A solution to the slowness described above could have been the precomputation of scores for all combinations appearing in the knowledge bases, which would turn the actual computation of the score during the challenge into a simple lookup. It also required significant preprocessing and data selection, so while this approach works great in many applications, in this case the choice was made, also due to the availability of the entire ClueWeb12 set on an Hadoop cluster, to do further implementations with Spark and Python.

7.3 Machine Learning cross-validation

Several different types of parameters and data are compared to each other. This section will cover each of those, and each subsection will start with a small overview of the exact data parameters used. It should be noted though that the results in this section cannot be directly compared to the results shown in the sections before and after this one, as the results in this section were generated using the previously described averaged F1 scores over cross-validation using data from a single entity group. Using the profession data as example, cross-validation will split this into a train and test set, and the F1 metric will show how well the model is able to predict the actual class of a profession snippet. The assumption is made that when a model cannot accurately classify snippets that do actually appear around a profession, it also will not be able to correctly classify snippets occurring around persons.

7.3.1 Spam ranking threshold

The goal is to determine whether increasing the spam threshold, thus removing more spammy documents, increases performance (F1) during cross-validation. The thresholds that are used here for the spam rankings are 0, 50, 75 and 88. Due to the even distribution of spam documents over the scores, each step above 0 removes 50% of the most spammy documents that are still in the dataset. The expected outcome is that it

will not change much on feature vectors generated from snippets containing the entity, but that it should have a positive influence on data without the entity mention in the middle of the snippet, as the context should be more focused and relevant.

The results below were generated with the previously mentioned thresholds and sample sizes, a minimum snippet length of 256 and n-gram length 5. Table 14a shows the average weighted F1 scores for five-fold cross-validation on the profession data, while Table 14b shows the results for nationalities, both taken from FACC1. These two tables show, that whether or not the entity mention is present in the snippets the feature vectors were generated from, thresholding on spam only reduces accuracy. Although it should be mentioned that for FACC1 professions this could also be due to the smaller sample sizes as the spam threshold gets stricter.

Table 14: F1 scores averaged over 5-fold cross-validation for spam thresholding with FACC1 data

(a) Profession data

	Entity presence	
Spam threshold	With	Without
0	**0.877**	**0.564**
50	0.861	0.540
75	0.843	0.527
88	0.825	0.517

(b) Nationality data

	Entity presence	
Spam threshold	With	Without
0	**0.780**	**0.482**
50	0.775	0.478
75	0.770	0.477
88	0.766	0.479

Tables 15a and 15b tell a different story however, on the FRANK1 data, without entity mentions in the snippet the n-gram features were generated from, thresholding on spam actually increases accuracy and while it is not by large amounts, 0.004 for professions and $0,015$ for nationalities, it still appears to make a difference, especially because when entity mentions are present in the snippet, this increase does not occur, and it actually decreases. This seems to indicate that, at least for this data, thresholding on spam provides better context clues.

Table 15: F1 scores averaged over 5-fold cross-validation for spam thresholding with FRANK1 data

(a) Profession data

	Entity presence	
Spam threshold	With	Without
0	**0.913**	0.468
50	0.907	0.466
75	0.901	0.468
88	0.892	**0.472**

(b) Nationality data

	Entity presence	
Spam threshold	With	Without
0	**0.816**	0.491
50	0.811	0.498
75	0.809	0.501
88	0.803	**0.506**

7.3.2 Varying N-gram lengths

Table 16 shows the results for varying the N-grams from 4 through 7 on the FACC1 snippets, while Table 17 shows the same experiments on the FRANK1 data, both using 75 as spam threshold. What this data shows is that in many cases a lower n-gram value is preferable when using snippets without entities, while a higher n-gram value seems to work better for snippets with entities. A reasonable explanation

for this is that a higher n-gram allows logistic regression to more specifically learn the n-grams that were generated over the entity mentions in the snippet, something which cannot happen which this mention is absent in the snippets the n-grams were generated with, in which case lower n values are preferable.

Table 16: F1 scores for varying N-gram lengths with FACC1 data

(a) Profession data

	Entity presence	
N-gram value	With	Without
4	0.831	0.519
5	0.843	**0.527**
6	**0.845**	0.526
7	0.839	0.519

(b) Nationality data

	Entity presence	
N-gram value	With	Without
4	**0.770**	0.471
5	**0.770**	**0.477**
6	0.767	0.475
7	0.761	0.467

Table 17: F1 scores for varying N-gram lengths with FRANK1 data

(a) Profession data

	Entity presence	
N-gram value	With	Without
4	0.889	0.456
5	0.901	0.468
6	0.906	**0.470**
7	**0.909**	0.466

(b) Nationality data

	Entity presence	
N-gram value	With	Without
4	0.805	0.491
5	**0.809**	**0.501**
6	**0.809**	0.500
7	0.805	0.493

7.4 Machine Learning triple ranking results

The full results generated based on the provided train and test data, on profession and nationality separately, using the official metrics, can be viewed in appendix C and D. Results in this section can be compared to those in section 7.1, as they both use the same data and metrics. Table 18 shows the highest accuracies achieved for spam thresholding and varying n-grams for all combinations of entity type, entity mention (whether or not the entity was present in the data before features were generated), annotation source, and train and test set.

7.4.1 Filtering before score scaling and linear or logarithmic mapping

The columns in the results describe different variations in entity filtering and score scaling. As previously described scaling the probabilities to scores can be done on all probabilities provided by logistic regression, or any combinations of person-profession that do not occur in the knowledge base can be filtered out. Secondly there are different methods of actually scaling these probabilities to scores, the previously described linear and logarithmic mapping, which can then be truncated to a smaller interval to further increase accuracy.

An interesting observation can be made that there appears to be a relation between whether or not the probabilities are filtered on the knowledge base and which mapping is used. For most datasets, logarithmic

mapping is preferred when using all probabilities predicted by logistic regression, while linear mapping is preferred when filtering predictions on the knowledge base.

Nearly all highest accuracies shown in Table 18 use the entire collection of probabilities generated by logistic regression scaled with maplog, indicating that both for the train as well as the test set these appear to produce the best results in general.

7.4.2 Restricting the output scores

Unsurprisingly, accuracies increase across the board when the scores are restricted to an interval of $[2, 5]$, as this is in no way able to decrease the accuracy scores given the 2 score window on both sides of a prediction where are values are considered true.

7.4.3 With or without entity mention in snippets

Unfortunately the full results do not show a convincing trend for whether or not models trained on feature vectors generated using snippet data with or without entity mentions in them perform better. Some datasets barely show a difference between the two, other datasets prefer one over another, while in other cases it depends on the scaling and filtering.

However, when using the results shown in Table 18 which contains the highest possible accuracies, some observations can be made. When predicting nationalities, the highest accuracies are always achieved when not using entity mentions in the snippets used for feature generation, though the difference varies from as little as 0.013 to as large as 0.04. Predictions for professions using FACC1 also achieve higher accuracies when no entity mention is present, though in this case the difference can be as small as 0.002. For profession predictions using FRANK1 on the other hand using the entity mention does in some cases result in a higher max accuracy, though only by at most 0.012.

7.4.4 FACC1 or FRANK1

Much like the entity mentions, the results do not show a clear preference for FACC1 or FRANK1. For profession train data FACC1 achieves the highest accuracy of 0.678 and FRANK1 achieves 0.724, for the test data it is 0.725 against 0.725. For the nationality data these are 0.827 and 0.815 for the train data and 0.761 and 0.751 for the test data. What does this show, however, is that the naive annotations are able to keep up with FACC1.

7.5 Generating and comparing challenge results

Only the method combining co-occurrence counts and wiki occurrences, as seen in section 6.2, has been formally submitted to the challenge, as further research was performed after the submission deadline. This section investigates the metrics that a simple baseline and the other approaches in this paper would have produced, by utilizing the exact same data, the combination of both the profession as well as the nationality test datasets, and metrics as used in the challenge.

Specific parameters and choices will be briefly described for each approach, while doing a slightly more extensive investigation into the results of the machine learning methods.

Table 18: Maximum accuracies achieved with spam ranking thresholding and different n-grams without truncating scores. The values next to accuracy indicate either the spam threshold (Thr.) or the n value (N) for n-grams. The entity column indicates whether the entity mention was present (+) or not (-) in the data used for the feature generation as described in sections 6.3.1 and 6.3.2. It should be noted that the probability filter and scaling method are not shown as all except for three utilize all probabilities scaled with maplog, the exceptions filter the probabilities using the knowledge base and are denoted with *.

				Spam		Ngram	
Data	Annotations	Train or test	Entity	Acc	Thr.	Acc	N
Prof	FACC1	Train	+	0.676	0	0.674	4
			-	0.678	0	0.676	4
		Test	+	0.704	0	0.686	4
			-	0.725	0	0.712	4
	FRANK1	Train	+	0.678	0	0.687	4
			-	0.724	88	0.712	6
		Test	+	0.725	0	0.715	4
			-	0.713	88	0.706	5
Nat	FACC1	Train	+	0.802	50	0.802*	4
			-	0.827	75	0.827	5
		Test	+	0.731	0	0.721	4
			-	0.751	0	0.761	7
	FRANK1	Train	+	0.802*	50	0.802*	4
			-	0.815	50	0.815	5
		Test	+	0.721	75	0.736	4
			-	0.751	75	0.751	5

7.5.1 Baseline

A sample baseline is tested by predicting a score of five for each triple. This baseline achieves an accuracy of 0.721, an average score difference (ASD) of 2.070 and a Kendall's tau (tau) of 0.460.

7.5.2 Co-occurrence counts

Results based on co-occurrence counts were generated using the full amount of data, both FACC1 and FRANK1 separately, without any thresholding based on spam scores or inner-document distances. Scaling counts to scores was done using the logarithmic approach, with a default score of 5. Using FACC1 data to generate these scores results in an accuracy of 0.644, an ASD of 2.101 and a tau of 0.427, while using FRANK1 results in respective metrics of 0.631, 2.149 and 0.461.

7.5.3 Co-occurrence counts and wiki occurrences

Out of the approaches described in this section this method was the only one formally submitted, using only a subset of the FACC1 data and four as default score. More information can be found in the workshop paper [9]. It achieved an accuracy of 0.630, an ASD of 1.969 and a tau of 0.353.

7.5.4 Machine Learning with ClueWeb12 snippets

Choosing which models are used is done using the data available in appendix C and D, by selecting the models from both which generate the highest accuracy scores on the profession and nationality training data.

FACC1 The highest accuracy achieved by a model trained on FACC1 profession data without truncating scores (reducing the score range from $[0, 7]$ to $[2, 5]$) is 0.678, using 5-grams, the full range of probabilities with logarithmic scaling, no spam thresholding and no entity mentions. For nationalities the best model achieves an accuracy of 0.827, using the same model properties as the profession model, except for the spam threshold which is set to 75.

Combining these two models for the mock submission results in an accuracy of 0.730, an average score difference of 1.787 and a Kendall's tau of 0.427, putting this combination of models at rank 13 out of 22 current submissions. Table 19 shows the confusion matrices for both profession as well as nationality predictions. For professions it shows that a large number of combinations, that in reality have a true score of 0, are actually given a higher relevancy. The data does also show, however, that the diagonal is still slightly more active, albeit skewed.

The results for nationalities show that almost all predicted scores are on the higher end, as none of the combinations are ranked as zero or one, and only very few are ranked as two. This indicates that logistic regression has a hard time differentiating between different nationalities, as the probabilities must be relatively close to each other.

Table 19: Confusion matrices for predicted professions and nationalities using the described FACC1 models.

(a) Profession data

True scores \ Predicted scores	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0	3	11	11	18	12	12	1	0
1	6	3	7	15	13	5	0	0
2	5	1	12	5	17	7	2	1
3	2	5	4	7	11	9	3	2
4	6	1	4	10	11	13	6	5
5	6	0	4	7	10	16	9	5
6	3	2	3	9	15	21	9	9
7	6	1	3	7	7	25	22	48

(b) Nationality data

True scores \ Predicted scores	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0	0	0	0	0	0	3	3	2
1	0	0	1	1	1	2	6	2
2	0	0	0	0	0	5	4	1
3	0	0	1	2	1	4	8	3
4	0	0	0	0	3	11	9	4
5	0	0	1	1	2	7	10	7
6	0	0	0	3	1	8	13	13
7	0	0	0	0	3	10	17	24

Score truncation, reducing the range of scores from $[0, 7]$ to $[2, 5]$, is able to provide an immediate increase in accuracy, due to the implementation of the accuracy metric. Feedback from the organizers of the challenge has also shown that there are approaches which have used this, so for completeness' sake truncation is also applied to the scores shown above, to see what influence this has on the predicted scores.

Truncating the scores of the combined models on the test data results in an accuracy of 0.777, an average score difference of 1.835 and a Kendall's tau of 0.611. These new metrics would put us at approximately rank 7 out of 22, whereas it was rank 13 before. Table 20 shows the confusion matrix of the

truncated scores, the contents of which should come as no surprise as the columns 0, 1 and 2 and 5, 6 and 7 have been merged.

Table 20: Confusion matrices for predicted professions and nationalities using the described FACC1 models and truncating the score to a smaller interval.

(a) Profession data

True scores \ Predicted scores	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0	0	0	25	18	12	13	0	0
1	0	0	16	15	13	5	0	0
2	0	0	18	5	17	10	0	0
3	0	0	11	7	11	14	0	0
4	0	0	11	10	11	24	0	0
5	0	0	10	7	10	30	0	0
6	0	0	8	9	15	39	0	0
7	0	0	10	7	7	95	0	0

(b) Nationality data

True scores \ Predicted scores	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0	0	0	0	0	0	8	0	0
1	0	0	1	1	1	10	0	0
2	0	0	0	0	0	10	0	0
3	0	0	1	2	1	15	0	0
4	0	0	0	0	3	24	0	0
5	0	0	1	1	2	24	0	0
6	0	0	0	3	1	34	0	0
7	0	0	0	0	3	51	0	0

FRANK1 The model achieving the highest accuracy on profession data when using FRANK1 data gets an accuracy score of 0.724. This was achieved using 5-grams, the full probability range with logarithmic mapping, without any entity mentions in the training data and a spam score threshold of 88. The nationality model achieves the highest accuracy of 0.815 with the same properties as the profession model, except for the spam threshold which is lowered to 75.

Combining these two models results in an accuracy of 0.724, an average score difference of 1.810 and a Kendall's tau of 0.446, ranking this 14th out of 22. The predictions of both models are visible in Table 21. Much like the profession predictions with the FACC1 data, in this case it also ranks many of the combinations with a true score of zero higher. The predicted score of 5 also seems to be the most common prediction, which is fortunate as this covers true scores of three through seven.

