Information Retrieval Project 2, Group 11

Michael Merki & Markus Greiner - December 2016

# Installation

See README.md how to install and run using sbt.

# Preprocessing / Tokenizing

Preprocessing and tokenizing posed two challenges:

* remove tokens that will not contribute, such as stopwords, underlines or numbers.
* correcting spelling mistakes

## Tokenizing

The content of the document is **split**. We enhanced the splitting function of Tipster (which, by the way, contained an error: ‘-‘ inside ‘[]’ will not match a dash but indicates a range), to also include brackets, all kind of quotes, underscore, and line-breaks (which appeared in titles).

**val** *Split* = """[ .,;:?!\*&$\-+\_/\s\(\)\[\]\{\}'\"\x60\n\r]+"""  
**val** *rDate* = "^\\d+[/-]\\d+[/-]\\d+$".r -> "<DATE>"  
**val** *rUSPhone* = "^\\d{3}\\W\\d+{3}\\W\\d{4}$".r -> "<USPHONE>"  
**val** *rNumber* = "^[-]?\\d+([.,]\\d+)\*$|^(one|two|three|four)$".r -> "<NUMBER>"  
**val** *rTwoNum* = "^\\d+[-/=]\\d+$".r -> "<NUMBER>"  
**val** *rOrdinal* = "^\\d+(th|1st|2nd|3rd)$".r -> "<ORDINAL>"  
**val** *rLine* = "--+".r -> "" // underlines like -----------  
**val** *rQuote* = """['\"\x60]+""".r -> "" // also ` = x60

After splitting, various number formats are removed, as are lines of dashes. All this is done using regular expressions which are static variables so that they only have to be computed once.

**Stop words** are also removed. They are a combination of the stop word provided in TinyIR, in the NLTK, and some observed most frequent words in the collection.

The TinyIR **Porter Stemmer** is also used. A simpler approach (chopping terms longer than length 6) did not improve the results.

## Spelling correction

The provided collection contains many spelling mistakes, and the most frequent observed one was concatenation of words by leaving out blanks. Several pre-processing steps were applied to improve the results.

**val** *MidCap* = "(.+)([A-Z].\*)".r  
**val** *patIonal* = "(.+ionally|.+ional|.+ions?)(.{3,})".r  
**val** *patIng* = "(.+ingly|.+ing|.+iously|.+ious|.+ment)(.{3,})".r  
**val** *patThe* = "(their|they|there|these|the|those|without|with|was|and|have)(.+)".r  
**val** *patEd* = "(.{4,}ed)(.{4,})".r  
**val** *patAnd* = "([Aa]nd)([^aeiory].{3,})".r  
**val** *patTheS* = "([Tt]he)([dhjpvxz].{3,})".r

Splitting of long words: The English language has an average word length of 8.23 characters[[1]](#footnote-1) and only 1% of the words contain 15 or more characters. So detecting a token of length > 15 very likely indicates two or more concatenated words. To split them, we took the following approaches:

* A capital letter inside a token indicates a new word
* Typical word endings inside a token: -ion, -ions, -ionally, -ional, -ing, -ious, -iously, -ment
* Most frequent words at the beginning of the token: there, those, have, …
* If this fails, we look for –ed followed by at least 4 more characters.

These measures already lead to very small F1 improvement. However, it was still observed that there are many misspellings connected to “**the**” and “**and**” at the beginning of a token even for tokens of length < 15.

Using a dictionary (manually), we discovered that in the English dictionary the following rules:

* “and” is always followed by a, e, I, o, r, or y.
* “the” is never followed by d, h, p, v, z, z

These rules can be applied using regular expressions to split shorter tokens.

|  |  |  |  |
| --- | --- | --- | --- |
| Collection size | ‘and’ splitting | ‘the’ splitting | Long-words splitting |
| 40’000 docs | 22’858 | 13’278 | 67’086 |
| 100’000 docs | 187’797 | 48’684 | 405’179 |

## Title detection

The collection contains documents from various sources AP, DO, FR, etc. Some of these documents contain a title. The intuition was that the terms of the title can be used in the tf scores with a higher weight (ex 5 times).

