“Haberim Var”
Turkish News Retrival
(Startup track)

Erol ÖZKAN
erolozkan@outlook.com

Content

- Introduction
- Track Information
- Competition
- Related Work
- Design
- Completed Tasks
- Remaining Tasks
- Evaluation Method and Results
- Demo

Introduction

- Information Retrieval (IR) applied to news has been a popular research area for decades.
- But, it is still far from a ‘solved’ problem.
- According to an international workshop held in 2016 called NewsIR [1], information retrieval algorithms are still improving in news domain.
- In this project, we aim is to build Turkish news retrieval system.
- With this project, users have an opportunity to find the latest Turkish news by searching for the keywords they are looking for.

Track Information

- “Haberim Var” is a startup track. It is a pure application-oriented project.
- There is a demo system that users are able search for the news they are looking for.
- Specifically;
  - Crawls the web for the latest news articles.
  - It does statistical analysis about the data.
  - Indexes the articles into an inverted index.
  - For given single query, it return top ranking documents with their scores.
  - Finally, it provides a simple and user-friendly search interface.

Competition

- There are some competitors such as Bundle [2], MSN [3], Google News [4] and Yandex News [5].
- These systems provide their users with some kind of news information.
- However;
  - Bundle doesn’t have any search capabilities.
  - MSN provides their users with a news feed. But, it doesn’t have any search capabilities. It redirects to Yandex search engine.
  - Yandex News has search capability. But, It is only available in Russian.
  - Finally, although Google News has news search functionality, it is not national and it is not totally accurate for Turkish.

Related Work

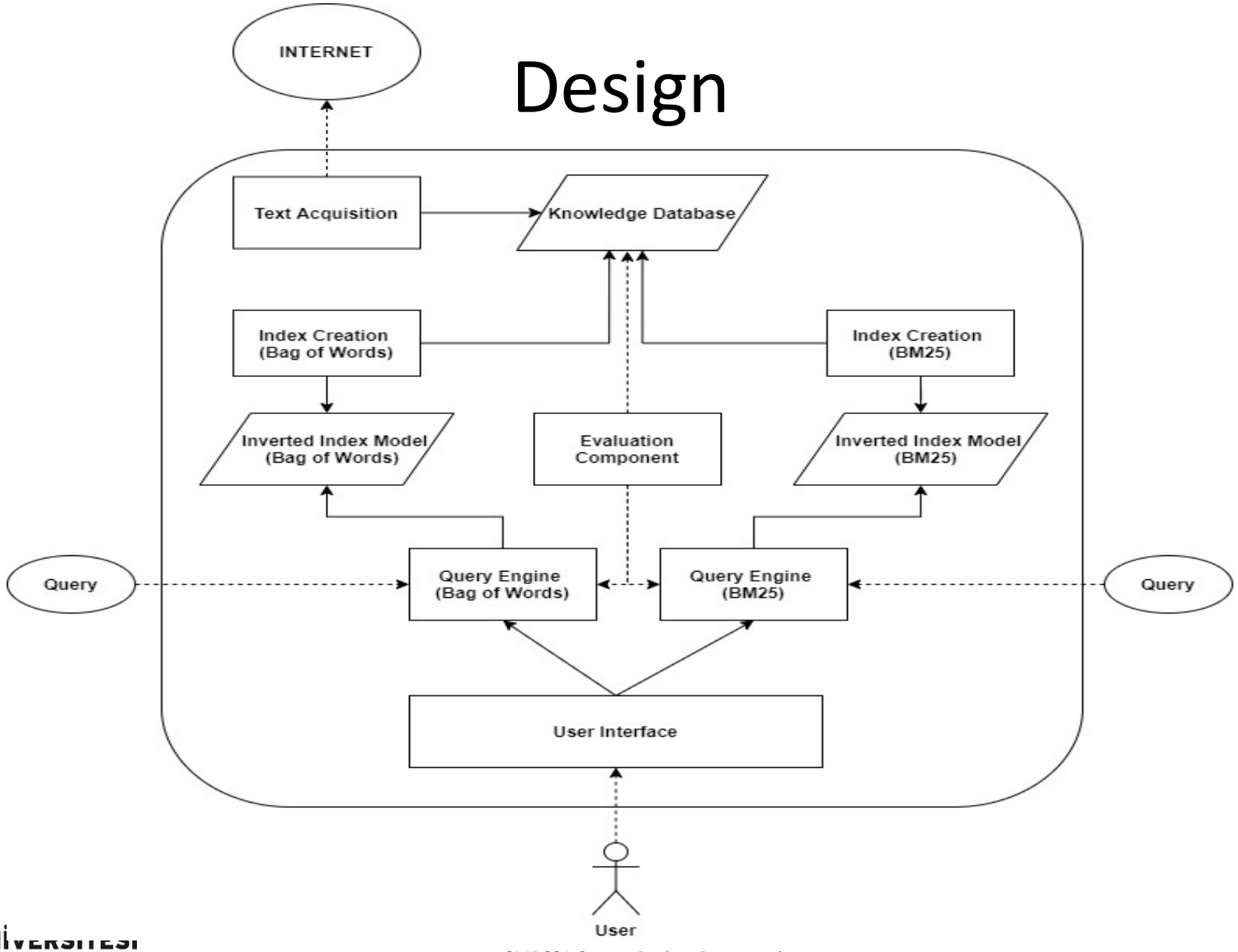
- Catena et al. [2] build a model called BM25P to improve the effectiveness of a news retrieval system.
  - They use different weights for the different passages in the news articles.
  - They show that, their method achieves better results than BM25.
- In order to evaluate their system, They use three different datasets (Aquaint, Signal and RCV1).
  - The Signal and RCV1 datasets do not provide any evaluation data.
  - As a result, they use a different methodology for these datasets. They assume that there is only one relevant article for each query. It is the article which the title belongs to. All other articles are considered non-relevant.
  - They use MRR metric to evaluate the performance on the Signal and RCV1 datasets.