The predictions for nationality all appear at the higher end of the scoring interval, mostly covering the scores 5, 6 and 7. This seems to indicate that, just like for FACC1, logistic regression for FRANK1 nationalities also is not able to find large differences between different nationalities for persons.

Truncating the previous scores results in an accuracy of 0.772, an average score difference of 1.868 and a Kendall's tau of 0.618, increasing the untruncated rank from 14 to 7. The truncated scores are visible in Table 22.

7.5.5 Significance testing

This section provides a brief overview of the previously mentioned accuracies and their significance levels when compared to the baseline, seen in Table 23. Significance levels are calculated using Wilcoxon signed-rank test, where each approach is tested against the baseline, using the score differences between the predictions and the true values.

These results show that all approaches are significantly different from solely recommending a score of 5. It should be noted however that a significant difference does not also equate to a higher accuracy.

Table 21: Confusion matrices for predicted professions and nationalities using the described FRANK1 models.

(a) Profession data

True scores \ Predicted scores	0	1	2	3	4	5	6	7
0	7	1	13	13	15	15	4	0
1	5	3	5	8	11	12	3	2
2	3	2	7	7	17	11	2	1
3	1	1	5	7	9	13	6	1
4	5	0	4	9	11	15	6	6
5	1	1	4	5	14	21	7	4
6	1	1	3	5	13	25	10	13
7	3	0	3	5	12	30	24	42

(b) Nationality data

True scores \ Predicted scores	0	1	2	3	4	5	6	7
0	0	0	0	0	1	4	2	1
1	0	0	0	1	4	2	4	2
2	0	0	0	0	1	4	4	1
3	0	0	0	1	3	4	4	7
4	0	0	0	0	1	9	13	4
5	0	0	0	2	3	4	12	7
6	0	0	0	1	4	8	13	12
7	0	0	0	2	2	9	18	23

Table 22: Confusion matrices for predicted professions and nationalities using the described FRANK1 models and truncating the score to a smaller interval.

(a) Profession data

True scores \ Predicted scores	0	1	2	3	4	5	6	7
0	0	0	21	13	15	19	0	0
1	0	0	13	8	11	17	0	0
2	0	0	12	7	17	14	0	0
3	0	0	7	7	9	20	0	0
4	0	0	9	9	11	27	0	0
5	0	0	6	5	14	32	0	0
6	0	0	5	5	13	48	0	0
7	0	0	6	5	12	96	0	0

(b) Nationality data

True scores \ Predicted scores	0	1	2	3	4	5	6	7
0	0	0	0	0	1	7	0	0
1	0	0	0	1	4	8	0	0
2	0	0	0	0	1	9	0	0
3	0	0	0	1	3	15	0	0
4	0	0	0	0	1	26	0	0
5	0	0	0	2	3	23	0	0
6	0	0	0	1	4	33	0	0
7	0	0	0	2	2	50	0	0

Table 23: Accuracies for the approaches described in this section together with an indicator whether or not there was a significant difference, in terms of score differences to the true scores, between the baseline and the approach (* $p \leq 0.05$, ** $p \leq 0.01$ and *** $p \leq 0.001$).

Approach	Accuracy
Baseline	0.721
Co-occurrences with FACC1	0.644***
Co-occurrences with FRANK1	0.631***
Co-occurrences with wiki occurrences	0.630***
Machine Learning with FACC1 snippets	0.730***
Machine Learning with FRANK1 snippets	0.724*

8 Conclusion

Two research goals were stated in the introduction, the first of which was whether or not web data is a valuable source of information for entity ranking. As shown in this thesis, web information can be a valuable source of contextual information, as one is able to get information from a large variety of different

webpages. It should be noted though, that there are still several challenges one has to tackle before a source can be used, for example finding the actual locations of entities within webpages, and being able to neatly extract these, removing any unwanted data like HTML or CSS structures.

Comparisons between sections 7.1 and 7.4 show how machine learning is nearly always able to achieve higher accuracies on the provided train and test datasets, with one exception for nationality, where an almost complete prediction consisting of the default score 5 appears to work exceptionally well.

Results in section 7.5 show what kind of an improvement machine learning is able to achieve over the initially submitted results consisting of FACC1 co-occurrence counts and Wiki abstract occurrences of entities as seen in section 7.2, increasing the initial submission with an accuracy of 0.63 to 0.73, or even 0.78 when scores are truncated to a smaller range.

The second research question was whether or not naive annotations of web data are able to compete with FACC1 annotations, and is it turns out, they likely can. There appears to only be a minimal difference in the mock submissions, as seen in the results in section 7.5, where using FRANK1 achieves an accuracy of 0.724 as opposed to 0.730. Full results on train and test data do however show that whether FACC1 or FRANK1 is better depends on the data being predicted. Significance levels also shown that FRANK1 is less significantly different from a baseline of five as opposed to other approaches. An important aspect to keep in mind however is that FRANK1 is, and currently can not be, a full replacement of FACC1, as FRANK1 does not contain the probabilities indicating whether or not the entity is tagged correctly, both on the text occurrence and the context as well as solely on the context.

It should be noted though that the machine learning results can vary a lot, as all different versions of the machine learning approach use sampled data which is often limited to 10,000 snippets, which in some cases is sampled from values ranging in the millions. Furthermore, while these annotations can easily be computed for other webcrawl datasets, it would also still need the spam rankings if the exact approaches in this paper were to be recreated. Fortunately though, the method that was used to generate the spam rankings is available.

For machine learning both the features, as well as handling of HTML data in ClueWeb documents, are still rather naive. Approaches using more advanced features, like using word embeddings [25], might actually perform better on the data as it looks at the meaning of the words, rather than just the characters that occur, especially assuming one is able leave out all HTML data.

HTML data on the other hand could also be taken into consideration when learning or predicting, for example placing higher emphasis on snippets which occur in header tags. Another interesting source of data to judge the quality of webpages, much like the spam rankings, would be PageRank scores.

In the data section it was stated that we opted not to use the provided annotated wiki-sentences dataset and to investigate a more generalizable approach. Unsurprisingly, this choice has likely been detrimental to the results with respect to the actual challenge. A brief review of the top three approaches with the highest accuries show that all of investigate the use of this data, often combined with additional data, varying from path data [8] or entity features [5] in Freebase, or textual information from full Wikipedia articles [26].

References

[1] R. Agrawal, A. Gupta, Y. Prabhu, and M. Varma. Multi-label learning with millions of labels. In *Proceedings of the 22nd international conference on World Wide Web - WWW '13*, pages 13–24, 2013. ISBN 9781450320351. doi: 10.1145/2488388.2488391. URL http://dl.acm.org/citation.cfm?id=2488391{%}0Ahttp://dl.acm.org/citation.cfm?doid=2488388.2488391.

[2] A. V. Aho and M. J. Corasick. Efficient string matching: an aid to bibliographic search. *Communications of the ACM*, 18(6):333–340, 1975. ISSN 00010782. doi: 10.1145/360825.360855. URL http://portal.acm.org/citation.cfm?doid=360825.360855.

[3] H. Bast, B. Buchhold, and E. Haussmann. Relevance scores for triples from type-like relations. In *SIGIR 2015 - Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 243–252, 2015. ISBN 9781450336215. doi: 10.1145/2766462.2767734. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-84953790766{&}partnerID=40{&}md5=036501b3e5be58cf533fdf91b26d80cf.

[4] H. Bast, B. Buchhold, and E. Haussmann. Overview of the Triple Scoring Task at WSDM Cup 2017. In *Proceedings of the 2nd WSDM Cup at the ACM WSDM Conference on Web Search and Data Mining (WSDM Cup 17)*, 2017.

[5] L.-W. Chen, B. Mangipudi, J. Bandlamudi, R. Sehgal, Y. Hao, M. Jiang, and H. Gui. Integrating Knowledge from Latent and Explicit Features for Triple Scoring—Team Radicchio's Triple Scorer at WSDM Cup 2017. In M. Potthast, S. Heindorf, and H. Bast, editors, *WSDM Cup 2017 Notebook Papers, February 10, Cambridge, UK*. CEUR-WS.org, 2017. URL http://www.wsdm-cup-2017.org/proceedings.html.

[6] G. V. Cormack, M. D. Smucker, and C. L. A. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *Information Retrieval*, 14(5):441–465, 2011. ISSN 13864564. doi: 10.1007/s10791-011-9162-z.

[7] M. Craven and J. Kumlien. Constructing biological knowledge bases by extracting information from text sources. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology (ISMB-99)*, pages 77–86, 1999. ISSN 1553-0833. doi: 090. URL http://www.aaai.org/Papers/ISMB/1999/ISMB99-010.pdf.

[8] B. Ding, Q. Wang, and B. Wang. Leveraging Text and Knowledge Bases for Triple Scoring: An Ensemble Approach—The BOKCHOY Triple Scorer at WSDM Cup 2017. In M. Potthast, S. Heindorf, and H. Bast, editors, *WSDM Cup 2017 Notebook Papers, February 10, Cambridge, UK*. CEUR-WS.org, 2017. URL http://www.wsdm-cup-2017.org/proceedings.html.

[9] F. Dorssers, A. P. de Vries, W. Alink, and R. Cornacchia. Ranking Triples using Entity Links in a Large Web Crawl—The Chicory Triple Scorer at WSDM Cup 2017. In M. Potthast, S. Heindorf, and H. Bast, editors, *WSDM Cup 2017 Notebook Papers, February 10, Cambridge, UK*. CEUR-WS.org, 2017. URL http://www.wsdm-cup-2017.org/proceedings.html.

[10] R. Fagin, R. Kumar, M. Mahdian, D. Sivakumar, and E. Vee. Comparing and aggregating rankings with ties. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems - PODS '04*, page 47, 2004. ISBN 158113858X. doi: 10.1145/1055558.1055568. URL http://portal.acm.org/citation.cfm?doid=1055558.1055568.

[11] E. Gabrilovich, M. Ringgaard, and A. Subramanya. FACC1: Freebase annotation of ClueWeb corpora, Version 1 (Release date 2013-06-26, Format version 1, Correction level 0). *Note: http://lemurproject.org/clueweb12/*, 5, 2013.

[12] S. Heindorf, M. Potthast, H. Bast, B. Buchhold, and E. Haussmann. WSDM Cup 2017: Vandalism Detection and Triple Scoring. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM 17)*. ACM, 2017. doi: http://dx.doi.org/10.1145/3018661.3022762.

[13] X. Kong, X. Shi, and P. S. Yu. Multi-Label Collective Classification. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, pages 618–629. 2011. ISBN 9780898719925. doi: 10.1137/1.9781611972818.53. URL http://epubs.siam.org/doi/abs/10.1137/1.9781611972818.53.

[14] X. Ling and D. Weld. Fine-Grained Entity Recognition. *Aaai*, pages 94–100, 2012. doi: 10.1.1.431.8777. URL http://www.aaai.org/ocs/index.php/aaai/aaai12/paper/download/5152/5124.

[15] P. Mcnamee and C. K. Nicholas. Textual Representations for Corpus-Based Bilingual Retrieval. 2008.

[16] M. Newman. Power laws, Pareto distributions and Zipf's law. *Power laws, Pareto distributions and Zipf's law. Contemporary physics*, 46(5):323–351, 2005. ISSN 0010-7514. doi: 10.1016/j.cities.2012.03.001. URL http://arxiv.org/abs/cond-mat/0412004{%}5Cnhttp://dx.doi.org/10.1016/j.cities.2012.03.001.

[17] Norbert Fuhr; Thomas Rölleke. A probabilistic relational algebra for the integration of information retrieval and database systems. *ACM Transactions on Information Systems*, 15(1):32–66, 1997. ISSN 10468188. doi: 10.1145/239041.239045.

[18] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. *World Wide Web Internet And Web Information Systems*, 54(1999-66):1–17, 1998. ISSN 1752-0509. doi: 10.1.1.31.1768. URL http://ilpubs.stanford.edu:8090/422.

[19] Y. Prabhu and M. Varma. FastXML: a fast, accurate and stable tree-classifier for extreme multi-label learning. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14*, pages 263–272, 2014. doi: 10.1145/2623330.2623651. URL http://doi.acm.org/10.1145/2623330.2623651.

[20] M. Schuhmacher, L. Dietz, and S. Paolo Ponzetto. Ranking Entities for Web Queries Through Text and Knowledge. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1461–1470, 2015. ISBN 978-1-4503-3794-6. doi: 10.1145/2806416.2806480. URL http://doi.acm.org/10.1145/2806416.2806480.

[21] A. Tonon, M. Catasta, G. Demartini, P. Cudré-Mauroux, and K. Aberer. TRank: Ranking entity types using the web of data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8218 LNCS, pages 640–656, 2013. ISBN 9783642413346. doi: 10.1007/978-3-642-41335-3_40.

[22] A.-M. Vercoustre, J. A. Thom, and J. Pehcevski. Entity Ranking in Wikipedia. *Knowledge Creation Diffusion Utilization*, page 1101, 2007. doi: 10.1145/1363686.1363943. URL http://arxiv.org/abs/0711.3128.

[23] K. Weinberger, A. Dasgupta, J. Attenberg, J. Langford, and A. Smola. Feature Hashing for Large Scale Multitask Learning. 2009. ISSN 1605585165. doi: 10.1145/1553374.1553516. URL http://arxiv.org/abs/0902.2206.

[24] J. Weston, a. Makadia, and H. Yee. Label partitioning for sublinear ranking. *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 28:181–189, 2013. URL http://machinelearning.wustl.edu/mlpapers/papers/weston13.

[25] Y. Yaghoobzadeh and H. Schütze. Corpus-level Fine-grained Entity Typing Using Contextual Information. 2016. URL http://arxiv.org/abs/1606.07901.

[26] I. Yamada, M. Sato, and H. Shindo. Ensemble of Neural Classifiers for Scoring Knowledge Base Triples—The Lettuce Triple Scorer at WSDM Cup 2017. In M. Potthast, S. Heindorf, and H. Bast, editors, *WSDM Cup 2017 Notebook Papers, February 10, Cambridge, UK*. CEUR-WS.org, 2017. URL http://www.wsdm-cup-2017.org/proceedings.html.

[27] D. Yogatama, D. Gillick, and N. Lazic. Embedding methods for fine grained entity type classification. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP '15)*, pages 291–296, 2015. URL http://www.aclweb.org/anthology/P15-2048.

