EROL ÖZKAN – N18245609
erolozkan@outlook.com

CMP 681 INFORMATION RETRIVAL COURSE PROJECT

SUBJECT: Haberim Var – A Next Generation News Search Engine – *Be the First to Know Any News!*

Information Retrieval (IR) applied to news has been a popular research area for decades [1]. But, it is still far from a 'solved' problem. According to an international workshop held in 2016 called NewsIR'16 [1], information retrieval algorithms in news domain are still improving. The same workshop points out that there are serious gaps in the state of the art in this field [1].

The purpose of my course project is to provide users a system to search for the latest news in the Turkish newspapers. With this project, users will have an opportunity to find the latest news by searching for the keywords they are looking for.

Related Work

Catena et al. [2] introduce a variant of the well-known BM25 model, called BM25 Passage (BM25P). BM25P is specifically build to improve the effectiveness of a news retrieval system [2]. Their method distinguishes the importance of different news passages by assigning different weights to different passages in news articles. They show that, by differently weighting news passages, their method improves NDCG and MRR scores with respect to BM25.

Dataset

In order to evaluate their system, Catena et al [2] use three different corpora of news articles. These are;

- the AQUAINT Corpus by Linguistic Data Consortium (Aquaint) [3],
- the Signal Media One-Million News Articles Dataset (Signal) [4],
- the Reuters Corpus, Volume 1, version 2 (RCV1) [5].

Aquaint [3] consists of newswire text data in English, drawn from three sources: the Xinhua News Service, the New York Times News Service, and the Associated Press Worldstream News Service. Signal [4] is released to facilitate conducting research on news articles. It is intended to serve the community for research on news retrieval in general. Finally, RCV1 [4] is Reuters's news dataset and it contains news stories for use in research and development of in natural language processing, information retrieval, and machine learning systems.

Please note that, although Aquaint dataset provides 50 queries and their associated relevance judgments, it is not available for free. It costs about $150.00 to download this corpora [3]. And, Signal and RCV1 datasets do not provide any evaluation data [4-5]. However, there is a different methodology to evaluate on these datasets [2].

In my project, I will be building my own Turkish news dataset without any evaluation data. In order to evaluate my system, I will be using the evaluation method described next.

Evaluation

Since Aquaint and RCV1 datasets do not provide any evaluation data, Catena et al. adopted a different methodology. They used the news titles as pseudo-queries. According to this methodology, there is only one relevant news article for each query which is the article to which the title belongs to. All other articles are considered to be non-relevant [2]. For Signal and RCV1 collections, they randomly selected 40, 000 documents to generate the same number of pseudo-queries [2].

In my project, I will be using the same evaluation methodology in order to evaluate my system on Turkish news.

Similar Applications

There are some applications such as Bundle [6], MSN [7], Google News [8] and Yandex News [9] provide their users with some kind of news resources. However;

- Bundle doesn't have any search capabilities.
- MSN provides only some news feed. But, it doesn't have search capabilities. It redirects to Yandex search engine which doesn't have any specific news search capabilities.
- Yandex News is not available in Turkish. It is only available in Russian.

- Finally, although, Google News provides specific news search capabilities, it is not national and it is not totally accurate for Turkish.

Also, there are some Turkish companies like Medya Takip Merkezi [10] that provide their users with media monitoring, such as; recording, sorting, reporting, providing results. They archive newspapers, magazines, internet media TV and radio channels. However, the search capabilities of these systems creates questions in mind. Furthermore, they are not open systems.

Track Information

"Haberim Var" will be an application-oriented project. So, it is a startup track. It will result in a demo system. Users will be able to find any news in Turkish. And, since there are no competitors known by the society except Google News and Google News is not national, it may fill a big gap in Turkey. It will be developed using Python programming language and usefulness of this system will be demonstrated by the known search engine evaluation methods.

Design Components

"Haberim Var" will have similar components to a search engine. Search engines usually have the following components [11];

- **Text Acquisition:** Identifies and acquires documents.
- **Text Transformation:** Parsing component responsible for processing the sequence of text.
- **Index Creation:** Gathers and records statistical information about words, features, and documents.
- **User Interaction:** Provides an interface and a parser for queries.
- **Ranking:** Calculates scores for documents using a ranking algorithm.
- **Evaluation:** Logs user's queries and interactions.

Project Plan:

To develop this project; Firstly, I will do a research in advanced IR technologies and I will examine the most interesting novel applications. Then, in order to achieve my goal, I will implement some of the most useful technologies and algorithms for my project. At the end of the semester, there will be a production ready application. So users will be using "Haberim Var" to search for the latest Turkish news. Also, I will gain experience in many IR technologies and algorithms.

- Domain Research – *1 week*
- Writing Related Work Report – *1 week*
- Text Acquisition – *1 week*
- Text Transformation – *1 week – End of April.*
- Index Creation – *2 weeks*
- User Interaction – *2 weeks*
- Ranking – *1 week – End of May.*
- Evaluation – *2 weeks.*
- Project Report – *2 weeks – End of this Semester.*

Referanslar:

[1] NewsIR16. 2016. Retrieved from: https://research.signal-ai.com/newsir16/.
[2] Matteo C., Ophir F., Cristina I.M., Franco M.N., Raffaele P., and Nicola T. 2019. Enhanced News Retrieval: Passages Lead the Way! In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'19). Association for Computing Machinery, New York, NY, USA, 1269–1272. DOI:https://doi.org/10.1145/3331184.3331373.
[3] Aquaint Dataset. 2020. https://catalog.ldc.upenn.edu/LDC2002T31
[4] Signal Dataset. 2020. https://research.signal-ai.com/newsir16/signal-dataset.html
[5] RCV1 Dataset. 2020. https://trec.nist.gov/data/reuters/reuters.html
[6] Bundle Haber. 2020. Retrieved from: https://www.bundlehaber.com/.
[7] MSN. 2020. Retrieved from: https://www.msn.com/tr-tr/.
[8] Google News. 2020. Retrieved from: https://news.google.com/.
[9] Yandex News. 2020. Retrieved from: https://yandex.ru/news/.
[10] Medya Takip Merkezi. 2020. Retrieved from: https://www.medyatakip.com.tr/medya-arama-motoru/.
[11] Croft B., Metzler D., Strohman T., 2010. Search Engines: Information Retrieval in Practice, Addison-Wesley.