EN 601.466/666 Final Project Report Group 1
Localized & Personalized Search Engine for COVID-19

Satish Palaniappan, Katarina Mayer, Darius Irani, Milind Agarwal
{spalani2, kmayer2, dirani2, magarw10}@jhu.edu

I. Introduction

Throughout the COVID-19 pandemic, local resources have shifted in scope and availability. Grocery stores are operating at reduced hours, businesses are offering increased delivery services, and shoppers' habits have evolved due to stay-at-home orders. Pandemics bring a host of additional challenges to search engines, including high volume of users querying specific and time-sensitive information, rapidly changing nature of information to be indexed, and keyword inconsistencies (Roberts et al.; Norgaard and Lazarus).

In the scientific community, open challenges and public releases of massive datasets like CORD-19 have enabled researchers to identify information helpful for combating the pandemic (Wang et al.; Roberts et al.). From these efforts, IR systems have rapidly been prototyped and deployed (Zhang et al.; MacAvaney et al.). Some efforts have also focused on developing information chatbots to answer questions related to the pandemic (JHU-COVID-QA).

Our work differs from this approach because we focus on retrieving information regarding the COVID-19 pandemic from local Baltimore news sources through an interactive, localized and personalized search engine. Our base system is a vector-space model that uses cosine similarity to determine relevance of a document to a query. We explore several extensions to this base model including context-free word embeddings, query expansion and user personalization methods.

The contributions of our work are two-fold:

- We develop and unleash web spiders to crawl local news sources for articles related to the COVID-19 pandemic. Specifically, we crawl articles from Baltimore Sun, WBALTV, and CBS Baltimore.
- We design and deploy a web-based personalized search engine for COVID-19 to retrieve documents that are relevant and personalized to user queries.

II. Methods

A. Data

Since evaluation is challenging for a system built directly on public news articles, we will take a two-fold approach: (1) we experiment and evaluate different systems on datasets with labeled relevance judgements and (2) we deploy the best performing system on data scraped from local Baltimore news sources. For development, we use the following labeled datasets:

- 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: Aerodynamics journal articles, queries, and relevance judgements.

We develop and unleash three web spiders to scrape news articles related to COVID-19 from the Baltimore Sun, WBALTV, and CBS Baltimore. The article texts and meta-data are extracted and the raw data is processed similarly to the development data.

B. Filtering

Prior to incorporating the scraped data into our knowledge base, the raw data is filtered based on keywords relevant to the COVID-19 pandemic and to local resources similar to the approach in Norgaard and Lazarus. A complete list of filtering keywords can be found in Appendix A.1.

C. Preprocessing

Preprocessing handles structured and unstructured data:

- Structured: Stemming (Porter Stemmer) & Stop Words Removal (scikit-learn's stopwords and punctuations list)
- Unstructured:We handle acronyms, contractions, and emoticons using scraped data from internetslang.com, urbandictionary.com and the emoticons wiki. We have added in Spell Correction (using Peter Norvig's spell checker) and better text tokenization (using Twokenize) capabilities as well.

D. Vectorization and Relevance

We employ a wealth of techniques from the course to build our system. We experiment with embedding schemes, weighting strategies and similarity metrics as outlined below:

Vectorization

- *Sentence Embeddings using Word Embeddings:*
  - *Word Embeddings:* One-hot, Word2Vec (Mikolov et al.), FastText (Bojanowski et al.), and GloVe (Pennington et al.).
  - *Weighting:* Mean, TF-IDF, Smooth Inverse Frequency (SIF) (Ethayarajh), Unsupervised SIF (uSIF) (Arora et al.).
- *Direct Sentence Embeddings:* Doc2Vec (Le and Mikolov)

Similarity Metric: Cosine similarity, as it works best with all the vector embeddings.

E. Query Expansion/Optimization

We allow for query expansion based on GloVe (glove-wiki-gigaword-100) which has 400K vectors in the vocabulary and is pretrained 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.

F. 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. (Note: D_r and D_{nr} represent documents relevant and non-relevant to the user profile.) (Degemmis et al.).

$$\overrightarrow{q_m} = \alpha\overrightarrow{q_0} + \beta\frac{1}{|D_r|}\sum_{\overrightarrow{d_j}\in D_r}\overrightarrow{d_j} - \gamma\frac{1}{|D_{nr}|}\sum_{\overrightarrow{d_j}\in D_{nr}}\overrightarrow{d_j}$$

III. Evaluation

The models that performed the best on the evaluation data were the One-Hot encoded and TF-IDF weighted document embeddings, and the Word2Vec (word2vec-google-news-300) word embeddings weighted with unsupervised smooth inverse frequency (uSIF). Table I summarizes key model permutations used to determine these model parameters. Table II summarizes the baseline model performance on the four datasets. We used precision and recall metrics to evaluate model performance.

