Group 1

Robin Helbig, ist1105007

robin.helbig@tecnico.ulisboa.pt

Karolina Kowalczyk, istxxxxxx

author’s e-mail address

author name, istxxxxxx

author’s e-mail address

# GENERAL DECISIONS

In the beginning of our report, we want to explain some general decisions that we made when developing our information retrieval system. This should help with understanding our code and the results it generates.

## STRUCTURE

We split the main functions that were listed in the project assignment (indexing, ranking, visualize, evaluation) into four files that each handle one functionality. We adjusted the parameters that were supposed to be passed to the functions slightly because some did not make that much sense for us. For example, the visualization function did not benefit from the summary maximum sentence and char length or the inverted index.

Besides the main functions we created data classes and some helper functions to support the main functionality.

## PREPROCESSING

For preprocessing we decided to remove terms that were just consisting of punctuation (e. g. “,”) because *nltk* seemed to include them in the output of the *word\_tokenize* function.

We extracted noun phrases and bigrams and added them to the list of terms to increase the number of informative terms that could be used for sentence evaluation. Furthermore, we decided to remove stopwords to increase the information density of sentences. We also turned every term lowercase to increase comparability between sentences.

## INVERTED INDEX

For the inverted index we decided to determine the inverted document frequencies of terms regarding the whole news article collection. We did that because it includes many samples that allow our system to deem words as more or less important when it comes to news. Only considering one article for that could have downsides. If we for example take a news article about renewable energies the term “energy” might arise very often. This does not mean though, that it is a term that is used a lot in the English language and that it therefore should be considered less for the evaluation. It might even be a very important term in the article that should have a significant part in the summary.

For the term frequencies we had to build a second inverted index per document. This is trivial because otherwise an evaluation per sentence would not have been possible.

## SUMMARIZATION

In general, we create the summaries by choosing the sentences that are most relevant to the document regarding a certain score.

In the preprocessing step we generate a list of terms for every sentence in the document which then can be used as a base for generating different scores.

We implemented the following five approaches for score generation. They can be chosen via the *args* parameter of the ranking function.

### TF

Do determine a TF (term frequency) based score for sentences or the whole document we went through the previously mentioned list of terms and counted the occurrences of every term. After that we diminished the impact of having the same term multiple times in the same sentence with the help of a logarithmic function:

To not weigh long sentences unsuitably strong we normalized the term frequencies values afterwards. To get the final score we multiplied the term frequencies of each sentence with the ones of the document.

### TF-IDF

The TF-IDF score was calculated like the TF score, except that all TF values were multiplied with the inverted document frequencies of the term in the corpus that were already mentioned before.

### BM25

We used the formular that is used to calculate the BM25 score for a term in a document to calculate it for a term in a sentence. This allowed us to take the sentence length into consideration. This makes sense here because the document length would be constant.

For the total score of the sentence in the document we summed up its BM25 score for each term appearing in the document. Consequently, terms appearing multiple times in the document are weighted stronger. This should have a positive impact though because words that are used often in the news article will probably very central to it.

### RRF

For the RRF score we considered the TF, TF-IDF and BM25. To include the TF might not be necessary because it is also included in the TF-IDF and BM25 already. A comparison of the four different RRF combinations that are possible with three scores was not conducted in the scope of this project though.

### MMR

To calculate the MMR, we used the cosine similarity provided by the multiplication of the TF-IDF values of the compared strings.

If lambda is close to 0 the MMR score is more focused on finding sentences that are similar to the document. If lambda is close to 1 it’s foremost relevant for a sentence to not be redundant in comparison to the already chosen summary sentences. As a compromise, we chose to set lambda to 0.5 for the start before exploring the impact of the value more closely later on.

## VISUALIZATION

Hubert

## EVALUATION

Karolina

# QUESTIONS

Here come the questions

## PREPROCESSING TERM IMPACT

Characterize the corpus D and summaries R. What is the distribution of informative terms before and after text processing?

Answer Hubert

## GENERAL PERFORMANCE

How does the developed summarization system perform for the full collection? And within each category? Any intuition for the observed differences?

Answer Karolina

## MMR λ

Consider the Maximal Marginal Relevance (MMR) stance. How λ impacts the redundancy and accuracy (against ideal extracts) of the summaries? Should λ be a fixed threshold or depend on the provided topic document (d-specific)?

Answer Robin

## LENGTH THRESHOLDS

Given specific length thresholds (p and l), is the system better at providing recall or preci- sion guarantees? How would you fix the length thresholds if the user prefers: minimizing false positives, minimizing false negatives, or maximizing true positives?

Answer Karolina

## PREPROCESSING EVALUATION

Does the inclusion of phrases and bigrams produces a positive impact on summarization?

Answer Robin

## IR MODELS

How IR models affect retrieval? Is Reciprocal Rank Fusion (RRF) useful to aid decisions?

Answer Robin

## LENGTH BIAS

Lengthier sentences are more verbose and therefore have higher chance of being judged as relevant for a randomly selected topic. If this is an undesirable bias of your summarization system, how would you extend your system to handle this bias?

Answer Hubert