Information Processing and Retrieval Project Report – Part 1

Instituto Superior Técnico – Universidade de Lisboa

Margarida Costa  
 83425

Marta Aparício  
 83525

Paulo Alves  
 83538

ABSTRACT

This report focuses on automatic keyphrase extraction, applied to two datasets: the 20-newsgroup collection and the Inspec dataset. To accomplish such task the TF-IDF and the BM25 information retrieval model are used, followed by an evaluation of the models. Lastly classifiers (…) are used.

Introduction

To address the problem – automatic keyphrase extraction – two main alternatives are explored: a simple approach based on TF-IDF; and a supervised approach.

The simple approach based on TF-IDF has three phases: First is implemented a simple baseline approach, that applies the keyphrase extraction method to an English textual document; Secondly its implementation is evaluated, based on metrics such as precision, recall, F1 measure, mean value for the precision@5 and mean average precision; Lastly, is improved by candidate selection, candidate scoring and other ideas where tested.

The supervised approach resorts to classification algorithms and a set of features (USED FEATURES), the results are then evaluated by (CONFUTION MATRIX AND SO ON).

The 20-newsgroup collection is only used for the baseline implementation of the simple approach based on TF-IDF. For the rest of approaches, the Inspec dataset is used.

1 Simple approach based on TF-IDF

1.1 Implementation

The libraries “sklearn” and “scipy” were used, together with the documents from the 20newsgroups collection. Only one document is being considered for test and the train was made over the whole 18000 documents and for just 30 documents, with the purpose of analyzing the differences.

When ***TfidfVectorizer***is being applied to the training set, the conversion of raw text to a TFIDF features matrix is being made. For better fit the purpose of the case study the following parameters were changed: ***ngram\_range(1, 3)***; the ***stop\_words=’english’***; the ***token\_pattern=r"(?u)\b[a-zA-Z][a-zA-Z-]\*[a-zA-Z]\b"*** that selects only single words or words followed by a hyphen and other words followed by hyphens. Lastly the ***max\_df***, ***min\_df*** and ***norm*** will be used on the exercise 2 for fine tuning. After applying the ***fit\_transform*** method, to all the documents, the vocabulary and the IDF scores are computed based on all the documents allowing for a full coverage of all words, making sure that the vocabulary also contains the words of the test document. Once this is done the ***transform***method is applied to the test document alone, making a TF-IDF-weighted matrix, where the line is the test document and the columns are the terms of the model.

Then, the scores of each term of the matrix are calculated by multiplying the TF-IDF value of the term by the number of words that it has (e.g. if the term is a tri-gram it would be *TF-IDF value\* 3*) or by the number of characters that the words have. Finally, the results, from this operation, are sorted and only the 5 most important results, thus candidates, are returned.

1.2 Results

The success or unsuccess of this exercise is undetermined since it is not possible to apply metrics. For the calculation of the final scores the multiplication by the number of words of the n-gram was the chosen option, since it gives more importance to the grams that have more words, thus the grams that will probably describe better the document.

The usage of the whole set of train documents played an important role on the results, since they would be more general when all the 18000 documents were included ([‘cms udel edu’, ‘phys psu edu’, ‘dev-null phys psu’ ‘espn’, ‘bed angry’]). However, when only one category of that set was used, the results included more words that defined better what the actual test document was about (['espn', 'psu edu', 'organization penn state', 'hockey', 'run']).

2 Evaluating the simple approach

2.1 Implementation

For evaluating the previous retrieval model, these documents are split into two sets, one train (80% of the documents) and other test (20% of the documents). After this sets are cleaned of xml tags; the train dataset is fitted using the same approach then before and the candidates keyphrases are calculated from the test dataset. Then the target keyphrases are obtained, so the predicted candidates can be compared and evaluated. The evaluation is done by the ***metrics(y\_true, y\_pred)***function, that calculates the f1-measure and the precision and recall measures, followed by the calculation of the mean average precision (MAP) value and the mean precision@5 (MP5) for the entire test collection.