Embedding	Weighting	$P_{0.25}$	P_{mean2}	R_{norm}	P_{norm}
one-hot	**TF-IDF**	**0.547**	**0.359**	**0.874**	**0.68**
one-hot	MEAN	0.458	0.296	0.854	0.622
word2vec-google-news-300	**uSIF**	**0.416**	**0.269**	**0.871**	**0.612**
word2vec-google-news-300	SIF	0.399	0.259	0.867	0.604
word2vec-google-news-300	TF-IDF	0.39	0.245	0.84	0.58

TABLE I: Select model permutations averaged over all datasets. For a complete list of model permutations, see Appendix A.3.

Dev Dataset	P@0.25	P_{mean2}	P_{norm}	R_{norm}
CACM	0.48	0.30	**0.87**	0.66
CISI	0.37	0.22	**0.81**	0.56
Medline	0.71	0.50	**0.91**	0.79
Cranfield	0.63	0.42	**0.91**	0.71

TABLE II: Baseline Results on Evaluation Datasets

IV. Results

Our search engine was successful in retrieving COVID-19 articles that were relevant and personalized to the user's query. Appendix A.2 showcases six sample queries that demonstrate the various features of the search engine.

We showcase results:

1) For basic queries
2) For queries with Acronyms/Abbreviations
3) For queries with Misspellings
4) That capture the semantics of the query using pre-trained word embeddings.
5) That are personalized towards the user's biases, preferences, search history.
6) That use query expansion to account for the query's topic in general.

V. Discussion

Future extensions to our search engine model include:

- Crawling more local news data from other sources and training a Word/Document embedding model, from scratch, specialized for COVID-19, as the current pre-trained embeddings we use are very generic and do not help COVID-19 specific search.
- Experimenting with relevance-based word embeddings to capture user profiles (Zamani and Croft).
- Using topic modeling to filter candidate documents during retrieval for more accurate and focused search results.
- Incorporating dimensionality reduction techniques like Singular Value Decomposition (SVD) and Local Linear Embeddings (LLE) into our experiments.

VI. Code

Code for this end-to-end system and accompanying results can be found in our GitHub repository at https://github.com/tpsatish95/covid19-search-engine. Instructions for setting up, deploying and using the search engine can be found in the README, along with must-try example queries demonstrating all the features of the search engine.

VII. Acknowledgements

We are grateful to Dr. Silvio Amir and the TAs, whose feedback during presentations and the entire semester was helpful in guiding our research, methodology and experiments.

References

[1] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. *ICLR*, page 16, 2017.

[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. ISSN 2307-387X.

[3] M Degemmis, P Lops, S Ferilli, N Di Mauro, T M A Basile, and G Semeraro. A Relevance Feedback Method for Discovering User Profiles from Text. page 12.

[4] Kawin Ethayarajh. Unsupervised random walk sentence embeddings: A strong but simple baseline. In

Proceedings of The Third Workshop on Representation Learning for NLP, pages 91–100, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-3012. URL https://www.aclweb.org/anthology/W18-3012.

[5] Quoc Le and Tomas Mikolov. Distributed Representations of Sentences and Documents. page 9.

[6] Sean MacAvaney, Arman Cohan, and Nazli Goharian. Sledge: A simple yet effective baseline for coronavirus scientific knowledge search, 2020.

[7] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[8] Ole Norgaard and Jeffrey V. Lazarus. Searching PubMed during a Pandemic. *PLoS ONE*, 5(4), April 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0010039.

[9] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162.

[10] Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen Voorhees, Lucy Lu Wang, and William R Hersh. TREC-COVID: Rationale and Structure of an Information Retrieval Shared Task for COVID-19. *Journal of the American Medical Informatics Association*, 2020. ISSN 1527-974X. doi: 10.1093/jamia/ocaa091. URL https://doi.org/10.1093/jamia/ocaa091.

[11] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. Cord-19: The covid-19 open research dataset, 2020.

[12] Hamed Zamani and W. Bruce Croft. Relevance-based Word Embedding. *arXiv:1705.03556 [cs]*, July 2017.

[13] Edwin Zhang, Nikhil Gupta, Rodrigo Nogueira, Kyunghyun Cho, and Jimmy Lin. Rapidly Deploying a Neural Search Engine for the COVID-19 Open Research Dataset: Preliminary Thoughts and Lessons Learned. page 10.