Appendices

A Co-occurrence scores with FACC1

A.1 Spam ranking thresholds

A.1.1 Log scaling

Table 24: Results for the profession training data when varying the spam ranking cutoff with log scaling

	Full data			Missing persons removed				
	Acc	Asd	Tau	Acc	Asd	Tau	Uniq	Ents
90-100	0.447	3.270	0.456	0.523	2.769	0.447	72	350
80-100	0.483	3.049	0.435	0.548	2.625	0.419	85	389
70-100	0.505	2.940	0.423	0.572	2.514	0.405	88	395
60-100	0.501	2.882	0.419	0.564	2.485	0.397	92	404
50-100	0.513	2.806	0.416	0.566	2.474	0.398	97	424
40-100	0.515	2.808	0.412	0.568	2.481	0.393	98	426
30-100	0.518	2.796	0.413	0.573	2.467	0.394	98	426
20-100	0.524	2.777	0.416	0.580	2.450	0.399	99	429
10-100	0.526	**2.769**	0.411	0.583	**2.441**	0.393	99	429
0-100	**0.528**	**2.769**	0.412	**0.585**	**2.441**	0.394	99	429

Table 25: Results for the profession test data when varying the spam ranking cutoff with log scaling

	Full data			Missing persons removed				
	Acc	Asd	Tau	Acc	Asd	Tau	Uniq	Ents
90-100	0.464	3.211	0.419	0.599	2.446	0.378	70	332
80-100	0.474	3.039	0.393	0.586	2.374	0.349	82	374
70-100	0.493	2.975	0.369	**0.608**	**2.319**	0.315	85	383
60-100	0.489	2.955	0.377	0.597	2.328	0.332	87	387
50-100	0.493	2.949	0.377	0.597	2.349	0.334	88	390
40-100	0.489	2.943	0.385	0.586	2.397	0.351	92	401
30-100	0.483	2.922	0.379	0.577	2.379	0.345	93	404
20-100	0.487	2.914	0.378	0.579	2.384	0.345	94	406
10-100	**0.497**	2.848	0.363	0.575	2.377	0.328	98	419
0-100	**0.497**	**2.838**	0.358	0.572	2.378	0.323	99	421

Table 26: Results for the nationality train data when varying the spam ranking cutoff with log scaling

	Full data			Missing persons removed				
	Acc	Asd	Tau	Acc	Asd	Tau	Uniq	Ents
90-100	0.550	2.663	0.408	0.667	1.944	0.407	59	126
80-100	0.637	2.294	0.442	0.697	1.944	0.434	67	142
70-100	0.650	2.112	0.410	0.699	1.822	0.401	69	146
60-100	0.669	2.075	0.416	0.700	1.860	0.410	71	150
50-100	0.662	2.087	0.449	0.693	1.873	0.445	71	150
40-100	0.669	2.069	0.435	0.700	1.853	0.430	71	150
30-100	0.669	2.087	0.457	0.700	1.873	0.454	71	150
20-100	0.694	1.962	0.427	0.717	1.803	0.423	72	152
10-100	0.688	1.962	0.433	0.711	1.803	0.430	72	152
0-100	**0.700**	**1.894**	0.420	**0.721**	**1.760**	0.417	73	154

Table 27: Results for the nationality test data when varying the spam ranking cutoff with log scaling

	Full data			Missing persons removed				
	Acc	Asd	Tau	Acc	Asd	Tau	Uniq	Ents
90-100	0.530	2.687	0.479	0.628	2.238	0.487	79	164
80-100	0.571	2.444	0.491	0.617	2.222	0.490	87	180
70-100	0.581	2.369	0.481	0.621	2.165	0.479	88	182
60-100	0.596	2.283	0.510	0.630	2.098	0.510	89	184
50-100	0.611	2.202	0.504	**0.640**	2.038	0.505	90	186
40-100	0.601	2.217	0.510	0.622	2.085	0.510	91	188
30-100	0.606	2.187	0.520	0.626	2.074	0.521	92	190
20-100	0.601	2.192	0.520	0.615	2.109	0.521	93	192
10-100	0.616	2.152	0.520	0.630	2.068	0.521	93	192
0-100	**0.631**	**2.081**	0.510	0.639	**2.026**	0.510	94	194

A.1.2 Lin scaling

Table 28: Results for the profession training data when varying the spam ranking cutoff with linear scaling

	Full data			Missing persons removed				
	Acc	Asd	Tau	Acc	Asd	Tau	Uniq	Ents
90-100	0.437	3.307	0.458	0.509	2.823	0.450	72	350
80-100	0.472	3.155	0.444	0.532	2.766	0.435	85	389
70-100	0.493	3.039	0.429	0.557	2.643	0.414	88	395
60-100	0.501	2.977	0.417	0.564	2.606	0.394	92	404
50-100	0.518	2.911	0.418	0.573	2.601	0.400	97	424
40-100	0.518	2.920	0.417	0.573	2.617	0.399	98	426
30-100	0.520	2.897	0.412	0.575	2.589	0.394	98	426
20-100	0.518	2.901	0.418	0.573	2.599	0.403	99	429
10-100	**0.522**	**2.878**	0.416	**0.578**	**2.571**	0.399	99	429
0-100	0.513	2.905	0.413	0.566	2.604	0.395	99	429

Table 29: Results for the profession test data when varying the spam ranking cutoff with linear scaling

	Full data			Missing persons removed				
	Acc	Asd	Tau	Acc	Asd	Tau	Uniq	Ents
90-100	0.425	3.355	0.412	0.539	2.669	0.363	70	332
80-100	0.456	3.181	0.383	0.561	2.570	0.331	82	374
70-100	0.472	3.111	0.363	0.580	2.501	0.305	85	383
60-100	0.483	3.035	0.381	0.589	2.434	0.337	87	387
50-100	0.487	3.019	0.373	**0.590**	2.441	0.327	88	390
40-100	0.485	3.008	0.375	0.581	2.479	0.338	92	401
30-100	0.485	2.996	0.381	0.579	2.473	0.347	93	404
20-100	0.487	2.963	0.374	0.579	2.446	0.339	94	406
10-100	0.501	2.895	0.366	0.580	2.434	0.332	98	419
0-100	**0.505**	**2.877**	0.361	0.582	**2.425**	0.327	99	421

Table 30: Results for the nationality train data when varying the spam ranking cutoff with linear scaling

	Full data			Missing persons removed				
	Acc	Asd	Tau	Acc	Asd	Tau	Uniq	Ents
90-100	0.506	2.906	0.434	0.611	2.254	0.441	59	126
80-100	0.550	2.550	0.439	0.599	2.232	0.431	67	142
70-100	0.581	2.337	0.428	**0.623**	**2.068**	0.421	69	146
60-100	0.581	2.337	0.421	0.607	2.140	0.416	71	150
50-100	0.588	2.331	0.432	0.613	2.133	0.428	71	150
40-100	0.594	2.344	0.445	0.620	2.147	0.442	71	150
30-100	0.575	2.388	0.472	0.600	2.193	0.470	71	150
20-100	0.594	2.275	0.432	0.612	2.132	0.429	72	152
10-100	0.594	2.263	0.432	0.612	2.118	0.429	72	152
0-100	**0.600**	**2.250**	0.437	0.617	2.130	0.434	73	154

Table 31: Results for the nationality test data when varying the spam ranking cutoff with linear scaling

	Full data			Missing persons removed				
	Acc	Asd	Tau	Acc	Asd	Tau	Uniq	Ents
90-100	0.500	2.838	0.469	0.591	2.421	0.474	79	164
80-100	0.556	2.596	0.465	0.600	2.389	0.462	87	180
70-100	0.561	2.510	0.468	0.599	2.319	0.465	88	182
60-100	0.571	2.475	0.489	0.603	2.304	0.488	89	184
50-100	0.591	2.394	0.503	**0.618**	**2.242**	0.503	90	186
40-100	**0.596**	2.384	0.505	0.617	2.261	0.505	91	188
30-100	0.591	2.394	0.505	0.611	2.289	0.505	92	190
20-100	0.581	2.424	0.510	0.594	2.349	0.510	93	192
10-100	**0.596**	2.384	0.506	0.609	2.307	0.506	93	192
0-100	0.581	**2.338**	0.501	0.588	2.289	0.501	94	194

A.1.3 Default score 5

Table 32: Results for the profession train data when varying the spam ranking cutoff with zero scores replaced with five

	Log scaling			Lin scaling		
	Acc	Asd	Tau	Acc	Asd	Tau
90-100	0.614	2.315	0.461	0.594	2.334	0.470
80-100	0.608	2.318	0.443	0.604	2.301	0.464
70-100	**0.619**	**2.250**	0.429	**0.612**	**2.256**	0.449
60-100	0.600	2.274	0.427	0.602	2.262	0.440
50-100	0.596	2.280	0.427	0.606	2.276	0.442
40-100	0.604	2.268	0.428	0.604	2.280	0.444
30-100	0.602	2.282	0.424	0.606	2.260	0.434
20-100	0.606	2.280	0.426	0.592	2.315	0.442
10-100	0.608	2.276	0.420	0.592	2.311	0.438
0-100	0.608	2.278	0.418	0.579	2.346	0.435

Table 33: Results for the profession test data when varying the spam ranking cutoff with zero scores replaced with five

	Log scaling			Lin scaling		
	Acc	Asd	Tau	Acc	Asd	Tau
90-100	**0.674**	2.146	0.424	0.637	2.250	0.431
80-100	0.647	2.175	0.412	0.634	2.251	0.417
70-100	0.665	2.101	0.375	0.649	2.181	0.397
60-100	0.661	2.082	0.378	0.655	2.119	0.405
50-100	0.673	**2.068**	0.374	0.659	2.119	0.403
40-100	0.663	2.099	0.382	**0.661**	2.107	0.403
30-100	0.655	2.092	0.379	0.655	**2.105**	0.405
20-100	0.657	2.097	0.377	0.647	2.125	0.399
10-100	0.645	2.131	0.371	0.651	2.113	0.397
0-100	0.639	2.150	0.368	0.645	2.121	0.393

Table 34: Results for the nationality train data when varying the spam ranking cutoff with zero scores replaced with five

	Log scaling			Lin scaling		
	Acc	Asd	Tau	Acc	Asd	Tau
90-100	0.750	**1.712**	0.408	**0.731**	1.869	0.434
80-100	**0.756**	1.725	0.442	0.700	1.900	0.434
70-100	0.738	**1.712**	0.410	0.700	**1.831**	0.429
60-100	0.738	1.731	0.416	0.675	1.906	0.417
50-100	0.731	1.744	0.449	0.681	1.900	0.433
40-100	0.731	1.756	0.435	0.688	1.913	0.446
30-100	0.725	1.806	0.457	0.669	1.956	0.473
20-100	0.731	1.775	0.427	0.675	1.906	0.428
10-100	0.725	1.775	0.433	0.662	1.944	0.428
0-100	0.731	1.738	0.420	0.669	1.931	0.432

Table 35: Results for the nationality test data when varying the spam ranking cutoff with zero scores replaced with five

	Log scaling			Lin scaling		
	Acc	Asd	Tau	Acc	Asd	Tau
90-100	**0.717**	**1.934**	0.479	**0.727**	**1.904**	0.469
80-100	0.682	1.970	0.491	0.697	1.960	0.462
70-100	0.672	1.975	0.481	0.692	1.934	0.464
60-100	0.672	1.965	0.510	0.677	2.015	0.486
50-100	0.672	1.960	0.504	0.677	2.056	0.499
40-100	0.657	1.990	0.510	0.682	2.035	0.501
30-100	0.662	1.970	0.520	0.677	2.056	0.501
20-100	0.652	1.990	0.520	0.652	2.141	0.507
10-100	0.657	1.980	0.520	0.657	2.131	0.503
0-100	0.657	1.965	0.510	0.631	2.136	0.498

A.2 Inner document thresholds

A.2.1 Log scaling

Table 36: Results for the profession training data when varying the co-occurence distance cutoff with log scaling

	Full data			Missing persons removed				
	Acc	Asd	Tau	Acc	Asd	Tau	Uniq	Ents
0-10	0.416	3.478	0.428	0.555	2.632	0.351	47	247
0-50	0.445	3.262	0.424	0.546	2.644	0.384	68	317
0-100	0.454	3.186	0.432	0.551	2.594	0.408	75	345
0-250	0.520	2.981	0.405	**0.620**	2.456	0.374	85	379
0-500	0.522	2.909	0.409	0.618	2.395	0.376	86	382
0-1000	0.522	2.849	0.403	0.603	**2.387**	0.370	88	393
0-2500	0.526	2.810	0.419	0.592	2.408	0.397	92	412
0-5000	0.518	2.827	0.435	0.579	2.454	0.421	94	416
0-10000	**0.528**	2.781	0.416	0.593	2.390	0.396	95	418
0-MAX	**0.528**	**2.769**	0.412	0.585	2.441	0.394	99	429

Table 37: Results for the profession test data when varying the co-occurence distance cutoff with log scaling

	Full data			Missing persons removed				
	Acc	Asd	Tau	Acc	Asd	Tau	Uniq	Ents
0-10	0.421	3.400	0.417	0.582	2.456	0.337	54	287
0-50	0.478	3.088	0.387	0.618	2.263	0.312	70	346
0-100	0.472	3.072	0.394	0.609	2.260	0.329	72	350
0-250	0.481	2.992	0.379	0.610	2.242	0.315	78	364
0-500	0.485	3.012	0.382	0.606	2.298	0.325	80	376
0-1000	0.509	2.903	0.359	**0.623**	2.243	0.299	85	387
0-2500	**0.515**	2.842	0.350	0.622	**2.209**	0.293	89	397
0-5000	0.503	2.856	0.349	0.600	2.266	0.298	92	403
0-10000	0.495	2.867	0.351	0.584	2.328	0.304	94	409
0-MAX	0.497	**2.838**	0.358	0.572	2.378	0.323	99	421

Table 38: Results for the nationality training data when varying the co-occurence distance cutoff with log scaling

	Full data			Missing persons removed				
	Acc	Asd	Tau	Acc	Asd	Tau	Uniq	Ents
0-10	0.619	2.375	0.280	0.742	1.648	0.238	60	128
0-50	0.669	2.006	0.308	0.725	1.620	0.289	67	142
0-100	**0.725**	1.887	0.299	**0.767**	**1.603**	0.278	69	146
0-250	0.700	1.913	0.316	0.727	1.727	0.303	71	150
0-500	0.706	1.913	0.336	0.733	1.727	0.325	71	150
0-1000	0.713	1.894	0.336	0.740	1.707	0.325	71	150
0-2500	0.700	1.900	0.374	0.724	1.737	0.367	72	152
0-5000	0.706	1.906	0.360	0.730	1.743	0.352	72	152
0-10000	0.706	**1.869**	0.367	0.727	1.734	0.361	73	154
0-MAX	0.700	1.894	0.420	0.721	1.760	0.417	73	154

Table 39: Results for the nationality test data when varying the co-occurence distance cutoff with log scaling