To detect titles in documents, the following patterns were used. When there was not title available, the first 100 characters of the content was used. Small improvements were detected.

name.substring(0, 2) **match** {  
 **case** "AP" => read(*doc*.getElementsByTagName("HEAD")) // AP  
 **case** "DO" => *bodyStart* **case** "FR" => *bodyStart* // read(doc.getElementsByTagName("ITAG"))// FR need tag 7  
 **case** "PT" => read(*doc*.getElementsByTagName("TTL")) // PT  
 **case** "SJ" => read(*doc*.getElementsByTagName("HEADLINE")) // SJM  
 **case** "WS" => read(*doc*.getElementsByTagName("HL"))  
 **case** "ZF" => read(*doc*.getElementsByTagName("TITLE"))  
 **case** \_ => *bodyStart*

# Term-based scoring models

Our term-based model consists of two different scoring approaches. On the one hand we implemented the scoring based on tf-idf weights described in the lecture (lecture 4, slide 14). In a first attempt we directly used raw term frequencies which resulted in a bias towards longer documents. In order to avoid such biases we extended the standard tf-idf weighting scheme by using the augmented term frequency instead of the raw term frequency. The augmented term frequency was calculated according to the following formula:

With this extension we were able to increase the performance metrics of ranking the provided training documents for the 40 test queries as shown in the following table:

|  |  |  |
| --- | --- | --- |
| Metrics | TFIDF with raw tf | TFIDF with augmented tf |
| Precision | 0.326 | 0.411 |
| Recall | 0.341 | 0.427 |
| F1 Score | 0.331 | 0.417 |
| MAP | 0.217 | 0.296 |

Table 1: Comparison between raw term frequency and augmented frequency weighting

On the other hand we used the idea of the vector space model described in the lecture (lecture 4, slides 15 – 17) and computed the cosine similarity between the documents and queries represented as vectors.

Both scoring approaches can be found in the class "TermBasedModel" and can be run with and without inverted index. In case of running the scoring with indexing we compute the needed extra information like the maximum frequency per document (in case of the augmented tf-idf weighting) or the norms of the document vectors (in case of the cosine similarity measure) directly based on the information available in the inverted index. In case of not using indexing we make two entire iterations over all documents. In the first round we precompute document frequencies for all terms in order to be able to compute the tf-idf score in the second round. During the second iteration over the document collection we compute the score for each document and keep a sequence of the top-n ranked documents.

# Language scoring models

As our language model we implemented the maximum likelihood estimation presented in the lecture (lecture 6, slides 5-8) and applied the proposed Jelinek-Mercer smoothing to get better estimates for small frequencies. Experimenting with varying the tuning parameter λ showed optimal precision, recall, f1 and MAP measures for λ = 0.2.

The implementation of the language model can be found in the class "LanguageModel". It is possible to run our language model with and without indexing. In case of using the inverted index the required information like maximum frequency per document is directly computed based on the index. For the variant without indexing we make two entire iterations over the document collection. During the first round we precompute collection frequencies as well as the total amount of terms in the document collection in order to be able to apply the described scoring with smoothing in the second iteration. Like in the case of the term-based models we just keep the top-n ranked documents after each iteration of the second round.

# Training data performance

Evaluating the performance of our three scoring models for the provided 40 test queries and 100'000 training documents resulted in the following precision, recall, f1 score and mean average precision (MAP) metrics:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | No pre-processing | | | Stop word, stemming | | | Stop words, stemming, splitting | | |
| Metrics | TFIDF | Cosine | Language | TFIDF | Cosine | Language | TFIDF | Cosine | Language |
| Precision | 0.379 | 0.316 | 0.402 | 0.400 | 0.342 | 0.407 | 0.411 | 0.331 | 0.425 |
| Recall | 0.392 | 0.330 | 0.418 | 0.415 | 0.355 | 0.421 | 0.427 | 0.345 | 0.441 |
| F1 Score | 0.384 | 0.321 | 0.408 | 0.405 | 0.346 | 0.412 | 0.417 | 0.336 | 0.431 |
| MAP | 0.262 | 0.169 | 0.282 | 0.277 | 0.198 | 0.280 | 0.296 | 0.179 | 0.303 |