VIII. APPENDIX

A.1: Filtering Keywords

The presence of the following keywords were used to determine relevance of the documents to COVID-19:

- **CBS Baltimore and WBALTV:** covid, coravirus, resources, covid-19, store, essential, relief, covid 19, food, grocery, groceries, help, quarantine, restaurants, grocery store, social distance.
- **Baltimore Sun:** covid, coronavirus, quarantine, social distance, masks, ventilator

A.2: Example Queries

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

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
```

Fig. 1: (1) Search results for a generic query: "masks"

```
##################################################

Query: stayy at homew ordero

1.  Governors in these states have issued statewide stay-at-home orders to stop spread o
f coronavirus
URL: https://www.wbaltv.com/article/california-governor-orders-statewide-stay-at-home-or
der-1/31795709

2.  Coronavirus Latest: More Marylanders Staying At Home Amid Pandemic, Researchers Say
URL: https://baltimore.cbslocal.com/2020/04/17/coronavirus-latest-more-marylanders-stayi
ng-at-home-amid-pandemic-researchers-say/

3.  With coronavirus at more than 80 nursing homes in Maryland, Gov. Hogan increases pro
tective measures
URL: https://www.baltimoresun.com/coronavirus/bs-md-hogan-measures-coronavirus-nursing-h
omes-20200405-dmeuao5cvbg3lbpuq7p6u3njyy-story.html

4.  Leaders in Maryland, Virginia, District of Columbia, work to get in sync on coronavi
rus response
URL: https://www.baltimoresun.com/politics/bs-md-pol-regional-coordination-coronavirus-2
0200331-oq3ju7f6qzdnlborevu467hvoa-story.html

5.  Baltimore Violence Continues Despite Governor's Coronavirus 'Stay At Home' Order
URL: https://baltimore.cbslocal.com/2020/04/14/13-killed-baltimore-violence-maryland-sta
y-at-home-order/
```

Fig. 2: (2) Search results for a query with misspellings: "stayy at homew ordero".

```
##################################################

Query: JHU

1.  JHU to transition to remote instruction for all classroom-based academic programs
URL: https://www.wbaltv.com/article/johns-hopkins-university-to-transition-to-remote-ins
truction-for-all-classroom-based-academic-programs/31368351

2.  Johns Hopkins' coronavirus-tracking map now shows cases by city, county
URL: https://www.wbaltv.com/article/coronavirus-johns-hopkins-university-tracking-map-ci
ty-county/31901179

3.  Johns Hopkins moves all classes to online, cancels in-person May commencement and va
cates campus amid coronavirus
URL: https://www.baltimoresun.com/coronavirus/bs-md-johns-hopkins-remainder-semester-202
00318-ctolvzheovhjzmdkfu4gfyfxvy-story.html

4.  Bloomberg, Maryland give $4M to Hopkins for coronavirus treatment research
URL: https://www.wbaltv.com/article/coronavirus-treatment-research-michael-bloomberg-mar
yland-donation-johns-hopkins-university/31960071

5.  Amid coronavirus fears, Maryland colleges get conflicting signals on decisions to ch
ange plans for large events
URL: https://www.baltimoresun.com/news/bs-md-hogan-games-coronavirus-fears-20200307-zh2y
u35vkjfl7mmzvu6jgbwk2e-story.html
```

Fig. 3: (3) Search results for a query with acronyms/abbreviations: "JHU".

```
##################################################

Query: employment

1.  Q&A: What are Maryland workers' rights, responsibilities in the coronavirus pandemic
?
URL: https://www.baltimoresun.com/coronavirus/bs-md-workers-rights-coronavirus-20200326-
ggppz2jk65f33ifcs5qfwlbxnq-story.html

2.  Maryland unemployment call center expands hours amid coronavirus outbreak
URL: https://www.wbaltv.com/article/coronavirus-outbreak-maryland-unemployment-call-cent
er-hours-expanded/31787367

3.  Coronavirus Latest: How To File For Unemployment In Maryland
URL: https://baltimore.cbslocal.com/2020/04/01/coronavirus-latest-how-to-file-for-unempl
oyment-in-maryland/

4.  Maryland legislature passes bill to extend temporary unemployment benefits during co
ronavirus pandemic
URL: https://www.baltimoresun.com/coronavirus/bs-md-pol-ga-covid-legisation-20200318-qux
hhgu2erhzvabv2kpotib4pm-story.html

5.  Here's what you need to know if you get laid off in Maryland during coronavirus pand
emic
URL: https://www.baltimoresun.com/coronavirus/bs-md-unemployment-coronavirus-20200318-3u
aqvtlqfvfwlaweah4sjnwgdy-story.html