	Full data			Missing persons removed				
	Acc	Asd	Tau	Acc	Asd	Tau	Uniq	Ents
0-10	0.571	2.495	0.384	**0.701**	1.844	0.370	74	154
0-50	0.626	2.192	0.384	0.695	**1.833**	0.386	84	174
0-100	0.626	2.157	0.403	0.663	1.973	0.401	89	184
0-250	0.652	2.096	0.435	0.676	1.968	0.431	91	188
0-500	**0.657**	2.061	0.440	0.681	1.931	0.437	91	188
0-1000	0.646	2.040	0.428	0.663	1.937	0.425	92	190
0-2500	0.646	2.061	0.441	0.656	1.984	0.439	93	192
0-5000	0.646	**2.035**	0.445	0.656	1.958	0.444	93	192
0-10000	0.631	2.045	0.447	0.641	1.969	0.445	93	192
0-MAX	0.631	2.081	0.510	0.639	2.026	0.510	94	194

A.2.2 Lin scaling

Table 40: Results for the profession training data when varying the co-occurence distance cutoff with linear scaling

	Full data			Missing persons removed				
	Acc	Asd	Tau	Acc	Asd	Tau	Uniq	Ents
0-10	0.417	3.483	0.428	0.559	2.644	0.350	47	247
0-50	0.449	3.324	0.425	0.552	2.744	0.386	68	317
0-100	0.454	3.274	0.430	0.551	2.725	0.404	75	345
0-250	0.491	3.060	0.406	0.580	2.565	0.375	85	379
0-500	0.499	3.000	0.411	**0.586**	**2.518**	0.379	86	382
0-1000	0.497	2.984	0.411	0.570	2.565	0.382	88	393
0-2500	0.509	2.961	0.419	0.570	2.597	0.397	92	412
0-5000	0.501	2.971	0.431	0.558	2.632	0.416	94	416
0-10000	0.507	2.946	0.430	0.567	2.593	0.415	95	418
0-MAX	**0.513**	**2.905**	0.413	0.566	2.604	0.395	99	429

Table 41: Results for the profession test data when varying the co-occurence distance cutoff with linear scaling

	Full data			Missing persons removed				
	Acc	Asd	Tau	Acc	Asd	Tau	Uniq	Ents
0-10	0.413	3.460	0.416	0.568	2.564	0.334	54	287
0-50	0.460	3.209	0.384	0.592	2.442	0.306	70	346
0-100	0.450	3.207	0.390	0.577	2.457	0.322	72	350
0-250	0.462	3.127	0.373	0.582	2.431	0.306	78	364
0-500	0.448	3.189	0.380	0.556	2.540	0.322	80	376
0-1000	0.470	3.078	0.360	0.571	2.475	0.301	85	387
0-2500	0.487	2.971	0.354	0.587	**2.375**	0.298	89	397
0-5000	0.497	2.951	0.350	**0.593**	2.387	0.299	92	403
0-10000	0.501	2.920	0.349	0.592	2.394	0.301	94	409
0-MAX	**0.505**	**2.877**	0.361	0.582	2.425	0.327	99	421

Table 42: Results for the nationality training data when varying the co-occurence distance cutoff with linear scaling

	Full data			Missing persons removed				
	Acc	Asd	Tau	Acc	Asd	Tau	Uniq	Ents
0-10	0.594	2.450	0.295	0.711	1.742	0.257	60	128
0-50	0.700	1.994	0.313	**0.761**	**1.606**	0.295	67	142
0-100	**0.706**	**1.906**	0.324	0.747	1.623	0.307	69	146
0-250	**0.706**	1.956	0.369	0.733	1.773	0.359	71	150
0-500	0.669	2.087	0.395	0.693	1.913	0.388	71	150
0-1000	0.656	2.050	0.386	0.680	1.873	0.378	71	150
0-2500	0.637	2.087	0.388	0.658	1.934	0.382	72	152
0-5000	0.625	2.138	0.404	0.645	1.987	0.398	72	152
0-10000	0.619	2.119	0.410	0.636	1.994	0.407	73	154
0-MAX	0.600	2.250	0.437	0.617	2.130	0.434	73	154

Table 43: Results for the nationality test data when varying the co-occurence distance cutoff with linear scaling

	Full data			Missing persons removed				
	Acc	Asd	Tau	Acc	Asd	Tau	Uniq	Ents
0-10	0.535	2.566	0.396	**0.656**	**1.935**	0.386	74	154
0-50	0.591	2.288	0.374	0.655	1.943	0.374	84	174
0-100	0.606	2.227	0.438	0.641	2.049	0.439	89	184
0-250	0.621	**2.167**	0.433	0.644	2.043	0.429	91	188
0-500	0.621	2.182	0.464	0.644	2.059	0.462	91	188
0-1000	**0.626**	2.182	0.438	0.642	2.084	0.435	92	190
0-2500	0.611	2.217	0.454	0.620	2.146	0.452	93	192
0-5000	0.611	2.207	0.456	0.620	2.135	0.454	93	192
0-10000	0.621	2.207	0.470	0.630	2.135	0.469	93	192
0-MAX	0.581	2.338	0.501	0.588	2.289	0.501	94	194

A.2.3 Default score 5

Table 44: Results for the profession train data when varying the co-occurence distance cutoff with zero scores replaced with five

	Log scaling			Lin scaling		
	Acc	Asd	Tau	Acc	Asd	Tau
0-10	0.625	2.239	0.447	0.625	2.247	0.451
0-50	0.621	2.254	0.445	**0.631**	2.241	0.436
0-100	0.629	2.204	0.440	**0.631**	2.231	0.435
0-250	**0.662**	2.140	0.408	0.625	**2.212**	0.423
0-500	0.650	2.140	0.419	0.614	2.241	0.436
0-1000	0.647	**2.115**	0.409	0.600	2.272	0.432
0-2500	0.637	2.159	0.426	0.592	2.315	0.448
0-5000	0.627	2.183	0.430	0.577	2.328	0.454
0-10000	0.623	2.212	0.428	0.571	2.365	0.461
0-MAX	0.608	2.278	0.418	0.579	2.346	0.435

Table 45: Results for the profession test data when varying the co-occurence distance cutoff with zero scores replaced with five

	Log scaling			Lin scaling		
	Acc	Asd	Tau	Acc	Asd	Tau
0-10	0.645	2.228	0.433	0.645	2.234	0.428
0-50	0.645	2.168	0.417	0.639	2.230	0.411
0-100	0.643	2.148	0.420	0.635	2.211	0.417
0-250	0.651	2.094	0.392	0.651	2.154	0.404
0-500	0.655	2.121	0.399	0.637	2.212	0.407
0-1000	**0.674**	2.062	0.378	0.647	2.175	0.397
0-2500	**0.674**	**2.047**	0.363	**0.659**	**2.086**	0.376
0-5000	0.661	2.078	0.361	0.655	2.099	0.376
0-10000	0.657	2.090	0.364	0.653	2.094	0.377
0-MAX	0.639	2.150	0.368	0.645	2.121	0.393

Table 46: Results for the nationality train data when varying the co-occurence distance cutoff with zero scores replaced with five

	Log scaling			Lin scaling		
	Acc	Asd	Tau	Acc	Asd	Tau
0-10	**0.831**	**1.400**	0.276	0.794	1.488	0.296
0-50	0.781	1.481	0.308	**0.806**	**1.450**	0.317
0-100	0.794	1.506	0.299	0.762	1.562	0.329
0-250	0.769	1.544	0.316	0.775	1.600	0.373
0-500	0.775	1.556	0.336	0.750	1.706	0.399
0-1000	0.769	1.600	0.336	0.731	1.712	0.391
0-2500	0.750	1.650	0.374	0.706	1.794	0.393
0-5000	0.756	1.656	0.360	0.700	1.800	0.399
0-10000	0.738	1.712	0.367	0.675	1.863	0.406
0-MAX	0.731	1.738	0.420	0.669	1.931	0.432

Table 47: Results for the nationality test data when varying the co-occurence distance cutoff with zero scores replaced with five

	Log scaling			Lin scaling		
	Acc	Asd	Tau	Acc	Asd	Tau
0-10	**0.768**	**1.631**	0.386	**0.727**	**1.697**	0.395
0-50	0.753	**1.631**	0.384	0.717	1.737	0.374
0-100	0.732	1.707	0.403	0.707	1.803	0.438
0-250	0.722	1.773	0.435	0.702	1.813	0.433
0-500	0.712	1.813	0.440	0.682	1.919	0.464
0-1000	0.692	1.833	0.428	0.687	1.909	0.438
0-2500	0.687	1.879	0.441	0.662	2.005	0.454
0-5000	0.687	1.854	0.445	0.662	1.995	0.456
0-10000	0.667	1.889	0.447	0.677	1.970	0.466
0-MAX	0.657	1.965	0.510	0.631	2.136	0.498

B Co-occurrence scores with FRANK1

B.1 Spam ranking thresholds

B.1.1 Log scaling

Table 48: Results for the profession training data when varying the spam ranking cutoff with log scaling

	Full data			Missing persons removed				
	Acc	Asd	Tau	Acc	Asd	Tau	Uniq	Ents
90-100	0.567	2.503	0.427	**0.611**	**2.209**	0.415	107	440
80-100	0.579	2.427	0.442	0.607	2.226	0.439	114	455
70-100	0.581	**2.388**	0.435	0.602	2.223	0.432	117	462
60-100	0.586	2.390	0.420	0.602	2.259	0.417	120	475
50-100	0.586	2.396	0.416	0.599	2.287	0.413	122	481
40-100	**0.588**	2.402	0.423	0.598	2.304	0.418	123	483
30-100	0.584	2.427	0.435	0.594	2.331	0.431	123	483
20-100	0.577	2.435	0.434	0.586	2.340	0.430	123	483
10-100	0.575	2.431	0.437	0.584	2.335	0.433	123	483
0-100	0.581	2.445	0.423	0.590	2.350	0.417	123	483

Table 49: Results for the profession test data when varying the spam ranking cutoff with log scaling

	Full data			Missing persons removed				
	Acc	Asd	Tau	Acc	Asd	Tau	Uniq	Ents
90-100	0.542	2.698	0.403	**0.616**	**2.217**	0.388	101	406
80-100	0.540	2.667	0.431	0.599	2.300	0.422	108	424
70-100	0.538	2.647	0.446	0.588	2.314	0.443	111	430
60-100	0.546	2.634	0.436	0.587	2.358	0.431	114	438
50-100	0.550	2.591	0.436	0.591	2.318	0.432	115	440
40-100	0.554	2.610	0.445	0.595	2.341	0.442	115	440
30-100	0.559	2.612	0.444	0.600	2.355	0.442	116	442
20-100	0.558	2.614	0.431	0.597	2.357	0.427	116	442
10-100	**0.561**	**2.589**	0.420	0.602	2.328	0.414	116	442
0-100	0.550	2.591	0.415	0.586	2.342	0.404	117	444

Table 50: Results for the nationality train data when varying the spam ranking cutoff with log scaling

	Full data			Missing persons removed				
	Acc	Asd	Tau	Acc	Asd	Tau	Uniq	Ents
90-100	0.588	2.581	0.439	0.700	1.946	0.449	61	130
80-100	0.631	2.225	0.443	0.678	1.952	0.437	69	146
70-100	0.675	2.100	0.431	0.707	1.907	0.426	71	150
60-100	0.694	2.094	0.445	0.727	1.900	0.441	71	150
50-100	0.681	2.062	0.453	0.713	1.867	0.450	71	150
40-100	0.688	2.013	0.452	0.720	1.813	0.449	71	150
30-100	0.706	1.962	0.439	0.740	1.760	0.435	71	150
20-100	**0.713**	**1.931**	0.446	**0.747**	**1.727**	0.442	71	150
10-100	0.706	1.969	0.462	0.740	1.767	0.459	71	150
0-100	**0.713**	1.962	0.464	**0.747**	1.760	0.462	71	150

Table 51: Results for the nationality test data when varying the spam ranking cutoff with log scaling

	Full data			Missing persons removed				
	Acc	Asd	Tau	Acc	Asd	Tau	Uniq	Ents
90-100	0.545	2.551	0.449	0.641	2.032	0.455	75	156
80-100	0.586	2.379	0.496	0.632	2.098	0.501	84	174
70-100	0.591	2.364	0.511	0.631	2.114	0.519	85	176
60-100	0.581	2.323	0.483	0.619	2.068	0.487	85	176
50-100	0.596	2.268	0.468	0.636	2.006	0.470	85	176
40-100	0.606	2.222	0.489	0.646	1.978	0.493	86	178
30-100	0.596	2.227	0.499	0.635	1.983	0.505	86	178
20-100	0.606	2.202	0.477	0.646	**1.955**	0.480	86	178
10-100	**0.611**	2.202	0.494	**0.652**	**1.955**	0.499	86	178
0-100	**0.611**	**2.197**	0.525	0.644	1.994	0.534	87	180

B.1.2 Lin scaling

Table 52: Results for the profession training data when varying the spam ranking cutoff with linear scaling

	Full data			Missing persons removed				
	Acc	Asd	Tau	Acc	Asd	Tau	Uniq	Ents
90-100	0.520	2.769	0.428	0.557	**2.520**	0.417	107	440
80-100	0.522	2.718	0.436	0.543	2.556	0.431	114	455
70-100	0.528	2.678	0.432	0.543	2.545	0.428	117	462
60-100	0.534	2.689	0.413	0.545	2.583	0.408	120	475
50-100	0.540	2.656	0.408	0.549	2.565	0.405	122	481
40-100	**0.550**	**2.652**	0.417	**0.557**	2.571	0.412	123	483
30-100	0.544	2.674	0.423	0.551	2.594	0.418	123	483
20-100	0.540	2.693	0.427	0.547	2.615	0.422	123	483
10-100	0.542	2.689	0.421	0.549	2.611	0.415	123	483
0-100	0.530	2.732	0.431	0.536	2.656	0.426	123	483

Table 53: Results for the profession test data when varying the spam ranking cutoff with linear scaling

	Full data			Missing persons removed				
	Acc	Asd	Tau	Acc	Asd	Tau	Uniq	Ents
90-100	0.478	2.967	0.406	0.534	2.557	0.391	101	406
80-100	0.495	2.928	0.447	0.545	2.616	0.443	108	424
70-100	0.507	2.858	0.455	0.551	2.565	0.453	111	430
60-100	0.515	2.807	0.427	0.550	2.562	0.421	114	438
50-100	0.517	2.776	0.428	0.552	2.534	0.423	115	440
40-100	0.515	2.768	0.427	0.550	2.525	0.421	115	440
30-100	0.517	2.750	0.419	0.550	2.516	0.413	116	442
20-100	0.524	2.725	0.413	0.559	2.486	0.405	116	442
10-100	0.528	2.737	0.422	0.563	2.500	0.417	116	442
0-100	**0.532**	**2.712**	0.415	**0.565**	**2.482**	0.405	117	444

Table 54: Results for the nationality train data when varying the spam ranking cutoff with linear scaling

