LOCALIZED AND PERSONALIZED SEARCH ENGINE FOR COVID-19

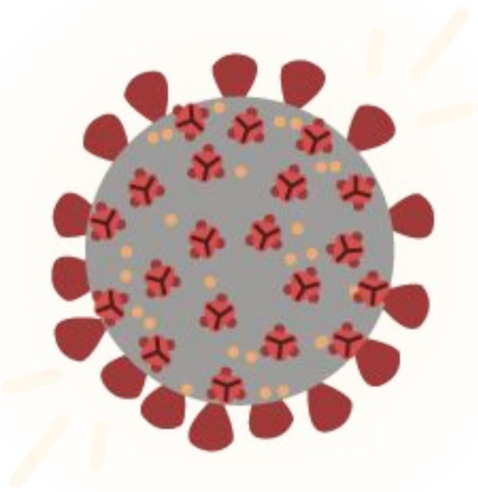
601.466/666 Information Retrieval & Web Agents

Milind, Darius, Satish, Katarina

Code: https://github.com/tpsatish95/covid19-search-engine

PROBLEM STATEMENT

- Throughout the COVID-19 pandemic, **people's needs have evolved** due to a myriad of closures and stay-at-home orders. Local services have had to adapt themselves to this everyday and this information **changes at a fast pace.**
- All of this **new and dynamic information** is difficult to sift through and not always straightforward to find.



OBJECTIVES

- Our goal is to crawl, aggregate, index and search/retrieve **information from local news sources in Baltimore** and report back relevant and personalized results to the user.
- To achieve this, we built a **search engine to retrieve relevant articles**. We expanded our search engine to simulate **user personalization** based on the user's profile, which can be mimicked through topics the user is biased towards, that are incorporated as a string of bias terms at runtime. This allows us to retrieve results personalized to the user needs.

APPROACH

Working with real-world datasets to build a search engine is **challenging since they do not contain relevance judgements**. We use the following approach to help us evaluate whether the articles returned by our search engine are indeed "relevant".

Common Information Retrieval Datasets → **Best Performing System** → **Search Engine** ← **Local News Sources**

Common Information Retrieval Datasets

We use CACM, CISI, Cranfield and Medline datasets for development

Best Performing System

After testing various permutations of the system on the common IR datasets, we choose to deploy the system with the which performs best on all 4 datasets

Search Engine

- Preprocessing
- Vectorization
- Scoring
- Query Expansion
- User Personalization

Local News Sources

We crawl news articles from WBALTV and CBS Baltimore and use this as the dataset for our final search engine deployment

DATASETS - EVALUATION

To find an appropriate system for real-world data, we considered **4 labelled datasets** (CACM, CISI, Medline, Cranfield) and conducted experiments on this data:

- **CACM**: abstracts and queries from Communications of ACM journal
- **CISI**: documents and queries from Centre for Inventions and Scientific Information
- **Medline**: collection of articles and queries from Medline journals
- **Cranfield**: commonly used IR dataset with aerodynamics journals articles, queries, and relevance judgements

DATASETS - DEPLOYMENT

- Then, we selected the best performing permutations from evaluation on development data to deploy on our COVID-19 news data.
- We crawled COVID-19 related articles from **CBS Baltimore** and **WBALTV** since they provide access to focused local information relevant to Baltimore.

METHODS

- Preprocessing
  - **Structured Text**: Stemming (Porter), Stopwords removal (using scikit-learn stopwords and punctuations list).
  - **Unstructured Text**:
    - Tokenization (using twokenize), Spell correction (Peter Norvig's spell checker).
    - Acronyms, Contractions and Emoticons: using scraped data from internetslang.com and urbandictionary.com.

METHODS

- <u>Vectorization</u>
  - **Sentence Embeddings using Word Embeddings:**
    - **Word Embeddings**: One-hot, Word2Vec, FastText, and GloVe.
    - **Weighting Techniques used to merge Word Embeddings:** Mean, TF-IDF, Smooth Inverse Frequency (SIF)[1], Unsupervised Smooth Inverse Frequency (uSIF)[2].
  - **Direct Sentence Embeddings:** Doc2Vec (from gensim)
- <u>Similarity Metric:</u> Cosine similarity, as it works best with all the vector embeddings.