6.  The US Department of Labor just issued tips for workplaces preparing for the coronav
irus
URL: https://www.wbaltv.com/article/the-us-department-of-labor-just-issued-tips-for-work
places-preparing-for-the-coronavirus/31334978

7.  Maryland Senate passes emergency legislation to extend temporary unemployment benefi
ts during coronavirus pandemic
URL: https://www.baltimoresun.com/politics/bs-md-pol-ga-coronavirus-legislation-20200316
-27v2qj3rjvbbxmcj4tbodb65l4-story.html

8.  'It's terrifying': Service workers fear for their jobs, health as coronavirus spread
s
URL: https://www.baltimoresun.com/coronavirus/bs-md-coronavirus-service-workers-20200313
-anhocn3mmbdi7k4tci3t2edr7m-story.html
```

Fig. 4: (4) Search results for query: "employment", *without* pre-trained word embeddings that capture the query semantics.

```
└─ python deploy.py --embedding "word2vec-google-news-300" --weighting_scheme "usif"
##################################################
Model details (embedding, weighting_scheme): (word2vec-google-news-300, usif)
Search engine initialized! Try the search engine:

Query: employment

1.  Now Hiring: Here's A List Of Businesses Hiring In Baltimore
URL: https://baltimore.cbslocal.com/2020/04/16/now-hiring-heres-a-list-of-businesses-hir
ing-in-baltimore/

2.  Here's what you need to know if you get laid off in Maryland during coronavirus pand
emic
URL: https://www.baltimoresun.com/coronavirus/bs-md-unemployment-coronavirus-20200318-3u
aqvtlqfvfwlaweah4sjnwgdy-story.html

3.  Coronavirus Resources: Officials Say Maryland Unemployment Website Is Improving
URL: https://baltimore.cbslocal.com/2020/05/06/coronavirus-resources-officials-say-maryl
and-unemployment-website-is-improving/

4.  Unemployment filings skyrocket in Carroll County in sign of economic damage from cor
onavirus
URL: https://www.baltimoresun.com/coronavirus/cc-carroll-unemployment-claims-coronavirus
-20200403-zqq7trxya5danjrtzotfheigpa-story.html

5.  Q&A: What are Maryland workers' rights, responsibilities in the coronavirus pandemic
?
URL: https://www.baltimoresun.com/coronavirus/bs-md-workers-rights-coronavirus-20200326-
ggppz2jk65f33ifcs5qfwlbxnq-story.html

6.  Carroll County unemployment claims decline from previous weeks but remain high durin
g coronavirus pandemic
URL: https://www.baltimoresun.com/coronavirus/cc-carroll-unemployment-claims-coronavirus
-20200423-qk2pr2vy35b6bh7bqv7hutdqqq-story.html

7.  Jobless claims double, with 84,000 Marylanders filing for unemployment as coronaviru
s shutters businesses
URL: https://www.baltimoresun.com/coronavirus/bs-md-coronavirus-march21-unemployment-202
00402-2rtvowsq45bsvmjkglebclguza-story.html

8.  Coronavirus Update: If Your Maryland Unemployment Claim Became Inactive This Week, H
ere's What You Need To Know
URL: https://baltimore.cbslocal.com/2020/05/05/coronavirus-update-if-your-maryland-unemp
loyment-claim-became-inactive-this-week-heres-what-you-need-to-know/
```

Fig. 5: (4) Search results for query: "employment", *with* pre-trained word embeddings that capture the query semantics.

```
##################################################

Query: social distancing

1.  Coronavirus Study: Maryland Among Top 20 States Where Self-Isolating Is Most Difficu
lt
URL: https://baltimore.cbslocal.com/2020/04/28/coronavirus-study-maryland-among-top-20-s
tates-where-self-isolating-is-most-difficult/

2.  The extrovert's guide to social distancing
URL: https://www.wbaltv.com/article/the-extroverts-guide-to-social-distancing/31898384

3.  Barber brightens clients' day with virtual haircuts to practice social distancing
URL: https://www.wbaltv.com/article/new-orelans-barber-virtual-haircuts-social-distancin
g-coronavirus-covid19/31946882

4.  'I have real serious concerns': Trump's comments about coronavirus prompted this Joh
ns Hopkins official to speak out
URL: https://www.baltimoresun.com/coronavirus/bs-md-johnshopkins-inglesby-coronavirus-se
curity-response-20200324-h7mbqlvucjh2df6pfc6czyvmqe-story.html

5.  When will the coronavirus peak in Maryland? Here's what to know about the prediction
s
URL: https://www.baltimoresun.com/coronavirus/bs-hs-faq-coronavirus-predictions-20200407
-20200408-uulylejcyjddnkyssqbbwdrlgu-story.html
```

Fig. 6: (5) Search results for the query: "social distancing", *without* user personalization based on search history.