	Full data			Missing persons removed				
	Acc	Asd	Tau	Acc	Asd	Tau	Uniq	Ents
90-100	0.525	2.913	0.469	**0.623**	2.354	0.486	61	130
80-100	0.562	2.587	0.480	0.603	2.349	0.478	69	146
70-100	0.581	2.519	0.486	0.607	2.353	0.485	71	150
60-100	0.581	2.544	0.499	0.607	2.380	0.499	71	150
50-100	0.588	2.494	0.486	0.613	2.327	0.485	71	150
40-100	0.581	2.450	0.477	0.607	2.280	0.475	71	150
30-100	0.588	2.413	0.478	0.613	2.240	0.476	71	150
20-100	**0.594**	**2.381**	0.491	0.620	**2.207**	0.490	71	150
10-100	0.556	2.519	0.530	0.580	2.353	0.532	71	150
0-100	0.531	2.606	0.543	0.553	2.447	0.546	71	150

Table 55: Results for the nationality test data when varying the spam ranking cutoff with linear scaling

	Full data			Missing persons removed				
	Acc	Asd	Tau	Acc	Asd	Tau	Uniq	Ents
90-100	0.510	2.778	0.455	0.596	2.321	0.462	75	156
80-100	0.525	2.667	0.478	0.563	2.425	0.481	84	174
70-100	0.540	2.601	0.497	0.574	2.381	0.503	85	176
60-100	0.551	2.535	0.490	0.585	2.307	0.495	85	176
50-100	0.581	2.444	0.480	0.619	2.205	0.483	85	176
40-100	**0.601**	2.348	0.468	**0.640**	2.118	0.470	86	178
30-100	0.586	2.374	0.448	0.624	2.146	0.448	86	178
20-100	**0.601**	**2.333**	0.462	**0.640**	**2.101**	0.464	86	178
10-100	0.591	2.364	0.457	0.629	2.135	0.458	86	178
0-100	0.571	2.429	0.509	0.600	2.250	0.516	87	180

B.1.3 Default score 5

Table 56: Results for the profession train data when varying the spam ranking cutoff with zero scores replaced with five

	Log scaling			Lin scaling		
	Acc	Asd	Tau	Acc	Asd	Tau
90-100	**0.625**	**2.173**	0.429	**0.550**	**2.472**	0.458
80-100	0.606	2.237	0.440	0.526	2.551	0.484
70-100	0.600	2.241	0.434	0.528	2.538	0.483
60-100	0.610	2.239	0.420	0.540	2.536	0.465
50-100	0.604	2.262	0.414	0.534	2.542	0.462
40-100	0.598	2.303	0.423	0.540	2.553	0.466
30-100	0.594	2.328	0.435	0.538	2.544	0.469
20-100	0.586	2.340	0.432	0.542	2.544	0.471
10-100	0.584	2.336	0.440	0.534	2.561	0.467
0-100	0.592	2.344	0.425	0.534	2.542	0.468

Table 57: Results for the profession test data when varying the spam ranking cutoff with zero scores replaced with five

	Log scaling			Lin scaling		
	Acc	Asd	Tau	Acc	Asd	Tau
90-100	**0.657**	**2.084**	0.409	0.558	2.450	0.450
80-100	0.645	2.129	0.440	0.559	2.505	0.489
70-100	0.639	2.136	0.455	0.563	2.474	0.492
60-100	0.639	2.162	0.437	0.563	2.462	0.471
50-100	0.628	2.181	0.440	0.554	2.474	0.471
40-100	0.634	2.191	0.447	0.554	2.452	0.472
30-100	0.634	2.211	0.445	0.558	2.441	0.466
20-100	0.634	2.211	0.430	0.569	2.396	0.460
10-100	0.637	2.197	0.421	**0.579**	2.386	0.468
0-100	0.622	2.214	0.414	0.575	**2.384**	0.467

Table 58: Results for the nationality train data when varying the spam ranking cutoff with zero scores replaced with five

	Log scaling			Lin scaling		
	Acc	Asd	Tau	Acc	Asd	Tau
90-100	**0.769**	1.725	0.439	**0.762**	**1.837**	0.465
80-100	0.744	1.700	0.443	0.713	1.887	0.475
70-100	0.750	1.738	0.431	0.706	1.944	0.482
60-100	**0.769**	1.731	0.445	0.700	1.988	0.494
50-100	0.750	1.731	0.453	0.706	1.962	0.486
40-100	0.750	1.712	0.452	0.688	1.956	0.472
30-100	0.762	1.694	0.439	0.681	1.981	0.473
20-100	**0.769**	**1.663**	0.446	0.688	1.950	0.486
10-100	0.750	1.762	0.462	0.650	2.087	0.526
0-100	0.756	1.756	0.464	0.631	2.144	0.539

Table 59: Results for the nationality test data when varying the spam ranking cutoff with zero scores replaced with five

	Log scaling			Lin scaling		
	Acc	Asd	Tau	Acc	Asd	Tau
90-100	**0.697**	**1.854**	0.449	**0.692**	**1.919**	0.455
80-100	0.672	1.949	0.496	0.636	2.101	0.476
70-100	0.667	1.985	0.511	0.631	2.126	0.495
60-100	0.652	1.970	0.483	0.631	2.101	0.488
50-100	0.657	1.955	0.468	0.646	2.056	0.478
40-100	0.657	1.960	0.489	0.667	1.980	0.466
30-100	0.646	1.965	0.499	0.652	2.005	0.447
20-100	0.657	1.939	0.477	0.657	2.005	0.460
10-100	0.662	1.939	0.494	0.657	2.005	0.455
0-100	0.657	1.970	0.525	0.636	2.101	0.507

B.2 Inner document thresholds

B.2.1 Log scaling

Table 60: Results for the profession training data when varying the co-occurence distance cutoff with log scaling

	Full data			Missing persons removed				
	Acc	Asd	Tau	Acc	Asd	Tau	Uniq	Ents
0-10	0.604	2.398	0.370	0.657	2.075	0.337	102	428
0-50	0.625	2.313	0.371	**0.666**	2.063	0.348	109	446
0-100	**0.627**	**2.252**	0.361	0.658	**2.048**	0.342	113	456
0-250	0.614	2.299	0.381	0.640	2.114	0.366	114	458
0-500	0.608	2.293	0.400	0.628	2.132	0.387	116	462
0-1000	0.598	2.328	0.411	0.617	2.187	0.402	119	470
0-2500	0.596	2.350	0.428	0.613	2.229	0.422	121	475
0-5000	0.586	2.367	0.419	0.602	2.248	0.413	121	475
0-10000	0.581	2.386	0.423	0.593	2.277	0.417	122	477
0-MAX	0.581	2.445	0.423	0.590	2.350	0.417	123	483

Table 61: Results for the profession test data when varying the co-occurence distance cutoff with log scaling

	Full data			Missing persons removed				
	Acc	Asd	Tau	Acc	Asd	Tau	Uniq	Ents
0-10	0.583	**2.464**	0.334	**0.674**	**1.904**	0.294	102	405
0-50	**0.585**	2.499	0.366	0.660	2.048	0.337	105	415
0-100	0.577	2.524	0.374	0.642	2.132	0.351	108	424
0-250	0.571	2.483	0.374	0.628	2.128	0.353	110	430
0-500	0.571	2.509	0.383	0.625	2.174	0.365	111	432
0-1000	0.565	2.474	0.381	0.615	2.156	0.365	113	436
0-2500	0.565	2.489	0.378	0.609	2.200	0.364	115	440
0-5000	0.561	2.515	0.373	0.602	2.242	0.360	116	442
0-10000	0.554	2.573	0.389	0.593	2.310	0.378	116	442
0-MAX	0.550	2.591	0.415	0.586	2.342	0.404	117	444

Table 62: Results for the nationality training data when varying the co-occurence distance cutoff with log scaling

	Full data			Missing persons removed				
	Acc	Asd	Tau	Acc	Asd	Tau	Uniq	Ents
0-10	0.600	2.481	0.304	0.730	1.656	0.274	57	122
0-50	0.662	2.231	0.334	0.743	1.757	0.303	64	136
0-100	0.681	2.038	0.370	0.743	**1.643**	0.350	66	140
0-250	0.694	1.975	0.390	0.740	1.705	0.379	69	146
0-500	**0.713**	**1.881**	0.422	**0.747**	1.673	0.416	71	150
0-1000	0.694	1.938	0.435	0.727	1.733	0.430	71	150
0-2500	0.694	1.950	0.435	0.727	1.747	0.430	71	150
0-5000	0.700	1.906	0.448	0.733	1.700	0.444	71	150
0-10000	0.694	1.981	0.468	0.727	1.780	0.465	71	150
0-MAX	**0.713**	1.962	0.464	**0.747**	1.760	0.462	71	150

Table 63: Results for the nationality test data when varying the co-occurence distance cutoff with log scaling

	Full data			Missing persons removed				
	Acc	Asd	Tau	Acc	Asd	Tau	Uniq	Ents
0-10	0.515	2.783	0.412	0.637	2.158	0.409	70	146
0-50	0.576	2.490	0.485	0.652	2.067	0.489	79	164
0-100	0.581	2.333	0.428	0.649	1.935	0.421	81	168
0-250	0.591	2.318	0.487	0.644	1.994	0.491	84	174
0-500	0.591	2.268	0.498	0.636	1.989	0.504	85	176
0-1000	0.611	2.192	0.496	0.652	1.944	0.502	86	178
0-2500	0.631	2.177	0.486	0.674	**1.927**	0.490	86	178
0-5000	0.636	**2.146**	0.498	0.672	1.939	0.504	87	180
0-10000	**0.641**	2.167	0.489	**0.678**	1.961	0.494	87	180
0-MAX	0.611	2.197	0.525	0.644	1.994	0.534	87	180

B.2.2 Lin scaling

Table 64: Results for the profession training data when varying the co-occurence distance cutoff with linear scaling

	Full data			Missing persons removed				
	Acc	Asd	Tau	Acc	Asd	Tau	Uniq	Ents
0-10	0.557	2.654	0.369	**0.600**	2.383	0.335	102	428
0-50	0.565	2.583	0.383	0.596	2.374	0.363	109	446
0-100	0.567	2.536	0.358	0.590	**2.368**	0.338	113	456
0-250	**0.575**	2.544	0.382	0.596	2.389	0.368	114	458
0-500	**0.575**	**2.513**	0.399	0.591	2.377	0.385	116	462
0-1000	0.561	2.548	0.415	0.577	2.428	0.406	119	470
0-2500	0.563	2.577	0.423	0.577	2.476	0.416	121	475
0-5000	0.551	2.604	0.426	0.564	2.505	0.420	121	475
0-10000	0.561	2.621	0.430	0.572	2.530	0.425	122	477
0-MAX	0.530	2.732	0.431	0.536	2.656	0.426	123	483

Table 65: Results for the profession test data when varying the co-occurence distance cutoff with linear scaling

	Full data			Missing persons removed				
	Acc	Asd	Tau	Acc	Asd	Tau	Uniq	Ents
0-10	0.540	2.684	0.339	0.620	**2.183**	0.299	102	405
0-50	**0.556**	2.610	0.361	**0.624**	2.186	0.331	105	415
0-100	**0.556**	2.598	0.361	0.616	2.222	0.335	108	424
0-250	**0.556**	**2.575**	0.368	0.609	2.237	0.346	110	430
0-500	0.550	2.591	0.384	0.600	2.271	0.366	111	432
0-1000	0.548	2.585	0.373	0.594	2.287	0.356	113	436
0-2500	0.534	2.612	0.362	0.573	2.343	0.346	115	440
0-5000	0.532	2.616	0.366	0.568	2.360	0.351	116	442
0-10000	0.532	2.647	0.384	0.568	2.396	0.372	116	442
0-MAX	0.532	2.712	0.415	0.565	2.482	0.405	117	444

Table 66: Results for the nationality training data when varying the co-occurence distance cutoff with linear scaling

	Full data			Missing persons removed				
	Acc	Asd	Tau	Acc	Asd	Tau	Uniq	Ents
0-10	0.544	2.669	0.296	0.656	1.902	0.263	57	122
0-50	0.625	2.344	0.377	0.699	1.890	0.354	64	136
0-100	**0.644**	2.212	0.371	**0.700**	**1.843**	0.351	66	140
0-250	0.637	2.206	0.400	0.678	1.959	0.389	69	146
0-500	**0.644**	**2.150**	0.407	0.673	1.960	0.400	71	150
0-1000	0.625	2.244	0.412	0.653	2.060	0.406	71	150
0-2500	0.606	2.331	0.449	0.633	2.153	0.445	71	150
0-5000	0.600	2.288	0.458	0.627	2.107	0.455	71	150
0-10000	0.588	2.419	0.493	0.613	2.247	0.492	71	150
0-MAX	0.531	2.606	0.543	0.553	2.447	0.546	71	150

Table 67: Results for the nationality test data when varying the co-occurence distance cutoff with linear scaling

	Full data			Missing persons removed				
	Acc	Asd	Tau	Acc	Asd	Tau	Uniq	Ents
0-10	0.475	2.904	0.407	0.582	2.322	0.401	70	146
0-50	0.520	2.737	0.503	0.585	2.366	0.510	79	164
0-100	0.545	2.591	0.444	0.607	2.238	0.440	81	168
0-250	0.535	2.611	0.485	0.580	2.328	0.489	84	174
0-500	0.540	2.566	0.489	0.580	2.324	0.493	85	176
0-1000	0.540	2.540	0.502	0.573	2.331	0.509	86	178
0-2500	0.551	2.439	0.491	0.584	2.219	0.496	86	178
0-5000	0.581	2.374	0.494	0.611	2.189	0.500	87	180
0-10000	**0.586**	**2.343**	0.467	**0.617**	**2.156**	0.469	87	180
0-MAX	0.571	2.429	0.509	0.600	2.250	0.516	87	180

B.2.3 Default score 5

Table 68: Results for the profession train data when varying the co-occurence distance cutoff with zero scores replaced with five

	Log scaling			Lin scaling		
	Acc	Asd	Tau	Acc	Asd	Tau
0-10	0.629	2.120	0.389	**0.604**	**2.181**	0.401
0-50	**0.641**	2.118	0.402	0.588	2.245	0.418
0-100	0.633	2.130	0.387	0.567	2.311	0.405
0-250	0.639	**2.111**	0.397	0.565	2.390	0.439
0-500	0.635	2.118	0.411	0.557	2.425	0.464
0-1000	0.623	2.171	0.420	0.538	2.474	0.471
0-2500	0.614	2.219	0.432	0.532	2.538	0.478
0-5000	0.612	2.214	0.420	0.520	2.573	0.480
0-10000	0.592	2.285	0.426	0.526	2.598	0.477
0-MAX	0.592	2.344	0.425	0.534	2.542	0.468