Table 2: Precision, recall, f1 score and MAP metrics for our three scoring models

Note: For computation of recall and MAP we bounded the denominator by

# Approaches for improving training data performance

We implemented several extensions to the above described scoring models with the goal to improve the training data performance.

## Character n-grams

Character n-grams: One of the major challenges when working with the tipster data set is in our opinion spelling errors. In order to become more tolerant against such spelling mistakes we tried to create an inverted index with character n-grams. Our class "TipsterParseSmart" has an option "ngramSize" which allows specifying the size of the n-grams to be extracted from the document tokens. Testing various n-gram sizes unfortunately didn't result in improving the training data performance.

## Query expansion[[2]](#footnote-2)

Query expansion through synonyms/terms from top ranked documents: Because the provided training queries were rather short (most often just a couple of terms) we had the idea to use query expansion to improve training data performance. Our first approach was based on doing a first round of scoring and only returning the top n ranked documents with n being quite small (e.g. 10). From these "very" relevant documents we than extracted the most frequent words and appended them to the query terms. After that we did a second round of scoring and returned the top 100 documents. Because this didn't lead to an improvement we also tried query expansion by synonyms. For this we used a dictionary with synonyms from the WordNet dataset (data.adj, data.noun, data.verb, data.adv) and extracted the entries with more than one synonym. Before evaluating a query, we extended the original query terms by synonyms from this dataset. Unfortunately, this didn't result in an improvement of the training data performance.

# Challenges during development

Our initial challenge was limited heap size. During all our developments we set maximum allowed heap size to 4 GB which seemed to be too restrictive when working directly with the class "TipsterStream" from the provided library "tinyir". These measures improved the situation:

* don’t keep the tokens in memory, just the term frequencies
* use functions (def) instead of lazy variables (val)
* try to use streams (.toSeq) also for term frequencies.

After solving the problem with the limited heap size the creation of the entire index took around 750 seconds which was not very practical for implementing and testing our scoring models. That was the reason why we made use of "LevelDB" and introduced a class "PersistentFreqIndex" which has the functionality to make the inverted index persistent and recreate the index from disk again. Recreating the inverted index for the entire document collection from the database takes around 50 seconds.

# Running times

The comparison of running the three scoring models with and without the inverted frequency index resulted in the following average running times per query:

| *Running Time* | *TFIDF* | *Cosine Similarity* | *Language Model* |
| --- | --- | --- | --- |
| Creating inverted index | ~ 750 secs | | |
| Recreating inverted index from the database | ~ 50 secs | | |
| Computing additional statistics based on the inverted index [[3]](#footnote-3) | < 20 secs | | |
| Average query running time **with** inverted frequency index | < 1 sec | < 1 sec | < 1 sec |
| Average query running time **without** inverted frequency index | ~ 3650 secs | ~ 2250 secs | ~ 2100 secs |

Table 3: Running times for index creation and running queries with/without index

Note: For these experiments we used a Windows Machine with Intelcore i/ 4500u processor, 1.80 GHz with 8 GB RAM. The maximum heap space was set to 4 GB.

1. <http://www.ravi.io/language-word-lengths> [↑](#footnote-ref-1)
2. internally called „query pimping“ [↑](#footnote-ref-2)
3. This includes calculating statistics like document frequencies which was done directly on the information available in the index (see chapters term-based/language scoring model for more details). Because running this routine didn't take longer than 20 seconds we didn't bother about improving the performance of this step. Obviously ideal would be to also create these statistics directly when creating the inverted index and potentially also make them persistent in LevelDB in order to faster respond to queries. [↑](#footnote-ref-3)