Dataset

- We build our own Turkish news dataset.
- It doesn't have any evaluation data.
- We evaluate our system using MRR metric using the same method.

	Number of Documents	Average Query Length	Average Text Length
Haberturk	165	7.93	382.18
Milliyet	453	7.06	431.66
Hurriyet	523	6.99	353.42
Ntv	57	7.78	385.84
Sputniknews (tr)	593	9.13	269.32
Cnnturk	280	6.55	298.53
TOTAL	**2071**	**7.57**	**353.49**

Design


INTERNET
Text Acquisition
Knowledge Database
Index Creation (Bag of Words)
Index Creation (BM25)
Inverted Index Model (Bag of Words)
Evaluation Component
Inverted Index Model (BM25)
Query
Query Engine (Bag of Words)
Query Engine (BM25)
Query
User Interface
User

Completed Tasks

- Text acquisition component is implemented.
- Turkish news dataset is built.
- Two different retrieval methods are implemented.
  - Basic bag of words model.
  - BM25 weighting model.
- Evaluation scripts are implemented.
- User interfaces are designed and implemented (Html, Javascript, CSS, and Bootstrap).

Remaining Tasks

- Project final report will be written.
- Project final report will be converted into LaTeX format.

Evaluation (Effectiveness)

- We use MRR metric to evaluate the effectiveness of our system.
  - BM25 model is almost two times better than classical bag of words approach.

	Bag of Words	BM 25
Haberturk	0.34	0.61
Milliyet	0.46	0.66
Hurriyet	0.42	0.65
Ntv	0.54	0.74
Sputniknews (tr)	0.52	0.87
Cnnturk	0.41	0.63
TOTAL	**0.45**	**0.71**