Table 69: Results for the profession test data when varying the co-occurence distance cutoff with zero scores replaced with five

	Log scaling			Lin scaling		
	Acc	Asd	Tau	Acc	Asd	Tau
0-10	0.655	2.045	0.359	**0.612**	**2.191**	0.376
0-50	**0.673**	**2.010**	0.380	0.610	2.209	0.396
0-100	0.667	2.029	0.384	0.608	2.222	0.400
0-250	0.653	2.027	0.386	0.598	2.248	0.414
0-500	0.657	2.045	0.392	0.593	2.283	0.437
0-1000	0.653	2.023	0.391	0.585	2.302	0.429
0-2500	0.653	2.039	0.389	0.565	2.351	0.413
0-5000	0.637	2.111	0.382	0.565	2.357	0.414
0-10000	0.634	2.154	0.396	0.563	2.382	0.431
0-MAX	0.622	2.214	0.414	0.575	2.384	0.467

Table 70: Results for the nationality train data when varying the co-occurence distance cutoff with zero scores replaced with five

	Log scaling			Lin scaling		
	Acc	Asd	Tau	Acc	Asd	Tau
0-10	**0.819**	**1.462**	0.304	0.769	1.606	0.300
0-50	0.794	1.575	0.330	**0.775**	1.594	0.375
0-100	0.762	1.606	0.370	0.756	1.650	0.366
0-250	0.762	1.631	0.390	**0.775**	**1.581**	0.395
0-500	0.750	1.681	0.422	0.731	1.762	0.402
0-1000	0.731	1.738	0.435	0.706	1.863	0.408
0-2500	0.731	1.762	0.435	0.694	1.919	0.445
0-5000	0.744	1.700	0.448	0.694	1.869	0.453
0-10000	0.744	1.744	0.468	0.681	1.988	0.489
0-MAX	0.756	1.756	0.464	0.631	2.144	0.539

Table 71: Results for the nationality test data when varying the co-occurence distance cutoff with zero scores replaced with five

	Log scaling			Lin scaling		
	Acc	Asd	Tau	Acc	Asd	Tau
0-10	**0.773**	**1.657**	0.414	**0.753**	**1.677**	0.410
0-50	0.722	1.798	0.485	0.677	2.015	0.505
0-100	0.712	1.747	0.428	0.677	1.985	0.444
0-250	0.697	1.808	0.487	0.657	1.995	0.485
0-500	0.667	1.889	0.498	0.652	2.000	0.485
0-1000	0.667	1.894	0.496	0.631	2.076	0.502
0-2500	0.692	1.864	0.486	0.626	2.030	0.491
0-5000	0.687	1.884	0.498	0.652	2.000	0.494
0-10000	0.692	1.904	0.489	0.657	1.980	0.467
0-MAX	0.657	1.970	0.525	0.636	2.101	0.507

C ML scores with FACC1

C.1 Spam ranking thresholding

Table 72: Spam ranking with FACC1 on profession train set with entity

		Maplin		Maplog		Maplin2		Maplog2	
Spam		Full	KB	Full	KB	Full	KB	Full	KB
0	Acc	0.584	0.625	0.676	0.643	0.662	0.718	0.705	0.705
	Asd	2.472	2.202	2.039	2.087	2.140	2.010	2.074	2.085
	Tau	0.392	0.344	0.364	0.348	0.485	0.395	0.462	0.509
50	Acc	0.571	0.633	0.674	0.631	0.652	0.722	0.703	0.685
	Asd	2.522	2.208	2.080	2.124	2.165	2.014	2.091	2.120
	Tau	0.394	0.341	0.379	0.341	0.482	0.415	0.476	0.505
75	Acc	0.557	0.614	0.662	0.600	0.645	0.703	0.691	0.656
	Asd	2.604	2.309	2.148	2.225	2.216	2.052	2.140	2.184
	Tau	0.405	0.365	0.398	0.365	0.530	0.436	0.497	0.544
88	Acc	0.548	0.600	0.647	0.573	0.635	0.693	0.674	0.637
	Asd	2.664	2.348	2.192	2.282	2.256	2.085	2.151	2.225
	Tau	0.420	0.379	0.401	0.380	0.551	0.435	0.493	0.558

Table 73: Spam ranking with FACC1 on profession train set without entity

		Maplin		Maplog		Maplin2		Maplog2	
Spam		Full	KB	Full	KB	Full	KB	Full	KB
0	Acc	0.590	0.625	0.678	0.652	0.680	0.717	0.707	0.709
	Asd	2.454	2.181	2.012	2.047	2.103	1.963	2.054	2.045
	Tau	0.378	0.339	0.363	0.347	0.482	0.400	0.480	0.486
50	Acc	0.559	0.621	0.676	0.643	0.649	0.711	0.703	0.697
	Asd	2.575	2.243	2.056	2.109	2.171	2.006	2.078	2.087
	Tau	0.401	0.343	0.373	0.347	0.505	0.401	0.493	0.518
75	Acc	0.553	0.604	0.662	0.614	0.643	0.691	0.691	0.672
	Asd	2.596	2.332	2.144	2.190	2.212	2.060	2.109	2.138
	Tau	0.400	0.369	0.398	0.364	0.553	0.424	0.477	0.531
88	Acc	0.546	0.586	0.637	0.579	0.633	0.680	0.662	0.645
	Asd	2.683	2.423	2.235	2.307	2.252	2.118	2.169	2.212
	Tau	0.427	0.385	0.405	0.390	0.573	0.457	0.482	0.544

Table 74: Spam ranking with FACC1 on profession test set with entity

		Maplin		Maplog		Maplin2		Maplog2	
Spam		Full	KB	Full	KB	Full	KB	Full	KB
0	Acc	0.583	0.655	0.704	0.684	0.713	0.766	0.750	0.735
	Asd	2.351	2.021	1.846	1.914	2.086	1.928	1.926	1.988
	Tau	0.377	0.325	0.347	0.325	0.469	0.402	0.494	0.560
50	Acc	0.583	0.661	0.684	0.667	0.713	0.768	0.731	0.719
	Asd	2.398	2.049	1.903	1.938	2.090	1.928	1.971	2.010
	Tau	0.401	0.347	0.383	0.343	0.463	0.395	0.510	0.557
75	Acc	0.583	0.637	0.682	0.659	0.704	0.749	0.729	0.717
	Asd	2.433	2.129	1.943	1.961	2.113	1.975	1.982	2.012
	Tau	0.412	0.375	0.388	0.376	0.500	0.420	0.523	0.569
88	Acc	0.571	0.632	0.674	0.657	0.698	0.745	0.721	0.713
	Asd	2.452	2.135	1.951	1.996	2.127	1.982	1.979	2.031
	Tau	0.408	0.373	0.401	0.373	0.503	0.434	0.534	0.595

Table 75: Spam ranking with FACC1 on profession test set without entity

		Maplin		Maplog		Maplin2		Maplog2	
Spam		Full	KB	Full	KB	Full	KB	Full	KB
0	Acc	0.593	0.659	0.725	0.696	0.715	0.762	0.762	0.745
	Asd	2.351	2.045	1.821	1.901	2.090	1.938	1.887	1.961
	Tau	0.377	0.336	0.362	0.332	0.480	0.402	0.474	0.519
50	Acc	0.575	0.641	0.708	0.676	0.702	0.758	0.749	0.735
	Asd	2.374	2.062	1.856	1.906	2.096	1.942	1.910	1.961
	Tau	0.400	0.362	0.395	0.357	0.502	0.431	0.496	0.523
75	Acc	0.567	0.632	0.684	0.665	0.692	0.752	0.729	0.727
	Asd	2.429	2.133	1.928	1.981	2.119	1.981	1.961	1.996
	Tau	0.389	0.360	0.388	0.359	0.517	0.447	0.503	0.556
88	Acc	0.579	0.634	0.659	0.649	0.692	0.747	0.708	0.712
	Asd	2.446	2.183	1.986	2.033	2.138	2.002	1.981	2.016
	Tau	0.399	0.387	0.412	0.395	0.537	0.460	0.531	0.597

Table 76: Spam ranking with FACC1 on nationality train set with entity

Spam		Maplin		Maplog		Maplin2		Maplog2	
		Full	KB	Full	KB	Full	KB	Full	KB
0	Acc	0.698	0.741	0.796	0.796	0.747	0.802	0.840	0.846
	Asd	1.827	1.605	1.525	1.488	1.963	1.827	1.784	1.790
	Tau	0.335	0.294	0.362	0.300	0.431	0.429	0.705	0.724
50	Acc	0.704	0.728	0.802	0.802	0.765	0.796	0.852	0.858
	Asd	1.833	1.667	1.500	1.488	1.938	1.827	1.778	1.778
	Tau	0.360	0.342	0.380	0.347	0.418	0.436	0.704	0.724
75	Acc	0.722	0.759	0.796	0.802	0.765	0.815	0.840	0.846
	Asd	1.852	1.648	1.537	1.463	1.938	1.846	1.784	1.784
	Tau	0.346	0.292	0.396	0.290	0.444	0.474	0.718	0.737
88	Acc	0.728	0.759	0.802	0.796	0.772	0.815	0.846	0.852
	Asd	1.827	1.623	1.543	1.488	1.951	1.833	1.784	1.790
	Tau	0.346	0.303	0.389	0.311	0.431	0.416	0.705	0.724

Table 77: Spam ranking with FACC1 on nationality train set without entity

Spam		Maplin		Maplog		Maplin2		Maplog2	
		Full	KB	Full	KB	Full	KB	Full	KB
0	Acc	0.580	0.704	0.809	0.802	0.636	0.772	0.852	0.852
	Asd	2.148	1.753	1.568	1.525	2.185	1.901	1.759	1.772
	Tau	0.364	0.336	0.506	0.339	0.389	0.454	0.678	0.704
50	Acc	0.611	0.710	0.815	0.796	0.654	0.765	0.852	0.846
	Asd	2.167	1.728	1.568	1.537	2.160	1.870	1.765	1.772
	Tau	0.374	0.319	0.469	0.317	0.388	0.473	0.698	0.711
75	Acc	0.617	0.704	0.827	0.796	0.667	0.778	0.864	0.858
	Asd	2.074	1.722	1.580	1.525	2.123	1.877	1.765	1.772
	Tau	0.400	0.341	0.514	0.338	0.427	0.446	0.711	0.724
88	Acc	0.611	0.710	0.827	0.802	0.660	0.778	0.864	0.864
	Asd	2.117	1.735	1.549	1.537	2.136	1.901	1.741	1.747
	Tau	0.376	0.350	0.444	0.347	0.397	0.403	0.678	0.691

Table 78: Spam ranking with FACC1 on nationality test set with entity

Spam		Maplin		Maplog		Maplin2		Maplog2	
		Full	KB	Full	KB	Full	KB	Full	KB
0	Acc	0.660	0.677	0.731	0.702	0.746	0.788	0.797	0.798
	Asd	2.086	2.045	1.787	1.843	1.914	1.828	1.761	1.742
	Tau	0.437	0.403	0.446	0.402	0.503	0.540	0.775	0.773
50	Acc	0.655	0.667	0.726	0.697	0.741	0.778	0.792	0.793
	Asd	2.071	2.040	1.787	1.859	1.893	1.823	1.756	1.747
	Tau	0.466	0.424	0.493	0.423	0.506	0.524	0.785	0.773
75	Acc	0.665	0.667	0.716	0.692	0.756	0.788	0.792	0.798
	Asd	2.056	2.005	1.792	1.864	1.868	1.798	1.751	1.737
	Tau	0.426	0.389	0.482	0.391	0.471	0.510	0.790	0.783
88	Acc	0.645	0.672	0.721	0.702	0.746	0.783	0.797	0.803
	Asd	2.046	2.000	1.777	1.869	1.898	1.798	1.756	1.732
	Tau	0.445	0.414	0.503	0.412	0.476	0.494	0.795	0.783

Table 79: Spam ranking with FACC1 on nationality train set without entity

Spam		Maplin		Maplog		Maplin2		Maplog2	
		Full	KB	Full	KB	Full	KB	Full	KB
0	Acc	0.604	0.646	0.751	0.732	0.695	0.763	0.817	0.818
	Asd	2.168	2.116	1.716	1.848	2.041	1.864	1.701	1.707
	Tau	0.494	0.436	0.569	0.438	0.522	0.538	0.759	0.773
50	Acc	0.629	0.662	0.741	0.732	0.726	0.778	0.812	0.818
	Asd	2.173	2.061	1.741	1.818	1.995	1.823	1.721	1.707
	Tau	0.450	0.412	0.539	0.418	0.502	0.555	0.771	0.764
75	Acc	0.650	0.677	0.741	0.732	0.726	0.778	0.817	0.823
	Asd	2.112	2.051	1.701	1.818	1.985	1.848	1.701	1.697
	Tau	0.438	0.396	0.518	0.397	0.476	0.570	0.801	0.804
88	Acc	0.645	0.677	0.746	0.727	0.721	0.778	0.822	0.828
	Asd	2.132	1.995	1.665	1.808	2.015	1.838	1.701	1.697
	Tau	0.417	0.374	0.473	0.373	0.467	0.545	0.790	0.794

C.2 Varying n-gram lengths

Table 80: N-gram length with FACC1 on profession train set with entity

N-gram		Maplin		Maplog		Maplin2		Maplog2	
		Full	KB	Full	KB	Full	KB	Full	KB
4	Acc	0.569	0.623	0.674	0.617	0.654	0.717	0.701	0.672
	Asd	2.532	2.243	2.103	2.155	2.163	2.014	2.101	2.124
	Tau	0.389	0.355	0.395	0.356	0.491	0.416	0.470	0.524
5	Acc	0.557	0.614	0.662	0.600	0.645	0.703	0.691	0.656
	Asd	2.604	2.309	2.148	2.225	2.216	2.052	2.140	2.184
	Tau	0.405	0.365	0.398	0.365	0.530	0.436	0.497	0.544
6	Acc	0.553	0.604	0.652	0.586	0.641	0.695	0.680	0.647
	Asd	2.617	2.320	2.190	2.276	2.239	2.066	2.179	2.219
	Tau	0.412	0.400	0.403	0.391	0.537	0.439	0.519	0.562
7	Acc	0.551	0.602	0.637	0.563	0.623	0.687	0.662	0.629
	Asd	2.631	2.350	2.233	2.330	2.268	2.099	2.198	2.245
	Tau	0.415	0.394	0.409	0.397	0.549	0.450	0.519	0.583