```
##################################################

Query: social distancing

1.  Coronavirus Study: Maryland Among Top 20 States Where Self-Isolating Is Most Difficu
lt
URL: https://baltimore.cbslocal.com/2020/04/28/coronavirus-study-maryland-among-top-20-s
tates-where-self-isolating-is-most-difficult/

2.  Costco, Home Depot limiting customers allowed in stores
URL: https://www.wbaltv.com/article/costco-home-depot-limiting-customers-allowed-in-stor
es/32005528

3.  The extrovert's guide to social distancing
URL: https://www.wbaltv.com/article/the-extroverts-guide-to-social-distancing/31898384

4.  Barber brightens clients' day with virtual haircuts to practice social distancing
URL: https://www.wbaltv.com/article/new-orelans-barber-virtual-haircuts-social-distancin
g-coronavirus-covid19/31946882

5.  'I have real serious concerns': Trump's comments about coronavirus prompted this Joh
ns Hopkins official to speak out
URL: https://www.baltimoresun.com/coronavirus/bs-md-johnshopkins-inglesby-coronavirus-se
curity-response-20200324-h7mbqlvucjh2df6pfc6czyvmqe-story.html
```

Fig. 7: (5) Search results for the query: "social distancing", *with* user personalization based on search history. **Note:** Here the user had already performed a query for "costco" hence, "social distancing" talks about limiting customers in the costco store in result no. 2.

```
##################################################

Query: recession

1.  Fed takes emergency action, including slashing rate by full percentage point
URL: https://www.wbaltv.com/article/fed-takes-emergency-action-including-slashing-rate-b
y-full-percentage-point/31645643

2.  What a 20% unemployment rate would mean for America
URL: https://www.wbaltv.com/article/what-a-20-unemployment-rate-would-mean-for-america/3
1746821

3.  Baltimore-area governments are weighing big cuts as they brace for a collapse in rev
enues due to coronavirus
URL: https://www.baltimoresun.com/coronavirus/bs-md-coronavirus-economic-crisis-counties
-20200501-heofwwlhofarzjdxmganprwuxe-story.html

4.  Millions of dads are stuck at home — which could be a game changer for working moms
URL: https://www.wbaltv.com/article/millions-of-dads-are-stuck-at-home-which-could-be-ga
me-changer-for-working-moms/32040142

5.  McConnell plan: $1,200 payments; $1T rescue takes shape
URL: https://www.wbaltv.com/article/mcconnell-plan-dollar1200-payments-dollar1t-rescue-t
akes-shape/31790273

6.  Family of 4 could get $3,000 under coronavirus relief plan, Treasury secretary says
URL: https://www.wbaltv.com/article/mnuchin-family-of-4-could-get-3k-under-virus-relief-
plan/31783643

7.  Maryland lawmakers consider ending legislative session early, advance emergency bill
 to fight coronavirus
URL: https://www.baltimoresun.com/coronavirus/bs-md-pol-ga-session-coronavirus-20200313-
qlbczkxchfew7bifyohsoyywyi-story.html

8.  World Health Organization declares coronavirus outbreak a pandemic
URL: https://www.wbaltv.com/article/world-health-organization-declares-coronavirus-outbr
eak-a-pandemic/31403007
```

Fig. 8: (6) Search results for the query "recession", *without* query expansion

```
##################################################
Model details (embedding, weighting_scheme): (one-hot, tf-idf)
Search engine initialized! Try the search engine:

Query: recession

1.  Fed takes emergency action, including slashing rate by full percentage point
URL: https://www.wbaltv.com/article/fed-takes-emergency-action-including-slashing-rate-b
y-full-percentage-point/31645643

2.  What a 20% unemployment rate would mean for America
URL: https://www.wbaltv.com/article/what-a-20-unemployment-rate-would-mean-for-america/3
1746821

3.  Millions of dads are stuck at home – which could be a game changer for working moms
URL: https://www.wbaltv.com/article/millions-of-dads-are-stuck-at-home-which-could-be-ga
me-changer-for-working-moms/32040142

4.  Baltimore-area governments are weighing big cuts as they brace for a collapse in rev
enues due to coronavirus
URL: https://www.baltimoresun.com/coronavirus/bs-md-coronavirus-economic-crisis-counties
-20200501-heofwwlhofarzjdxmganprwuxe-story.html

5.  'It's an unknown time': Baltimore bars and restaurants brace for decrease in busines
s amid coronavirus concern
URL: https://www.baltimoresun.com/coronavirus/bs-fo-covid19-bars-restaurants-20200312-iv
2otiqvm5b7zji3zfjw6racja-story.html