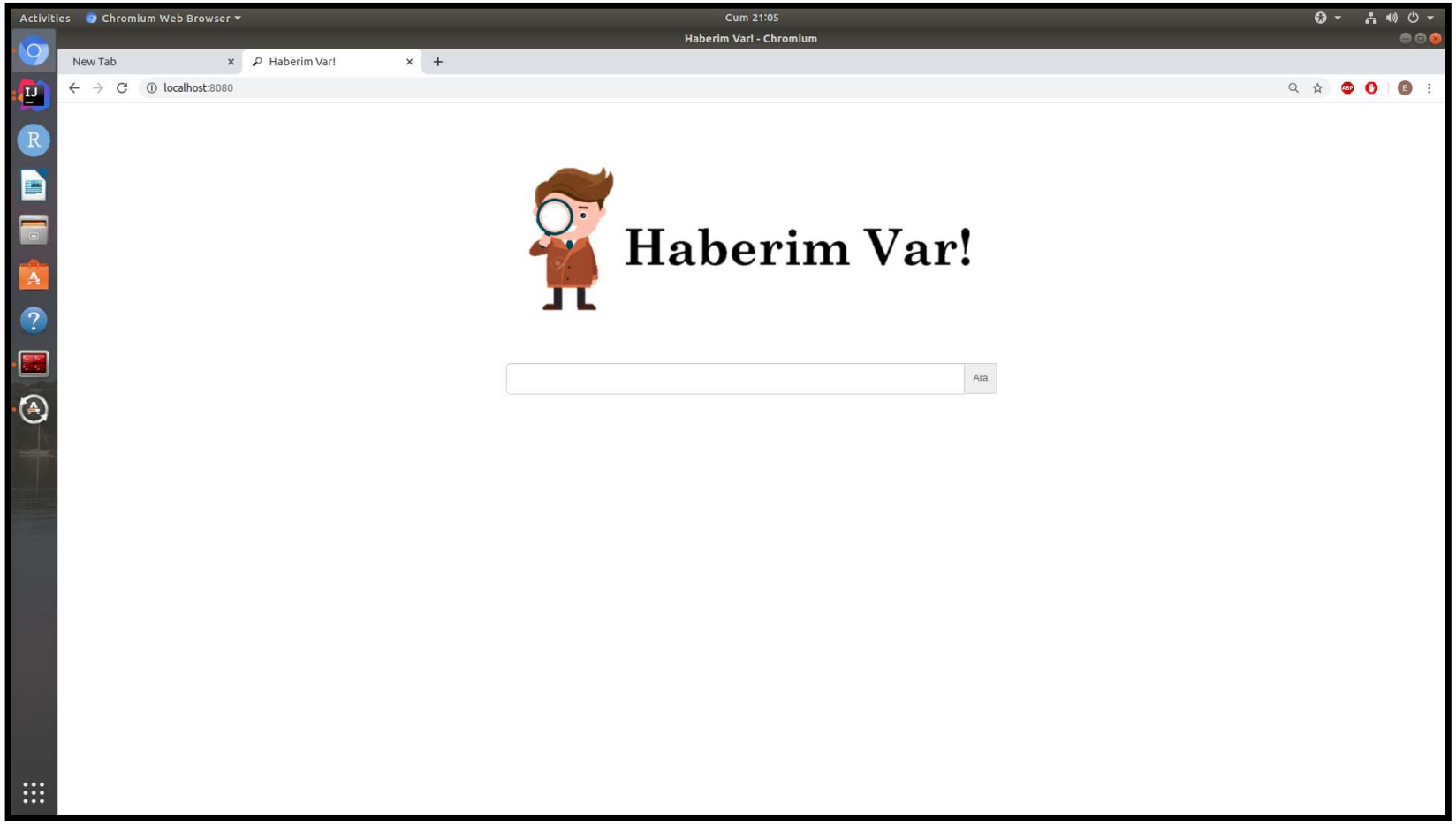
Evaluation (Efficiency)

- We measure the index creation and retrieval times for two methods.
  - Bag of words model is not very efficient in terms of both index creation and retrieval.
  - BM25 model is very efficient in terms of both index creation and retrieval.