Table 81: N-gram length with FACC1 on profession train set without entity

N-gram		Maplin		Maplog		Maplin2		Maplog2	
		Full	KB	Full	KB	Full	KB	Full	KB
4	Acc	0.559	0.608	0.676	0.623	0.649	0.695	0.707	0.683
	Asd	2.581	2.297	2.062	2.136	2.190	2.033	2.054	2.091
	Tau	0.402	0.362	0.370	0.366	0.532	0.411	0.458	0.510
5	Acc	0.553	0.604	0.662	0.614	0.643	0.691	0.691	0.672
	Asd	2.596	2.332	2.144	2.190	2.212	2.060	2.109	2.138
	Tau	0.400	0.369	0.398	0.364	0.553	0.424	0.477	0.531
6	Acc	0.561	0.614	0.643	0.592	0.639	0.693	0.672	0.652
	Asd	2.565	2.318	2.198	2.268	2.219	2.066	2.161	2.194
	Tau	0.417	0.376	0.394	0.383	0.549	0.429	0.476	0.549
7	Acc	0.569	0.612	0.619	0.571	0.641	0.691	0.647	0.633
	Asd	2.579	2.353	2.266	2.344	2.225	2.097	2.221	2.258
	Tau	0.416	0.387	0.412	0.391	0.553	0.449	0.505	0.562

Table 82: N-gram length with FACC1 on profession test set with entity

N-gram		Maplin		Maplog		Maplin2		Maplog2	
		Full	KB	Full	KB	Full	KB	Full	KB
4	Acc	0.589	0.639	0.686	0.665	0.715	0.754	0.733	0.721
	Asd	2.396	2.119	1.928	1.932	2.094	1.957	1.979	1.992
	Tau	0.409	0.369	0.401	0.362	0.489	0.416	0.513	0.548
5	Acc	0.583	0.637	0.682	0.659	0.704	0.749	0.729	0.717
	Asd	2.433	2.129	1.943	1.961	2.113	1.975	1.982	2.012
	Tau	0.412	0.375	0.388	0.376	0.500	0.420	0.523	0.569
6	Acc	0.565	0.630	0.680	0.653	0.690	0.735	0.731	0.708
	Asd	2.464	2.140	1.969	2.002	2.144	2.012	1.988	2.057
	Tau	0.397	0.379	0.379	0.382	0.512	0.436	0.522	0.593
7	Acc	0.569	0.620	0.682	0.655	0.686	0.727	0.727	0.710
	Asd	2.466	2.175	1.996	2.029	2.160	2.025	1.994	2.060
	Tau	0.393	0.375	0.380	0.380	0.512	0.431	0.542	0.604

Table 83: N-gram length with FACC1 on profession test set without entity

N-gram		Maplin		Maplog		Maplin2		Maplog2	
		Full	KB	Full	KB	Full	KB	Full	KB
4	Acc	0.571	0.641	0.712	0.688	0.694	0.752	0.752	0.743
	Asd	2.433	2.092	1.895	1.928	2.115	1.963	1.926	1.973
	Tau	0.387	0.364	0.389	0.358	0.514	0.433	0.483	0.546
5	Acc	0.567	0.632	0.684	0.665	0.692	0.752	0.729	0.727
	Asd	2.429	2.133	1.928	1.981	2.119	1.981	1.961	1.996
	Tau	0.389	0.360	0.388	0.359	0.517	0.447	0.503	0.556
6	Acc	0.569	0.624	0.661	0.653	0.688	0.739	0.710	0.715
	Asd	2.423	2.152	1.981	2.029	2.136	2.000	1.994	2.039
	Tau	0.401	0.371	0.405	0.370	0.537	0.453	0.519	0.594
7	Acc	0.575	0.626	0.645	0.641	0.684	0.735	0.692	0.704
	Asd	2.417	2.175	2.031	2.078	2.146	2.008	2.025	2.062
	Tau	0.421	0.398	0.419	0.400	0.539	0.445	0.550	0.617

Table 84: N-gram length with FACC1 on nationality train set with entity

N-gram		Maplin		Maplog		Maplin2		Maplog2	
		Full	KB	Full	KB	Full	KB	Full	KB
4	Acc	0.735	0.772	0.790	0.802	0.784	0.833	0.833	0.852
	Asd	1.858	1.660	1.562	1.481	1.957	1.852	1.809	1.802
	Tau	0.341	0.307	0.435	0.303	0.459	0.443	0.705	0.724
5	Acc	0.722	0.759	0.796	0.802	0.765	0.815	0.840	0.846
	Asd	1.852	1.648	1.537	1.463	1.938	1.846	1.784	1.784
	Tau	0.346	0.292	0.396	0.290	0.444	0.474	0.718	0.737
6	Acc	0.722	0.741	0.790	0.790	0.778	0.809	0.840	0.840
	Asd	1.796	1.673	1.537	1.519	1.914	1.858	1.790	1.796
	Tau	0.326	0.310	0.409	0.314	0.424	0.488	0.737	0.750
7	Acc	0.716	0.753	0.802	0.802	0.765	0.809	0.846	0.846
	Asd	1.772	1.617	1.537	1.500	1.914	1.846	1.778	1.778
	Tau	0.335	0.297	0.419	0.298	0.431	0.488	0.750	0.750

Table 85: N-gram length with FACC1 on nationality train set without entity

N-gram		Maplin		Maplog		Maplin2		Maplog2	
		Full	KB	Full	KB	Full	KB	Full	KB
4	Acc	0.599	0.704	0.821	0.796	0.654	0.784	0.858	0.858
	Asd	2.111	1.747	1.549	1.543	2.142	1.889	1.759	1.772
	Tau	0.373	0.354	0.468	0.351	0.395	0.441	0.674	0.700
5	Acc	0.617	0.704	0.827	0.796	0.667	0.778	0.864	0.858
	Asd	2.074	1.722	1.580	1.525	2.123	1.877	1.765	1.772
	Tau	0.400	0.341	0.514	0.338	0.427	0.446	0.711	0.724
6	Acc	0.642	0.716	0.802	0.790	0.691	0.784	0.840	0.846
	Asd	1.981	1.673	1.574	1.537	2.043	1.858	1.772	1.772
	Tau	0.403	0.329	0.522	0.325	0.439	0.475	0.730	0.730
7	Acc	0.636	0.710	0.796	0.778	0.685	0.778	0.840	0.846
	Asd	1.957	1.691	1.568	1.543	2.043	1.895	1.765	1.765
	Tau	0.389	0.328	0.503	0.325	0.413	0.488	0.730	0.730

Table 86: N-gram length with FACC1 on nationality test set with entity

N-gram		Maplin		Maplog		Maplin2		Maplog2	
		Full	KB	Full	KB	Full	KB	Full	KB
4	Acc	0.675	0.677	0.721	0.697	0.772	0.793	0.792	0.798
	Asd	2.091	2.025	1.746	1.848	1.888	1.808	1.746	1.742
	Tau	0.452	0.400	0.479	0.403	0.492	0.494	0.772	0.773
5	Acc	0.665	0.667	0.716	0.692	0.756	0.788	0.792	0.798
	Asd	2.056	2.005	1.792	1.864	1.868	1.798	1.751	1.737
	Tau	0.426	0.389	0.482	0.391	0.471	0.510	0.790	0.783
6	Acc	0.665	0.657	0.706	0.687	0.751	0.778	0.792	0.798
	Asd	2.046	2.015	1.812	1.864	1.898	1.808	1.756	1.742
	Tau	0.429	0.393	0.472	0.391	0.520	0.525	0.801	0.794
7	Acc	0.645	0.657	0.706	0.697	0.751	0.778	0.802	0.808
	Asd	2.056	2.030	1.812	1.864	1.883	1.803	1.746	1.737
	Tau	0.414	0.403	0.495	0.402	0.481	0.528	0.780	0.773

Table 87: N-gram length with FACC1 on nationality train set without entity

N-gram		Maplin		Maplog		Maplin2		Maplog2	
		Full	KB	Full	KB	Full	KB	Full	KB
4	Acc	0.645	0.672	0.746	0.727	0.726	0.768	0.812	0.818
	Asd	2.132	2.051	1.675	1.803	1.970	1.838	1.706	1.707
	Tau	0.433	0.396	0.490	0.395	0.466	0.550	0.764	0.778
5	Acc	0.650	0.677	0.741	0.732	0.726	0.778	0.817	0.823
	Asd	2.112	2.051	1.701	1.818	1.985	1.848	1.701	1.697
	Tau	0.438	0.396	0.518	0.397	0.476	0.570	0.801	0.804
6	Acc	0.665	0.672	0.751	0.727	0.736	0.778	0.827	0.828
	Asd	2.096	1.990	1.706	1.808	1.980	1.833	1.695	1.697
	Tau	0.419	0.379	0.563	0.381	0.497	0.581	0.795	0.804
7	Acc	0.670	0.687	0.761	0.737	0.731	0.783	0.838	0.838
	Asd	2.000	1.980	1.701	1.833	1.949	1.833	1.706	1.702
	Tau	0.462	0.390	0.557	0.390	0.543	0.620	0.837	0.835

D ML scores with FRANK1

D.1 Spam ranking thresholding

Table 88: Spam ranking with FRANK1 on profession train set with entity

Spam		Maplin		Maplog		Maplin2		Maplog2	
		Full	KB	Full	KB	Full	KB	Full	KB
0	Acc	0.573	0.616	0.678	0.664	0.676	0.709	0.701	0.713
	Asd	2.478	2.212	1.878	1.950	2.140	2.014	1.986	1.994
	Tau	0.443	0.391	0.419	0.395	0.519	0.459	0.492	0.531
50	Acc	0.586	0.619	0.678	0.670	0.680	0.713	0.699	0.718
	Asd	2.433	2.202	1.887	1.951	2.118	2.016	1.992	1.996
	Tau	0.433	0.378	0.409	0.372	0.516	0.454	0.504	0.536
75	Acc	0.588	0.627	0.678	0.662	0.682	0.717	0.701	0.709
	Asd	2.408	2.190	1.895	1.967	2.111	2.004	1.990	1.996
	Tau	0.445	0.387	0.398	0.390	0.506	0.452	0.525	0.561
88	Acc	0.584	0.610	0.674	0.658	0.682	0.711	0.699	0.713
	Asd	2.425	2.229	1.901	1.969	2.117	2.014	1.990	1.992
	Tau	0.446	0.395	0.406	0.393	0.505	0.432	0.511	0.543

Table 89: Spam ranking with FRANK1 on profession train set without entity

		Maplin		Maplog		Maplin2		Maplog2	
Spam		Full	KB	Full	KB	Full	KB	Full	KB
0	Acc	0.583	0.650	0.717	0.687	0.670	0.728	0.740	0.732
	Asd	2.417	2.062	1.819	1.883	2.150	1.975	1.915	1.946
	Tau	0.384	0.332	0.380	0.326	0.518	0.436	0.494	0.579
50	Acc	0.581	0.637	0.709	0.687	0.668	0.724	0.740	0.732
	Asd	2.367	2.062	1.823	1.885	2.134	1.983	1.930	1.961
	Tau	0.371	0.319	0.358	0.320	0.509	0.435	0.497	0.587
75	Acc	0.590	0.647	0.711	0.683	0.674	0.734	0.744	0.728
	Asd	2.338	2.045	1.821	1.895	2.117	1.963	1.903	1.942
	Tau	0.367	0.316	0.356	0.324	0.510	0.410	0.465	0.572
88	Acc	0.563	0.631	0.724	0.705	0.656	0.722	0.763	0.753
	Asd	2.412	2.080	1.810	1.856	2.146	1.971	1.882	1.924
	Tau	0.369	0.307	0.366	0.301	0.512	0.410	0.459	0.550

Table 90: Spam ranking with FRANK1 on profession test set with entity

		Maplin		Maplog		Maplin2		Maplog2	
Spam		Full	KB	Full	KB	Full	KB	Full	KB
0	Acc	0.614	0.669	0.725	0.696	0.713	0.764	0.758	0.750
	Asd	2.298	2.066	1.805	1.842	2.094	1.965	1.891	1.904
	Tau	0.399	0.378	0.381	0.381	0.499	0.411	0.504	0.565
50	Acc	0.602	0.653	0.712	0.690	0.708	0.754	0.749	0.745
	Asd	2.310	2.068	1.819	1.860	2.097	1.967	1.904	1.924
	Tau	0.407	0.384	0.407	0.391	0.495	0.421	0.518	0.588
75	Acc	0.608	0.669	0.708	0.688	0.708	0.764	0.749	0.745
	Asd	2.302	2.070	1.828	1.838	2.113	1.975	1.906	1.914
	Tau	0.409	0.387	0.400	0.382	0.504	0.430	0.515	0.583
88	Acc	0.598	0.661	0.704	0.684	0.706	0.764	0.747	0.745
	Asd	2.318	2.096	1.832	1.858	2.109	1.969	1.910	1.901
	Tau	0.415	0.394	0.398	0.392	0.493	0.423	0.530	0.580

Table 91: Spam ranking with FRANK1 on profession test set without entity

		Maplin		Maplog		Maplin2		Maplog2	
Spam		Full	KB	Full	KB	Full	KB	Full	KB
0	Acc	0.591	0.673	0.708	0.698	0.692	0.764	0.741	0.743
	Asd	2.333	1.945	1.813	1.811	2.119	1.922	1.912	1.914
	Tau	0.365	0.320	0.348	0.309	0.503	0.402	0.496	0.579
50	Acc	0.591	0.671	0.710	0.700	0.690	0.766	0.739	0.743
	Asd	2.329	1.975	1.832	1.817	2.129	1.936	1.926	1.932
	Tau	0.396	0.333	0.347	0.313	0.511	0.425	0.497	0.577
75	Acc	0.589	0.671	0.706	0.692	0.688	0.766	0.737	0.737
	Asd	2.329	2.014	1.854	1.844	2.140	1.949	1.936	1.936
	Tau	0.366	0.341	0.349	0.334	0.493	0.404	0.496	0.583
88	Acc	0.579	0.663	0.713	0.690	0.688	0.766	0.750	0.743
	Asd	2.353	2.019	1.834	1.858	2.133	1.942	1.926	1.943
	Tau	0.364	0.333	0.374	0.324	0.518	0.440	0.502	0.590

Table 92: Spam ranking with FRANK1 on nationality train set with entity

		Maplin		Maplog		Maplin2		Maplog2	
Spam		Full	KB	Full	KB	Full	KB	Full	KB
0	Acc	0.673	0.753	0.784	0.796	0.722	0.802	0.833	0.846
	Asd	2.099	1.710	1.630	1.519	2.068	1.846	1.765	1.778
	Tau	0.393	0.389	0.444	0.395	0.448	0.461	0.678	0.743
50	Acc	0.648	0.722	0.796	0.802	0.698	0.772	0.840	0.852
	Asd	2.111	1.741	1.599	1.531	2.111	1.901	1.772	1.772
	Tau	0.387	0.397	0.431	0.395	0.461	0.461	0.691	0.743
75	Acc	0.679	0.741	0.796	0.802	0.722	0.790	0.833	0.846
	Asd	2.056	1.667	1.605	1.512	2.074	1.858	1.772	1.772
	Tau	0.375	0.363	0.459	0.369	0.464	0.456	0.717	0.756
88	Acc	0.660	0.741	0.772	0.784	0.716	0.796	0.815	0.833
	Asd	2.074	1.691	1.648	1.549	2.068	1.852	1.796	1.778
	Tau	0.419	0.406	0.470	0.408	0.464	0.438	0.743	0.756