6.  Maryland Gov. Larry Hogan wants to tap millions of dollars in emergency funds to pre
pare for coronavirus
URL: https://www.baltimoresun.com/coronavirus/bs-hs-emergency-coronavirus-funding-202003
04-vfmzeekakfap7nn7yooz4x5ijq-story.html

7.  Unemployment filings skyrocket in Carroll County in sign of economic damage from cor
onavirus
URL: https://www.baltimoresun.com/coronavirus/cc-carroll-unemployment-claims-coronavirus
-20200403-zqq7trxya5danjrtzotfheigpa-story.html

8.  McConnell plan: $1,200 payments; $1T rescue takes shape
URL: https://www.wbaltv.com/article/mcconnell-plan-dollar1200-payments-dollar1t-rescue-t
akes-shape/31790273
```

Fig. 9: (6) Search results for the query "recession", *with* query expansion which accounts for the query's topic in general which can be about taxes, stocks, unemployment, relief funds, stimulus checks etc.

A.3: Table of All Model Permutations

dataset	embedding	weighting	p_0.25	p_0.5	p_0.75	p_1.0	p_mean1	p_mean2	r_norm	p_norm
cacm	one-hot	mean	0.319	0.1823	0.1066	0.0531	0.2026	0.2098	0.8372	0.5562
cacm	one-hot	tf-idf	0.4571	0.2723	0.1751	0.079	0.3015	0.3014	0.8681	0.6566
cisi	one-hot	mean	0.2673	0.1401	0.0822	0.0372	0.1632	0.1618	0.7782	0.5017
cisi	one-hot	tf-idf	0.3622	0.2137	0.1003	0.0388	0.2254	0.2189	0.8052	0.5577
med	one-hot	mean	0.6726	0.4527	0.285	0.0874	0.4701	0.442	0.9001	0.7507
med	one-hot	tf-idf	0.7346	0.5342	0.3599	0.0921	0.5429	0.4973	0.9122	0.7895
cran	one-hot	mean	0.5722	0.3527	0.2151	0.0941	0.38	0.3704	0.901	0.6807
cran	one-hot	tf-idf	0.6323	0.4226	0.2611	0.119	0.4386	0.4202	0.9111	0.7163
cacm	word2vec-google-news-300	mean	0.2535	0.129	0.0634	0.0178	0.1487	0.1656	0.8469	0.5152
cacm	word2vec-google-news-300	tf-idf	0.2882	0.147	0.0682	0.0167	0.1678	0.1776	0.8465	0.5183
cacm	word2vec-google-news-300	sif	0.2636	0.142	0.0641	0.0214	0.1566	0.1693	0.8548	0.5224
cacm	word2vec-google-news-300	usif	0.2994	0.1686	0.0879	0.039	0.1853	0.1906	0.8624	0.539
cacm	glove-twitter-100	mean	0.1647	0.0775	0.0345	0.0098	0.0922	0.1063	0.7798	0.406
cacm	glove-twitter-100	tf-idf	0.1846	0.0742	0.0339	0.0089	0.0976	0.1091	0.772	0.4015
cacm	glove-twitter-100	sif	0.1709	0.0808	0.0335	0.0099	0.0951	0.1099	0.7643	0.395
cacm	glove-twitter-100	usif	0.1747	0.0778	0.0339	0.0093	0.0955	0.1097	0.7906	0.4169
cacm	glove-wiki-gigaword-100	mean	0.2026	0.0972	0.0447	0.01	0.1148	0.1274	0.8333	0.4593
cacm	glove-wiki-gigaword-100	tf-idf	0.232	0.1143	0.0423	0.0092	0.1295	0.1397	0.8292	0.4636
cacm	glove-wiki-gigaword-100	sif	0.1939	0.094	0.0411	0.0127	0.1097	0.1253	0.806	0.4419
cacm	glove-wiki-gigaword-100	usif	0.2082	0.1031	0.0442	0.012	0.1185	0.1314	0.8366	0.4673
cacm	glove-wiki-gigaword-200	mean	0.2394	0.1183	0.0502	0.0123	0.136	0.1501	0.8512	0.4986
cacm	glove-wiki-gigaword-200	tf-idf	0.2838	0.1323	0.0491	0.011	0.1551	0.1642	0.8484	0.5055
