Heidelberg University
Institute of Computer Science
Database Systems Research Group

Project Proposal for the lecture Text Analytics

Clustering of scientific papers for easy information retrieval

Team Member: Daniela Fichiu, Matriculation Number, Course of Study
email address
Team Member: Christian Homeyer, 3606476, PhD Computer Science
ox182@uni-heidelberg.de
Team Member: Jessica Kaechele, 3588787, MSc Applied Computer Science
Uo251@stud.uni-heidelberg.de
Team Member: Jonas Reinwald, 3600238, MSc Applied Computer Science
am248@stud.uni-Heidelberg.de

1 Motivation

Motivate your project and state the *real-world problem* you want to solve.

2 Research Topic Summary

In order to get an overview of a research field without extensive research and reading countless papers, there are various approaches to simplify this. By browsing through digital libraries and search engines you can find papers by matching search strings, but to get an overview of an entire research field this is not enough. In [7] a tool is presented that uses different methods to gain insights into research fields. Through the citation network, papers can be divided into clusters. Furthermore, trends, gaps and outliers can be identified. Another possibility to get insights is the citation context, because the key statements of the paper are often summarized concisely. To know the key statements of the paper, it is still necessary to read the whole paper or at least all citation contexts. For many papers this is still a very large amount of work. To solve this problem Multi-Document Summarization is used here. This summarization is only applied to abstract and citation context. The dataset used is ACL Anthology Network (AAN).[1] The dataset contains the network of citations, as well as the full text of each article, its metadata, summary, references and citation sentences.

With the help of the techniques mentioned, it should be possible to gain insight into an entire research field more quickly. In this project we will concentrate on only one method: the clustering. For many years, attempts have been made to cluster papers in order to simplify research. In 1973, for example, it had already been tried to cluster journals by comparing reference patterns and looking at mutual references [6].

In [5] the context of the citations is used in addition to the citations to cluster. First a citation has to be recognized and the text has to be extracted on both sides of the citation. Then link-based clustering approaches, term-based clustering approaches and hierarchical document clustering, as well as a combination of all three, are applied and compared. In addition, this technique is also applied to the entire document and compared to the approach of citation context.

But there are also approaches where citations are not used. In [8] abstracts and titles are used. For this purpose the bibliography with various queries is downloaded from Web of Science [3]. Two types of pre-processing are then performed on the texts. The first method treats each word as a token, and stopwords are deleted. The second method uses term-clumping

to find noun phrases with significant commonality. In addition, several topic modeling algorithms are used. The Latent Dirichlet allocation(LDA), Correlated Topic Models (CTM), Hierarchical Latent Dirichlet Allocation (Hierarchical LDA) and Hierarchical Dirichlet Process (HDP) are tested. The clusters created by the algorithms have to be named manually. Abstracts are also used in [4] to perform clustering. First tokenization is used, then stopwords are removed and a stemming algorithm is performed. Because of the shortness of abstracts, the words must have a higher frequency than in a general balanced corpus of the given language. Then the keywords are grouped and weighted and the closeness of two documents is calculated using cosine measure. Additionally, clustering methods are applied to the whole abstract. Three algorithms from three different approaches are used: the k-medoid method from the example-based approach, the nearest neighbor method from the hierarchy-based approach and the MajorClust method from the density-based approach. As data source 48 abstracts from [2] are used, which have been classified by a human.

3 Evaluation

?? In this section, several evaluation approaches will be presented. Furthermore the whole processing pipeline is illustrated.

3.1 Goal

What is the goal of this work? Do we do unsupervised clustering of papers? Then we will have different evaluations on cluster statistics, number of clusters etc.

Do we do supervised clustering/classification? Then we could compute something like a cross entropy loss, Kullback-Leibler divergence, etc.

I remember that our original idea was to utilize the existing key words to supervise our feature pipeline/ clustering algorithm. That means, that we define broader categories before and then compare relationships that we identify with the ones from the key words? We could also compute a similarity measure based on our ground truth and then compute a difference with the prediction.

3.2 What evaluation losses are possible?

When evaluating, we mean the direct comparison between our work and other works or the comparison of a ground truth and predictions. What metrics can be used for the task of key word generation? What losses are possible to compute for key word classification, general assignment problem?

Unsupervised: Label density, cluster variances, number of clusters, etc.

Supervised: Precision, Recall, F1-Score, KL-divergence, etc.

3.3 Overall processing pipeline

What is the overall processing pipeline? After identifying this, we can draw a nice picture.

3.4 Outlook

Outlook to evaluation on other tasks, that benefit from our task. This should be reorganized into the other .tex file in the end I guess. What I mean by this, is that we could evaluate our computed clusters on other tasks. Sort of how good recommended key words based on the classification are compared to ground truth key words. However, depending on how we organize our approach, this seems sort of like a full circle.

4 Section

Use sections to organize your contents. Read the project proposal guidelines available on Moodle to get more information on the contents your proposal should cover. Do not forget to cite online sources [?], books [?] or articles you are referencing! It may also be useful to integrate charts or figures in your proposal as seen in Figure 1.

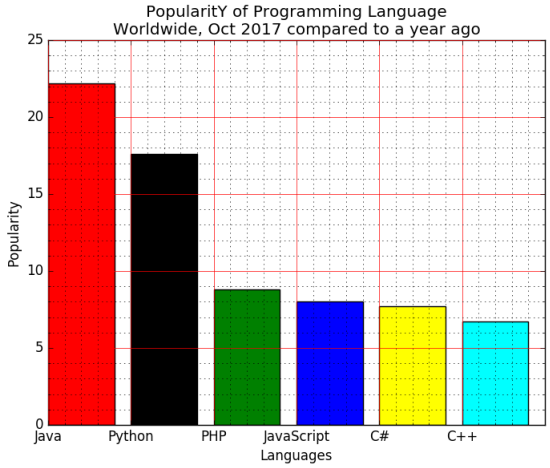

Figure 1: An example chart showing the change of popularity of various programming languages[1].

In the Latex source provided together with this PDF, you also find hints on how to work on one Latex project collaboratively.

References

[1] Acl anthology network (aan). `http://aan.how/`.

[2] Cicling: International conference on computational linguistics and intelligent text processing. `https://www.cicling.org/`.

[3] Web of science. `www.webofknowledge.com`.

[4] Mikhail Alexandrov, Alexander Gelbukh, and Paolo Rosso. An approach to clustering abstracts. In Andrés Montoyo, Rafael Muńoz, and Elisabeth

[1] `https://www.w3resource.com/w3r_images/matplotlib-barchart-exercise-4.png`

Métais, editors, *Natural Language Processing and Information Systems*, pages 275–285, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[5] Bader Aljaber, Nicola Stokes, James Bailey, and Jian Pei. Document clustering of scientific texts using citation contexts. *Inf. Retr.*, 13:101–131, 04 2010.

[6] Mark P. Carpenter and Francis Narin. Clustering of scientific journals. *Journal of the American Society for Information Science*, 24(6):425–436, 1973.

[7] Cody Dunne, Ben Shneiderman, Robert Gove, Judith Klavans, and Bonnie Dorr. Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization. *Journal of the American Society for Information Science and Technology*, 63(12):2351–2369, 2012.

[8] c-k Yau, Alan Porter, Nils Newman, and Arho Suominen. Clustering scientific documents with topic modeling. *Scientometrics*, 100:767–786, 09 2014.