[1] Arora et al. (2017): A Simple but Tough-to-Beat Baseline for Sentence Embeddings. https://openreview.net/pdf?id=SyK00v5xx
[2] Ethayarajh, K. (2019). Unsupervised Random Walk Sentence Embeddings: A Strong but Simple Baseline. 91–100. https://doi.org/10.18653/v1/w18-3012

METHODS

- User Personalization
  - To simulate personalization, we added the ability to include bias terms which characterizes a user's profile, at runtime. A user can enter in topics which **simulate a bias towards a user's behavior or preferences.**
  - To incorporate this user preference into the query, we perform an initial search using the bias terms as a query of its own. Then, we use a modified **Rocchio relevance feedback** mechanism to update the original query vector by moving it closer to the centroid of the documents relevant to the bias terms. (**Note:** D_r and D_{nr} represent documents relevant and non-relevant to the bias terms).

$$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j$$

METHODS

- Query Expansion
  - We allow for **query expansion based on GloVe** (glove-wiki-gigaword-100) which has ~400K vectors in the vocabulary and is pre-trained on Wikipedia-2014 data with 6B uncased tokens.
  - For each term in the query we get the top K words/vectors that are at least 70% similar to the query term (cosine similarity) and incorporate them back into the query for reformulating it.

RESULTS

- The models that we'll use for deployment since they performed best on the evaluation/development data are:
  - One-Hot encoded word vectors with TF-IDF Weighting
  - Word2Vec Word Embeddings (*word2vec-google-news-300*) with Unsupervised Smooth-Inverse Frequency (uSIF) Weighting.

No.	Embedding	Weighting Scheme	$P_{0.25}$	$P_{0.50}$	$P_{0.75}$	$P_{1.0}$	P_{mean1}	P_{mean2}	R_{norm}	P_{norm}
1.	**one-hot**	**TF-IDF**	**0.547**	**0.361**	**0.224**	**0.082**	**0.377**	**0.359**	**0.874**	**0.68**
2.	one-hot	Mean	0.458	0.282	0.172	0.068	0.304	0.296	0.854	0.622
3.	**word2vec-google-news-300**	**uSIF**	**0.416**	**0.261**	**0.15**	**0.058**	**0.276**	**0.269**	**0.871**	**0.612**
4.	word2vec-google-news-300	SIF	0.399	0.25	0.138	0.051	0.262	0.259	0.867	0.604
5.	word2vec-google-news-300	TF-IDF	0.39	0.232	0.124	0.043	0.249	0.245	0.84	0.58

***Note:** All of the above metrics were averaged over the system's performance on all 4 evaluation datasets. For all results and permutations other than the top-5 see this and this.

SEARCH ENGINE

```
└─ python deploy.py
####################################################
Model details (embedding, weighting_scheme): (one-hot, tf-idf)
Search engine initialized! Try the search engine:

Query: ventilators

1.  Ford to build 50,000 ventilators in 100 days
URL: https://www.wbaltv.com/article/ford-to-build-50-000-ventilators-in-100-days/31983486

2.  'I am willing to give up my ventilator': Woman makes changes to living will amid coro
navirus outbreak
URL: https://www.wbaltv.com/article/pittsburgh-woman-made-changes-to-living-will-in-event
-medical-professionals-must-decide-who-gets-life-saving-equipment/31989167

3.  Coronavirus Latest: Johns Hopkins Working On Device So Patients Can Share Ventilators
URL: https://baltimore.cbslocal.com/2020/04/02/coronavirus-latest-johns-hopkins-working-o
n-device-so-patients-can-share-ventilators/

4.  Some states receive masks with dry rot, broken ventilators
URL: https://www.wbaltv.com/article/some-states-receive-masks-with-dry-rot-broken-ventila
tors/32038844

5.  Watchdog report finds severe shortages and significant challenges to hospitals' coron
avirus responses
URL: https://www.wbaltv.com/article/watchdog-report-finds-severe-shortages-and-significan
t-challenges-to-hospitals-coronavirus-responses/32050699
```

DEMO

Code: https://github.com/tpsatish95/covid19-search-engine

REFERENCES

- Arora et al. (2017): A Simple but Tough-to-Beat Baseline for Sentence Embeddings. https://openreview.net/pdf?id=SyK00v5xx
- Ethayarajh, K. (2019). Unsupervised Random Walk Sentence Embeddings: A Strong but Simple Baseline. 91–100. https://doi.org/10.18653/v1/w18-3012
- Amir, Silvio, Glen Coppersmith, Paula Carvalho, Mário J. Silva, and Byron C. Wallace. "Quantifying Mental Health from Social Media with Neural User Embeddings." ArXiv:1705.00335 [Cs], April 30, 2017.
- Norgaard, Ole, and Jeffrey V. Lazarus. "Searching PubMed during a Pandemic." PLoS ONE 5, no. 4 (April 7, 2010).
- Rastegari, Hamid, and Siti Mariyam Shamsuddin. "Web Search Personalization Based on Browsing History by Artificial Immune System," n.d., 20.
- Sugiyama, Kazunari, Kenji Hatano, and Masatoshi Yoshikawa. "Adaptive Web Search Based on User Profile Constructed without Any Effort from Users." In Proceedings of the 13th Conference on World Wide Web - WWW '04, 675. New York, NY, USA: ACM Press, 2004.