Localized & Personalized Search Engine for COVID-19

601.466/666 Information Retrieval and Web Agents

Satish Palaniappan

Katarina Mayer

Darius Irani

Milind Agarwal

BACKGROUND Throughout the COVID-19 pandemic, the people and local services have had to adapt quickly everyday. This poses the challenge of being able to find the most up-to-date information in this dynamic environment.

OBJECTIVE Create a search engine that reports back relevant and personalized results to the user from Baltimore's local news sources.

METHODS

We use the following approach to help mitigate the lack of relevance judgements for building a search engine on real-world data:

Common IR Datasets

CACM + CISI + Cranfield + Medline for evaluation

↓

High Performing System

Permutations that perform well on all 4 datasets →

Local News Sources

Crawl Baltimore Sun, WBALTV, CBS Baltimore for deployment

↓

Search Engine

- Preprocessing
- Vectorization
- Query Expansion
- User Personalization

DISCUSSION

The models that performed the best on the evaluation data were the One-Hot Encoding and Word2Vec word embeddings with TF-IDF and uSIF weightings, respectively. These models were used for deployment on the crawled news data.

Get **personalized** and **latest local news** with our **COVID-19** Search Engine!

Example Query: 5 most relevant articles retrieved by our search engine on query 'masks'

```
Query: masks

1.  Fact Check: The arguments for and against widespread face mask use during the corona
virus outbreak
URL: https://www.wbaltv.com/article/fact-check-the-arguments-for-and-against-widespread-
face-mask-use-during-the-coronavirus-outbreak/32065758

2.  Coronavirus Resources: How To Make Your Own Face Mask
URL: https://baltimore.cbslocal.com/2020/04/05/coronavirus-resources-how-to-make-your-ow
n-face-mask/

3.  'Save us so we can save you': Maryland doctors, nurses battling coronavirus increasi
ngly desperate for protective supplies
URL: https://www.baltimoresun.com/coronavirus/bs-md-doctors-nurses-needing-supplies-2020
0327-iwpf3fd3k5ddzaapumujnmnz5e-story.html

4.  'Why not make masks?': Westminster cleaners giving out face coverings for essential
workers
URL: https://www.baltimoresun.com/coronavirus/cc-cleaners-masks-20200418-bnwbeavff5e6ph2
7d44vu6ao2e-story.html

5.  Masks sent to Maryland from federal stockpile in coronavirus crisis 'technically pas
t' suggested shelf life
URL: https://www.baltimoresun.com/coronavirus/bs-md-pol-facemasks-shelf-life-coronavirus
-20200327-2fyapp3ssnej5h3pz6oxurjiji-story.html
```

Evaluation results: used to help select best models for deployment

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

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

EXTRA DETAILS

PREPROCESSING

- **Structured:** Porter stemmer, stopword removal
- **Unstructured:** *twokenize* tokenization, spell correction (Peter Norvig's spell checker); Acronyms, Contractions and Emoticons (using scraped data from internetslang & urbandictionary)

VECTORIZATION

- **Word to Sentence Embeddings:**
  - **Word Embeddings:** One-hot, Word2Vec, FastText, and GloVe.
  - **Weighting:** Mean, TF-IDF, SIF[1], uSIF[2]
- **Sentence Embeddings:** Doc2Vec (from gensim)

Cosine similarity measure works best with all the vector embeddings.

USER PERSONALIZATION

To simulate personalization, we added the ability to keep track of the user's search history which characterizes a user's profile/preference, during runtime.

To incorporate this into the current query, we average the query vectors from the user's search history and perform an initial search using profile query vector. 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 user profile's query vector.

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.

REFERENCES

1. Arora et al. (2017): A Simple but Tough-to-Beat Baseline for Sentence Embeddings.
2. Ethayarajh, K. (2019). Unsupervised Random Walk Sentence Embeddings: A Strong but Simple Baseline. 91–100.
3. Arora et al. (2017): A Simple but Tough-to-Beat Baseline for Sentence Embeddings. https://openreview.net/pdf?id=SyK00v5xx
4. Ethayarajh, K. (2019). Unsupervised Random Walk Sentence Embeddings: A Strong but Simple Baseline. 91–100. https://doi.org/10.18653/v1/w18-3012
5. 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.
6. Norgaard, Ole, and Jeffrey V. Lazarus. "Searching PubMed during a Pandemic." PLoS ONE 5, no. 4 (April 7, 2010).
7. Rastegari, Hamid, and Siti Mariyam Shamsuddin. "Web Search Personalization Based on Browsing History by Artificial Immune System," n.d., 20.
8. 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.