cacm	glove-wiki-gigaword-200	sif	0.2192	0.11	0.0477	0.0171	0.1256	0.1427	0.8365	0.4874
cacm	glove-wiki-gigaword-200	usif	0.2355	0.1163	0.0513	0.0153	0.1344	0.1491	0.8599	0.5101
cacm	fasttext-wiki-news-subwords-300	mean	0.2874	0.1335	0.0648	0.013	0.1619	0.1689	0.8157	0.4918
cacm	fasttext-wiki-news-subwords-300	tf-idf	0.3072	0.1444	0.0645	0.0121	0.172	0.1788	0.8136	0.4974
cacm	fasttext-wiki-news-subwords-300	sif	0.2912	0.1404	0.0684	0.0128	0.1667	0.1757	0.8293	0.5009
cacm	fasttext-wiki-news-subwords-300	usif	0.304	0.1454	0.0718	0.013	0.1738	0.1825	0.8384	0.5163
cisi	word2vec-google-news-300	mean	0.2293	0.1285	0.0821	0.037	0.1466	0.1523	0.7901	0.5053
cisi	word2vec-google-news-300	tf-idf	0.2378	0.1331	0.0844	0.0366	0.1517	0.1585	0.7957	0.5127
cisi	word2vec-google-news-300	sif	0.2657	0.1526	0.0944	0.04	0.1709	0.1741	0.8289	0.5428
cisi	word2vec-google-news-300	usif	0.2592	0.1546	0.096	0.0409	0.1699	0.1749	0.8316	0.5451
cisi	glove-twitter-100	mean	0.1626	0.1028	0.0618	0.0318	0.109	0.1133	0.7235	0.4343
cisi	glove-twitter-100	tf-idf	0.1584	0.0936	0.0594	0.0319	0.1038	0.1095	0.7139	0.424
cisi	glove-twitter-100	sif	0.1748	0.1115	0.0676	0.0346	0.118	0.1237	0.7641	0.4617
cisi	glove-twitter-100	usif	0.1808	0.1123	0.0691	0.035	0.1207	0.1268	0.7663	0.4645
cisi	glove-wiki-gigaword-100	mean	0.2019	0.1189	0.0749	0.0344	0.1319	0.1372	0.7706	0.4828
cisi	glove-wiki-gigaword-100	tf-idf	0.2132	0.1226	0.0753	0.0344	0.137	0.1413	0.7727	0.4846
cisi	glove-wiki-gigaword-100	sif	0.227	0.1367	0.084	0.0379	0.1493	0.1538	0.8088	0.517
cisi	glove-wiki-gigaword-100	usif	0.2308	0.137	0.0837	0.0383	0.1505	0.1544	0.8093	0.5168
cisi	glove-wiki-gigaword-200	mean	0.2134	0.1257	0.0776	0.0346	0.1389	0.1439	0.7769	0.4945
cisi	glove-wiki-gigaword-200	tf-idf	0.2289	0.1278	0.0806	0.0344	0.1458	0.1519	0.7812	0.4993
cisi	glove-wiki-gigaword-200	sif	0.2431	0.1481	0.0909	0.0392	0.1607	0.1661	0.8205	0.5356
cisi	glove-wiki-gigaword-200	usif	0.243	0.1497	0.0909	0.0396	0.1612	0.1653	0.8213	0.5352
cisi	fasttext-wiki-news-subwords-300	mean	0.1967	0.1082	0.061	0.0333	0.122	0.1276	0.7219	0.4554
cisi	fasttext-wiki-news-subwords-300	tf-idf	0.234	0.1225	0.0672	0.0325	0.1412	0.1469	0.7398	0.4773
cisi	fasttext-wiki-news-subwords-300	sif	0.2006	0.1136	0.0716	0.0362	0.1286	0.1353	0.7678	0.4851
cisi	fasttext-wiki-news-subwords-300	usif	0.1996	0.1128	0.0703	0.0359	0.1276	0.1351	0.7653	0.4838
med	word2vec-google-news-300	mean	0.5805	0.389	0.2021	0.0633	0.3905	0.3676	0.8896	0.6992
med	word2vec-google-news-300	tf-idf	0.5335	0.3686	0.1866	0.0552	0.3629	0.337	0.8797	0.6779
med	word2vec-google-news-300	sif	0.5861	0.429	0.2347	0.0771	0.4166	0.3904	0.9042	0.7249
med	word2vec-google-news-300	usif	0.626	0.4336	0.2505	0.0806	0.4367	0.4034	0.9078	0.7325
med	glove-twitter-100	mean	0.3746	0.1858	0.0972	0.0417	0.2192	0.2157	0.7805	0.5414
med	glove-twitter-100	tf-idf	0.3622	0.1511	0.0803	0.0365	0.1979	0.1955	0.7652	0.5203
