Unsupervised Approaches for Keyword Extraction using word2vec

Sonu Agarwal, IITK and Viveka Kulharia, IITK

March 18, 2015

1 Proposal

Here we propose various unsupervised keyword extraction methods. Keyword extraction means extracting words/phrases that can describe the meaning of the document. It's important for many text mining applications as it helps in indexing and thus classifying the document.

We need automatic keyword extraction techniques as manual assignment of high quality keywords is expensive and time consuming. Here we focus on unsupervised approaches, as in many cases not enough number of documents are labelled/tagged for supervised methods to learn.

2 Recent Works

Till now, most of the research has focused on using supervised algorithms like SVM which has given pretty good results. In unsupervised learning, most of the work has been done by using wordnet, Conditional Random Fields and by assessing the correlation among words.

3 Approach

In our project we aim to implement at least two methods.

3.1 Method 1

In this method, we will be using statistical Information on co-occurrence of words for keyword extraction. Two terms in a sentence are considered to co-occur once. For a document, frequency of each term is calculated and then few most frequent terms are taken into account. After this co-occurrence of each term with these frequent terms are calculated.

Implementation of this algorithm consists of six steps:[1]

1. Pre-processing
2. Selection of frequent terms
3. Clustering frequent terms

4. Calculation of expected probability
5. Calculation of X'^2 value which is a measure of bias
6. Output Keywords

In the paper, two clustering methods are mentioned which are similarity-based clustering and pairwise clustering. We propose to use **word2vec**[2] method for clustering purpose.

3.2 Method 2

Here, we will be implementing TFIDF(term frequency, inverse document frequency) algorithm[3]. The concept is that frequency, in a text, of the overall least frequent word determines the nature of the text. Here, the frequency of the similar words can also be taken into account. So, we can use the **word2vec** to find the cosine distances between different words and then use their combined frequency to get the result.

3.3 Evaluation

To compare the accuracy of these two methods, we intend to use F-measure (F_1 score). To compare phrases, we will break them into words and then compare the words

4 Datasets

References

[1] Y. Matsuo and M. Ishizuka. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13:2004, 2004.

[2] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546, 2013.

[3] Feifan Liu, Deana Pennell, Fei Liu, and Yang Liu. Unsupervised approaches for automatic keyword extraction using meeting transcripts. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 620–628, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.