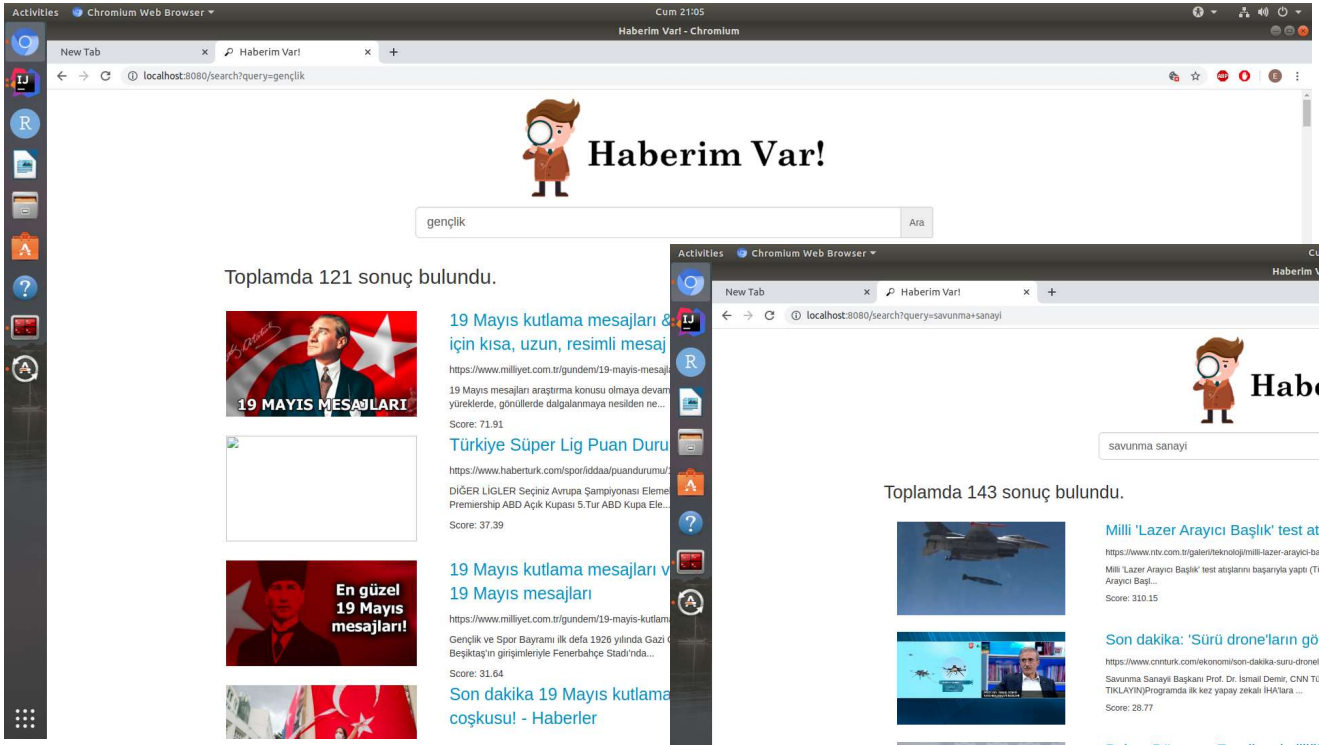
	Bag of Words	BM 25
Index Creation	10 minute 35 seconds	10 seconds
Query	17 minutes 39 seconds	01 minutes 33 seconds

