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, Matriculation Number, Course of Study
email address
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

GitHub Repository: `https://github.com/DonatJR/ita_ws20`

1 Motivation

When starting a new project or trying to improve existing bodies of work it is often vitally important to do extensive research. This both helps to familiarize oneself with the topic that is to be worked on and provides the necessary new information that is needed to solve the problem at hand.

In comparison to more general search engines, where finding information for most topics is rather easy and comfortable, doing research on scientific papers can be more cumbersome. There are multiple websites with different papers and search interfaces available, and to get a comprehensive overview one has to go through all of them and adapt their 'search-procedure' to the available tools.

We want to propose a solution to this by taking scientific texts and clustering them into relevant subgroups, which can then be more easily presented to and explored by people looking for specific topics and terms. While we intend to first focus on a rather narrow subset of papers from a topic like Deep Learning or something similar, more or less specific subgroups for clustering therein can be explored to find a good balance. The pipeline should later be usable on a broader range of fields.

Furthermore, we think a solution to the stated problem could later be used on a grander scale by building good group visualization tools and providing existing websites with this technology. In addition, a meta search site incorporating data from these other sources with a common, easily digestible search and presentation interface could be developed.

2 Research Topic Summary

There exist various approaches to simplify getting an overview of a research field without extensive research and reading countless papers. 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 similarity 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.

With these information in mind there are multiple approaches we could take in our own project. The final goal is of course to achieve a better clustering than previous attempts. In regards to [4], [5] and [8] there is the possibility to not only use abstracts or citation contexts for the clusters but

combine both information sources into the same process. There is also an opportunity in taking different clustering techniques (e.g. k-means, hierarchical clustering or one of the ones mentioned above) and comparing their respective performance with each other. In addition to all of the methods mentioned above we could also use a neural network architecture to either automate the process of generating word vectors from the texts or to automate the clustering from word vectors that are obtained in a more traditional fashion, but we are not yet sure if this will be possible considering the time constraints.

2.1 Pipeline

Our pipeline will consist of the stages that can be seen in figure 1. First of all data from two different sources will be downloaded, temporarily stored and cleaned up to disregard records with missing data. All cleaned up records will be stored permanently so we don't have to download and clean up the data every time we change some downstream settings. From this data storage the text we are going to use in the end will be extracted and some clustering algorithms will be used to build the final result.

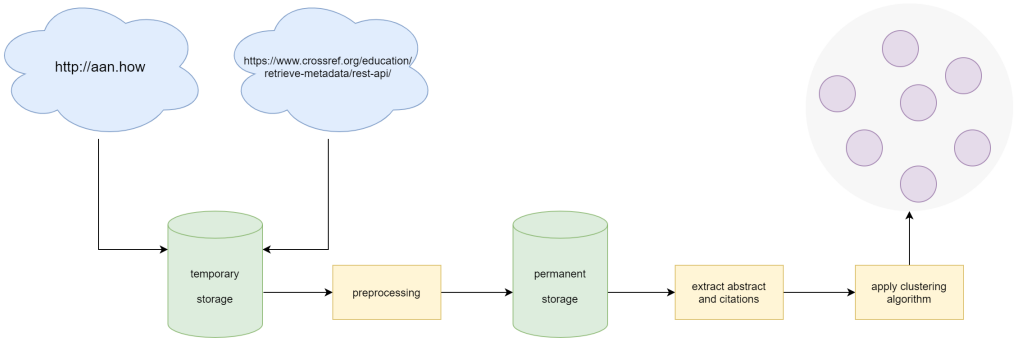

Figure 1: Coarse overview over the intended pipeline.

3 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 2.

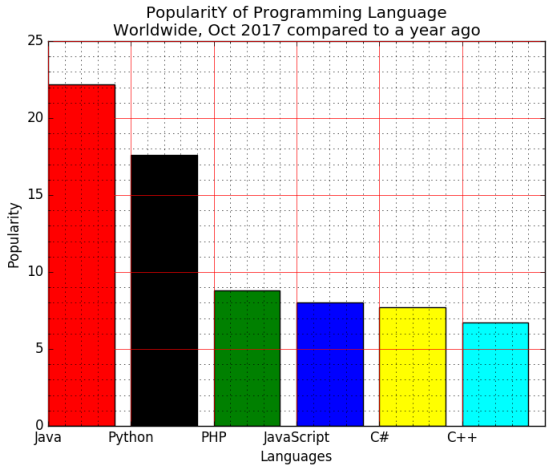

Figure 2: An example chart showing the change of popularity of various programming languages[1].

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 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.

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

[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.