Table 93: Spam ranking with FRANK1 on nationality train set without entity

		Maplin		Maplog		Maplin2		Maplog2	
Spam		Full	KB	Full	KB	Full	KB	Full	KB
0	Acc	0.654	0.722	0.809	0.802	0.716	0.784	0.858	0.858
	Asd	2.037	1.747	1.617	1.574	2.025	1.864	1.747	1.759
	Tau	0.426	0.427	0.442	0.428	0.480	0.480	0.709	0.735
50	Acc	0.648	0.716	0.815	0.802	0.710	0.784	0.864	0.864
	Asd	2.031	1.759	1.623	1.586	2.025	1.864	1.753	1.759
	Tau	0.444	0.432	0.463	0.442	0.512	0.500	0.735	0.761
75	Acc	0.698	0.735	0.815	0.802	0.747	0.796	0.858	0.858
	Asd	1.957	1.667	1.599	1.543	2.025	1.870	1.753	1.765
	Tau	0.413	0.374	0.443	0.377	0.513	0.499	0.735	0.774
88	Acc	0.685	0.716	0.815	0.796	0.735	0.784	0.852	0.852
	Asd	2.000	1.673	1.593	1.525	2.037	1.877	1.747	1.759
	Tau	0.402	0.342	0.454	0.351	0.460	0.486	0.735	0.761

Table 94: Spam ranking with FRANK1 on nationality test set with entity

		Maplin		Maplog		Maplin2		Maplog2	
Spam		Full	KB	Full	KB	Full	KB	Full	KB
0	Acc	0.670	0.682	0.706	0.717	0.756	0.783	0.797	0.818
	Asd	2.096	2.025	1.827	1.808	1.904	1.833	1.787	1.742
	Tau	0.487	0.442	0.500	0.444	0.560	0.560	0.793	0.796
50	Acc	0.675	0.682	0.716	0.707	0.761	0.793	0.797	0.813
	Asd	2.081	1.985	1.802	1.803	1.919	1.813	1.787	1.747
	Tau	0.448	0.443	0.469	0.444	0.551	0.571	0.793	0.791
75	Acc	0.690	0.687	0.721	0.707	0.766	0.798	0.797	0.813
	Asd	2.102	2.045	1.807	1.813	1.909	1.838	1.766	1.727
	Tau	0.468	0.473	0.510	0.474	0.554	0.560	0.780	0.778
88	Acc	0.675	0.682	0.711	0.717	0.766	0.788	0.807	0.823
	Asd	2.107	2.035	1.817	1.818	1.883	1.818	1.777	1.737
	Tau	0.474	0.464	0.492	0.465	0.580	0.582	0.799	0.800

Table 95: Spam ranking with FRANK1 on nationality train set without entity

Spam		Maplin		Maplog		Maplin2		Maplog2	
		Full	KB	Full	KB	Full	KB	Full	KB
0	Acc	0.695	0.712	0.721	0.697	0.772	0.818	0.797	0.803
	Asd	2.051	1.995	1.827	1.843	1.893	1.798	1.766	1.753
	Tau	0.479	0.469	0.503	0.467	0.558	0.552	0.798	0.827
50	Acc	0.670	0.682	0.746	0.707	0.751	0.803	0.817	0.818
	Asd	2.030	1.955	1.761	1.818	1.888	1.758	1.731	1.732
	Tau	0.524	0.503	0.567	0.506	0.551	0.551	0.812	0.842
75	Acc	0.680	0.697	0.751	0.727	0.756	0.803	0.827	0.833
	Asd	2.030	1.980	1.746	1.828	1.904	1.768	1.716	1.712
	Tau	0.538	0.507	0.546	0.506	0.619	0.609	0.781	0.800
88	Acc	0.716	0.732	0.751	0.737	0.787	0.823	0.827	0.833
	Asd	1.954	1.879	1.731	1.808	1.863	1.737	1.701	1.702
	Tau	0.506	0.474	0.548	0.475	0.551	0.562	0.781	0.810

D.2 Varying n-gram lengths

Table 96: N-gram length with FRANK1 on profession train set with entity

N-gram		Maplin		Maplog		Maplin2		Maplog2	
		Full	KB	Full	KB	Full	KB	Full	KB
4	Acc	0.596	0.639	0.687	0.670	0.683	0.726	0.711	0.718
	Asd	2.402	2.159	1.887	1.977	2.109	1.973	1.996	2.008
	Tau	0.454	0.380	0.399	0.381	0.519	0.448	0.523	0.565
5	Acc	0.588	0.627	0.678	0.662	0.682	0.717	0.701	0.709
	Asd	2.408	2.190	1.895	1.967	2.111	2.004	1.990	1.996
	Tau	0.445	0.387	0.398	0.390	0.506	0.452	0.525	0.561
6	Acc	0.586	0.619	0.676	0.643	0.682	0.709	0.699	0.695
	Asd	2.437	2.216	1.915	2.017	2.118	2.012	1.994	2.029
	Tau	0.451	0.399	0.420	0.398	0.518	0.449	0.524	0.565
7	Acc	0.588	0.614	0.672	0.643	0.680	0.703	0.693	0.687
	Asd	2.462	2.291	1.957	2.039	2.130	2.050	2.031	2.056
	Tau	0.435	0.406	0.422	0.408	0.517	0.447	0.525	0.551

Table 97: N-gram length with FRANK1 on profession train set without entity

N-gram		Maplin		Maplog		Maplin2		Maplog2	
		Full	KB	Full	KB	Full	KB	Full	KB
4	Acc	0.579	0.643	0.707	0.678	0.662	0.728	0.738	0.720
	Asd	2.373	2.054	1.835	1.901	2.118	1.959	1.920	1.944
	Tau	0.377	0.333	0.372	0.339	0.500	0.410	0.504	0.580
5	Acc	0.590	0.647	0.711	0.683	0.674	0.734	0.744	0.728
	Asd	2.338	2.045	1.821	1.895	2.117	1.963	1.903	1.942
	Tau	0.367	0.316	0.356	0.324	0.510	0.410	0.465	0.572
6	Acc	0.604	0.658	0.713	0.680	0.682	0.740	0.744	0.726
	Asd	2.307	2.025	1.841	1.911	2.103	1.951	1.918	1.957
	Tau	0.355	0.315	0.350	0.321	0.506	0.407	0.464	0.565
7	Acc	0.612	0.662	0.703	0.676	0.689	0.744	0.732	0.720
	Asd	2.250	1.990	1.872	1.932	2.078	1.932	1.951	1.979
	Tau	0.347	0.308	0.354	0.312	0.490	0.403	0.493	0.589

Table 98: N-gram length with FRANK1 on profession test set with entity

N-gram		Maplin		Maplog		Maplin2		Maplog2	
		Full	KB	Full	KB	Full	KB	Full	KB
4	Acc	0.596	0.674	0.715	0.698	0.700	0.768	0.752	0.750
	Asd	2.322	2.051	1.834	1.838	2.121	1.963	1.916	1.916
	Tau	0.397	0.381	0.383	0.379	0.486	0.412	0.515	0.569
5	Acc	0.608	0.669	0.708	0.688	0.708	0.764	0.749	0.745
	Asd	2.302	2.070	1.828	1.838	2.113	1.975	1.906	1.914
	Tau	0.409	0.387	0.400	0.382	0.504	0.430	0.515	0.583
6	Acc	0.589	0.641	0.702	0.684	0.702	0.745	0.747	0.743
	Asd	2.349	2.138	1.852	1.865	2.109	1.994	1.926	1.930
	Tau	0.430	0.398	0.420	0.390	0.506	0.449	0.540	0.584
7	Acc	0.581	0.634	0.688	0.674	0.690	0.735	0.735	0.733
	Asd	2.374	2.168	1.885	1.910	2.127	2.006	1.940	1.957
	Tau	0.426	0.400	0.418	0.393	0.520	0.481	0.507	0.583

Table 99: N-gram length with FRANK1 on profession test set without entity

N-gram		Maplin		Maplog		Maplin2		Maplog2	
		Full	KB	Full	KB	Full	KB	Full	KB
4	Acc	0.596	0.671	0.696	0.692	0.694	0.764	0.733	0.741
	Asd	2.335	2.031	1.875	1.873	2.129	1.947	1.955	1.953
	Tau	0.369	0.354	0.368	0.346	0.495	0.418	0.515	0.573
5	Acc	0.589	0.671	0.706	0.692	0.688	0.766	0.737	0.737
	Asd	2.329	2.014	1.854	1.844	2.140	1.949	1.936	1.936
	Tau	0.366	0.341	0.349	0.334	0.493	0.404	0.496	0.583
6	Acc	0.593	0.674	0.694	0.680	0.692	0.768	0.727	0.727
	Asd	2.304	1.990	1.879	1.869	2.133	1.942	1.947	1.961
	Tau	0.364	0.347	0.362	0.335	0.493	0.406	0.507	0.618
7	Acc	0.604	0.686	0.667	0.661	0.696	0.772	0.708	0.712
	Asd	2.275	1.949	1.920	1.928	2.131	1.940	1.986	2.004
	Tau	0.377	0.339	0.362	0.336	0.510	0.414	0.526	0.633

Table 100: N-gram length with FRANK1 on nationality train set with entity

N-gram		Maplin		Maplog		Maplin2		Maplog2	
		Full	KB	Full	KB	Full	KB	Full	KB
4	Acc	0.654	0.741	0.796	0.802	0.698	0.790	0.833	0.846
	Asd	2.093	1.704	1.586	1.525	2.099	1.858	1.772	1.772
	Tau	0.413	0.396	0.441	0.392	0.459	0.459	0.678	0.743
5	Acc	0.679	0.741	0.796	0.802	0.722	0.790	0.833	0.846
	Asd	2.056	1.667	1.605	1.512	2.074	1.858	1.772	1.772
	Tau	0.375	0.363	0.459	0.369	0.464	0.456	0.717	0.756
6	Acc	0.685	0.759	0.796	0.796	0.728	0.802	0.833	0.840
	Asd	1.969	1.617	1.586	1.506	2.031	1.846	1.772	1.778
	Tau	0.392	0.354	0.420	0.356	0.455	0.460	0.743	0.756
7	Acc	0.673	0.765	0.796	0.796	0.716	0.809	0.833	0.840
	Asd	1.926	1.586	1.574	1.494	2.006	1.827	1.765	1.772
	Tau	0.418	0.360	0.419	0.367	0.461	0.486	0.743	0.756

Table 101: N-gram length with FRANK1 on nationality train set without entity

N-gram		Maplin		Maplog		Maplin2		Maplog2	
		Full	KB	Full	KB	Full	KB	Full	KB
4	Acc	0.648	0.728	0.790	0.796	0.716	0.790	0.846	0.852
	Asd	2.074	1.698	1.630	1.543	2.086	1.901	1.759	1.759
	Tau	0.367	0.347	0.440	0.348	0.495	0.468	0.737	0.756
5	Acc	0.698	0.735	0.815	0.802	0.747	0.796	0.858	0.858
	Asd	1.957	1.667	1.599	1.543	2.025	1.870	1.753	1.765
	Tau	0.413	0.374	0.443	0.377	0.513	0.499	0.735	0.774
6	Acc	0.704	0.735	0.802	0.784	0.753	0.796	0.846	0.846
	Asd	1.864	1.599	1.593	1.543	1.969	1.821	1.759	1.765
	Tau	0.405	0.392	0.456	0.403	0.470	0.482	0.761	0.774
7	Acc	0.716	0.753	0.796	0.784	0.753	0.802	0.840	0.840
	Asd	1.802	1.599	1.605	1.562	1.920	1.821	1.778	1.778
	Tau	0.418	0.406	0.479	0.420	0.495	0.521	0.787	0.787

Table 102: N-gram length with FRANK1 on nationality test set with entity

N-gram		Maplin		Maplog		Maplin2		Maplog2	
		Full	KB	Full	KB	Full	KB	Full	KB
4	Acc	0.680	0.677	0.736	0.722	0.761	0.798	0.812	0.833
	Asd	2.112	2.061	1.772	1.783	1.924	1.843	1.751	1.712
	Tau	0.477	0.473	0.493	0.472	0.534	0.530	0.761	0.763
5	Acc	0.690	0.687	0.721	0.707	0.766	0.798	0.797	0.813
	Asd	2.102	2.045	1.807	1.813	1.909	1.838	1.766	1.727
	Tau	0.468	0.473	0.510	0.474	0.554	0.560	0.780	0.778
6	Acc	0.670	0.687	0.711	0.707	0.756	0.798	0.797	0.813
	Asd	2.122	2.040	1.832	1.808	1.919	1.828	1.772	1.732
	Tau	0.468	0.462	0.520	0.464	0.577	0.607	0.791	0.795
7	Acc	0.660	0.687	0.721	0.712	0.751	0.793	0.802	0.813
	Asd	2.127	1.955	1.807	1.773	1.914	1.808	1.761	1.732
	Tau	0.504	0.421	0.526	0.422	0.571	0.597	0.812	0.816

Table 103: N-gram length with FRANK1 on nationality train set without entity

N-gram		Maplin		Maplog		Maplin2		Maplog2	
		Full	KB	Full	KB	Full	KB	Full	KB
4	Acc	0.655	0.667	0.731	0.722	0.741	0.778	0.827	0.833
	Asd	2.066	1.995	1.761	1.803	1.919	1.808	1.721	1.712
	Tau	0.493	0.483	0.518	0.487	0.590	0.589	0.780	0.799
5	Acc	0.680	0.697	0.751	0.727	0.756	0.803	0.827	0.833
	Asd	2.030	1.980	1.746	1.828	1.904	1.768	1.716	1.712
	Tau	0.538	0.507	0.546	0.506	0.619	0.609	0.781	0.800
6	Acc	0.690	0.702	0.741	0.717	0.756	0.803	0.817	0.823
	Asd	2.041	1.965	1.787	1.869	1.893	1.763	1.736	1.732
	Tau	0.549	0.528	0.573	0.526	0.635	0.624	0.811	0.830
7	Acc	0.675	0.687	0.741	0.712	0.741	0.793	0.822	0.828
	Asd	2.010	1.970	1.787	1.879	1.909	1.783	1.726	1.727
	Tau	0.547	0.516	0.582	0.519	0.658	0.647	0.811	0.830