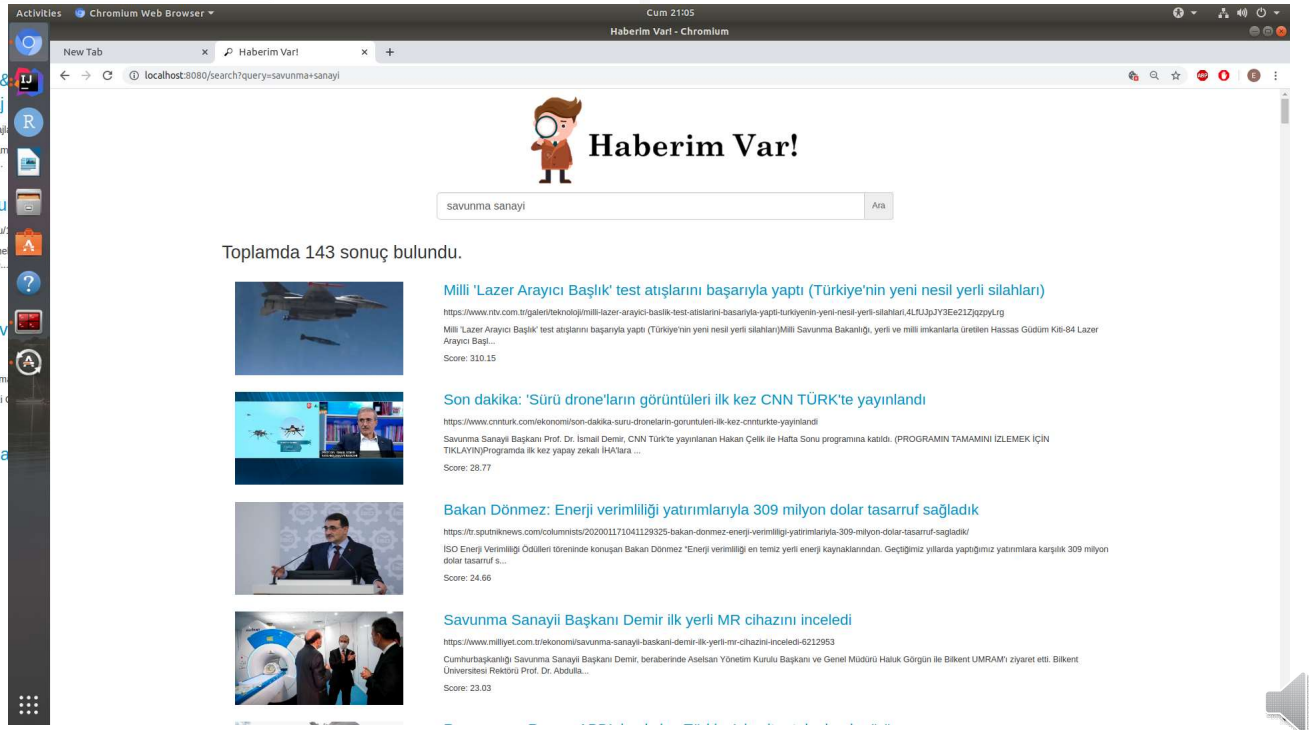
Intel Core i7-4700HQ is used.

“Haberim Var”


Activities
Chromium Web Browser
Cum 21:05
Haberim Var! - Chromium
New Tab
Haberim Var!
localhost:8080
Haberim Var!
Ara

“Haberim Var”


Activities
Chromium Web Browser
Cum 21:05
Haberim Var! - Chromium
New Tab
Haberim Var!
localhost:8080/search?query=gençlik
Haberim Var!
gençlik
Ara
Toplamda 121 sonuç bulundu.
19 MAYIS MESAJLARI
19 Mayıs kutlama mesajları &
için kısa, uzun, resimli mesaj
https://www.milliyet.com.tr/gundem/19-mayis-mesaj
19 Mayıs mesajları araştırma konusu olmaya devam
yüreklerde, gönüllerde dalgalanmaya nesilden ne...
Score: 71.91
Türkiye Süper Lig Puan Duru
https://www.haberturk.com/spor/iddaa/puandurumu/
DİĞER LİGLER Seçiniz Avrupa Şampiyonası Eleme
Premiership ABD Açık Kupası 5.Tur ABD Kupa Ele...
Score: 37.39
En güzel
19 Mayıs
mesajları!
19 Mayıs kutlama mesajları v
19 Mayıs mesajları
https://www.milliyet.com.tr/gundem/19-mayis-kutlam
Gençlik ve Spor Bayramı ilk defa 1926 yılında Gazi
Beşiktaş'ın girişimleriyle Fenerbahçe Stadı'nda...
Score: 31.64
Son dakika 19 Mayıs kutlama
coşkusu! - Haberler