1.2 Results

Different approaches were tested. First, an approach where ***TfidfVectorizer*** used the original word from (e.g. studying) the text in order to calculate the candidates, was implemented. Then this was changed so the ***TfidfVectorizer***used the word lemmas (e.g. study) instead, and the obtained results improved in comparison to the previous approach. This, because the words were reduced to the form of the word that is chosen by convention. Lastly, instead of the utilization of lemmas, the word stems (e.g. stud) were calculated and used by the ***TfidfVectorizer****.* After each approach, fine tuning was done with the ***max\_df***, ***min\_df*** = 1 and ***norm*** = ‘l2’ parameter of the ***fit*\_*transform***method, improving some of the results.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Word  (max\_df = 3) | Lemma  (max\_df = 2) | Stem  (max\_df = 9 ) |
| MAP | 0.122 | 0.112 | 0.167 |
| MP5 | 0.054 | 0.036 | 0.060 |

3 Evaluating the simple approach

The updated template, user manuals, samples, and required fonts, all are available at the URL <https://www.acm.org/publications/proceedings-template>. It contains said information for all three versions of MS Word (Windows and 2 versions of Mac). There are also separate links to the user guide, which can be referred to by the user. This URL also contains some useful video links, which describe how to add the template, structure the paper, and generate the layout, in different clips. **Display Formula with Number**

 (1)

**Continuation part of Paragraph Text** The user must style this paragraph in **ParaContinue** style, which follows immediately after the **DisplayFormula** (numbered equation). The **DisplayFormula** style is applied only in case of a numbered equation. A numbered equation always has a number to its right. Insert paragraph text here. **Display Formula without Number**



The **DisplayFormulaUnnum** style is applied only in case of an unnumbered equation. An unnumbered display equation never contains an equation number to its right, and this unique property distinguishes it from a numbered equation.



Figure 1: Figure Caption and Image above the caption [In draft mode, Image will not appear on the screen]

**Theorem/Proof/Lemma.** Insert text here for the enunciation or Math statement. Insert text here for the enunciation or Math statement. Insert text here for the enunciation or Math statement. Insert text here for the enunciation or Math statement. Insert text here for the enunciation or Math statement.

....Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract.

1.1 Heading Level 2

In the below paragraph, it is explained how alt-txt value is placed in **MS Word 2010**. To add alternative text to a picture in Word 2010, follow these steps:

1. In a Word 2010 document, insert a picture.
2. Right click on the inserted picture and select the **Format Picture** option.
3. Select the **Alt Txt** option from the left-side panel options.
4. In the "Title:" and "Description:" text boxes, type the text you want to represent the picture, and then click "Close".

Below are steps to place alt-txt value in **MS Word 2013/2016**. To add alternative text to a picture in Word 2013/2016, follow these steps:

1. In a Word 2013/2016 document, insert a picture.
2. Right click on the inserted picture and select the **Format Picture** option.
3. In the settings at the right side of the window, click on the "Layout & Properties" icon (3rd option).
4. Expand **Alt Txt** option.
5. In the "Title:" and "Description:" text boxes, type the text you want to represent the picture, and then click "Close".

*1.1.1 Heading Level 3.* Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here.

*1.1.1.1 Heading Level 4.*Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here.

CONSTRUÇÃO

Por exercise:

* 1. How was implemented (justify decision taken best parameters,…)
  2. Results (examples of success and unsuccess (why it happen)) [graphics//confusion matrix…]

Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here.

REFERENCES

[1] FNM Surname (2018). Article Title. Journal Title, 10(3), 1–10.

[2] F.N.M Surname, Article Title, https://www.acm.org/publications/proceedings-template.

[3] F.N.M Surname and F.N.M Surname, 2018 Article Title, The title of book two (2nd. ed.). Publisher Name, City, State, Country.

[4] Ian Editor (Ed.). 2018. *The title of book two* (2nd. ed.). University of XXX Press, City, Chapter 100. DOI: <http://dx.doi.org/10.1000/0-000-00000-0>.

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00