Activities
Chromium Web Browser
Cum 21:05
Haberim Var! - Chromium
New Tab
Haberim Var!
localhost:8080/search?query=savunma+sanayi
Haberim Var!
savunma sanayi
Ara
Toplamda 143 sonuç bulundu.
Milli 'Lazer Arayıcı Başlık' test atışlarını başarıyla yaptı (Türkiye'nin yeni nesil yerli silahları)
https://www.ntv.com.tr/galeri/teknoloji/milli-lazer-arayici-baslik-test-atislarini-basariyla-yapti-turkiyenin-yeni-nesil-yerli-silahlari,4LfUJpJY3Ee212ZjqzpyLrg
Milli 'Lazer Arayıcı Başlık' test atışlarını başarıyla yaptı (Türkiye'nin yeni nesil yerli silahları)Milli Savunma Bakanlığı, yerli ve milli imkanlarla üretilen Hassas Güdüm Kiti-84 Lazer Arayıcı Başl...
Score: 310.15
Son dakika: 'Sürü drone'ların görüntüleri ilk kez CNN TÜRK'te yayınlandı
https://www.cnnturk.com/ekonomi/son-dakika-suru-dronelarin-goruntuleri-ilk-kez-cnnturkte-yayinlandi
Savunma Sanayii Başkanı Prof. Dr. İsmail Demir, CNN Türk'te yayınlanan Hakan Çelik ile Hafta Sonu programına katıldı. (PROGRAMIN TAMAMINI İZLEMEK İÇİN TIKLAYIN)Programda ilk kez yapay zekalı İHA'lara ...
Score: 28.77
Bakan Dönmez: Enerji verimliliği yatırımlarıyla 309 milyon dolar tasarruf sağladık
https://tr.sputniknews.com/columnists/202001171041129325-bakan-donmez-enerji-verimliligi-yatirimlariyla-309-milyon-dolar-tasarruf-sagladik/
İSO Enerji Verimliliği Ödülleri töreninde konuşan Bakan Dönmez "Enerji verimliliği en temiz yerli enerji kaynaklarından. Geçtiğimiz yıllarda yaptığımız yatırımlara karşılık 309 milyon dolar tasarruf s...
Score: 24.66
Savunma Sanayii Başkanı Demir ilk yerli MR cihazını inceledi
https://www.milliyet.com.tr/ekonomi/savunma-sanayii-baskani-demir-ilk-yerli-mr-cihazini-inceledi-6212953
Cumhurbaşkanlığı Savunma Sanayii Başkanı Demir, beraberinde Aselsan Yönetim Kurulu Başkanı ve Genel Müdürü Haluk Görgün ile Bilkent UMRAM'ı ziyaret etti. Bilkent Üniversitesi Rektörü Prof. Dr. Abdulla...
Score: 23.03

“Haberim Var”

Conclusion

- According to an international workshop called NewsIR [1], information retrieval algorithms in news domain are still improving.
- In this project, we built a Turkish news retrieval system.
  - We implemented text acquisition component.
  - We built a Turkish news dataset.
  - We implemented bag of words and BM25 retrieval models.
  - We implemented evaluation scripts using MRR metric.
  - Finally, we designed and implemented user interfaces with Html, Javascript, CSS, and Bootstrap.
- With this project, users can find the latest Turkish news by searching for the keywords they are looking for.

References (1)

[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/.

[11] Croft B., Metzler D., Strohman T., 2010. Search Engines: Information Retrieval in Practice, Addison-Wesley.

References (2)

[12] Toine Bogers and Antal van den Bosch. 2007. Comparing and Evaluating Information Retrieval Algorithms for News Recommendation. In Proc. RecSys. ACM, 141–144.

[13] Jose M. Chenlo and David E. Losada. 2014. An empirical study of sentence features for subjectivity and polarity classification. Inf. Sc. 280 (2014), 275 – 288.

[14] Dipanjan Das and André F.T. Martins. 2007. A survey on automatic text summarization. Lit. Survey for the Lang. and Stat. II course at CMU 4 (2007), 192–195.

[15] Chin-Yew Lin. 1999. Training a Selection Function for Extraction. In Proc. CIKM.

ACM, 55–62.

[16] Sean MacAvaney, Andrew Yates, Kai Hui, and Ophir Frieder. 2019. Content-Based Weak Supervision for Ad-Hoc Re-Ranking. In SIGIR 2019.

[17] Saket Mengle and Nazli Goharian. 2009. Passage detection using text classification. JASIST 60, 4 (2009), 814–825.

[18] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1996. Okapi at TREC-3. 109–126.

[19] Stephen E. Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. Found. Trends Inf. Retr. 3, 4 (April 2009), 333–389.