med	glove-twitter-100	sif	0.396	0.2031	0.1099	0.0456	0.2363	0.2308	0.7975	0.5609
med	glove-twitter-100	usif	0.4065	0.2081	0.1121	0.0474	0.2422	0.2373	0.7997	0.5652
med	glove-wiki-gigaword-100	mean	0.4346	0.2505	0.122	0.0439	0.269	0.2648	0.8224	0.5943
med	glove-wiki-gigaword-100	tf-idf	0.3977	0.2706	0.1086	0.0435	0.259	0.25	0.812	0.5808
med	glove-wiki-gigaword-100	sif	0.4503	0.3096	0.1261	0.0495	0.2953	0.2831	0.8335	0.6102
med	glove-wiki-gigaword-100	usif	0.4504	0.2856	0.1233	0.0488	0.2865	0.2801	0.8314	0.6069
med	glove-wiki-gigaword-200	mean	0.5305	0.3399	0.1635	0.054	0.3446	0.3219	0.855	0.649
med	glove-wiki-gigaword-200	tf-idf	0.5041	0.3239	0.1466	0.0518	0.3249	0.3029	0.8446	0.6337
med	glove-wiki-gigaword-200	sif	0.5424	0.368	0.1717	0.0577	0.3607	0.3348	0.8667	0.6644
med	glove-wiki-gigaword-200	usif	0.5287	0.3698	0.1685	0.0573	0.3557	0.3343	0.8651	0.6621
med	fasttext-wiki-news-subwords-300	mean	0.5118	0.3082	0.1415	0.0348	0.3205	0.3051	0.8144	0.6247
med	fasttext-wiki-news-subwords-300	tf-idf	0.5307	0.3069	0.1346	0.0383	0.3241	0.3041	0.8238	0.6286
med	fasttext-wiki-news-subwords-300	sif	0.5357	0.3425	0.1784	0.0515	0.3522	0.3293	0.8639	0.6618
med	fasttext-wiki-news-subwords-300	usif	0.5402	0.3552	0.18	0.0519	0.3585	0.3338	0.8647	0.6649
cran	word2vec-google-news-300	mean	0.4683	0.2577	0.1434	0.0527	0.2898	0.2873	0.8579	0.6009
cran	word2vec-google-news-300	tf-idf	0.4992	0.2789	0.157	0.0569	0.3117	0.3068	0.8395	0.5966
cran	word2vec-google-news-300	sif	0.4801	0.2782	0.1569	0.0672	0.3051	0.301	0.8816	0.6247
cran	word2vec-google-news-300	usif	0.4811	0.2882	0.1646	0.0697	0.3113	0.3066	0.8835	0.6305
cran	glove-twitter-100	mean	0.3118	0.1393	0.0735	0.0237	0.1749	0.1891	0.783	0.4847
cran	glove-twitter-100	tf-idf	0.3161	0.1354	0.0656	0.0263	0.1724	0.1869	0.7561	0.4629
cran	glove-twitter-100	sif	0.35	0.1575	0.0846	0.0275	0.1974	0.2077	0.8093	0.5113
cran	glove-twitter-100	usif	0.3525	0.1576	0.0836	0.0277	0.1979	0.2084	0.8108	0.5136
cran	glove-wiki-gigaword-100	mean	0.3486	0.1532	0.0792	0.0278	0.1937	0.2046	0.8099	0.514
cran	glove-wiki-gigaword-100	tf-idf	0.3495	0.1568	0.0812	0.0326	0.1958	0.2073	0.7903	0.5011
cran	glove-wiki-gigaword-100	sif	0.3606	0.1659	0.087	0.0338	0.2045	0.2159	0.8326	0.5335
cran	glove-wiki-gigaword-100	usif	0.3525	0.1581	0.0847	0.033	0.1984	0.2101	0.8299	0.5289
cran	glove-wiki-gigaword-200	mean	0.4123	0.2047	0.1061	0.0389	0.241	0.2465	0.8393	0.5612
cran	glove-wiki-gigaword-200	tf-idf	0.4353	0.216	0.1143	0.0471	0.2552	0.2596	0.8211	0.554
cran	glove-wiki-gigaword-200	sif	0.4267	0.225	0.1184	0.05	0.2567	0.2625	0.8638	0.5869
cran	glove-wiki-gigaword-200	usif	0.4225	0.2249	0.1168	0.0482	0.2547	0.2605	0.8618	0.5829
cran	fasttext-wiki-news-subwords-300	mean	0.4306	0.2156	0.1048	0.0318	0.2503	0.2563	0.804	0.5489
cran	fasttext-wiki-news-subwords-300	tf-idf	0.4517	0.218	0.1085	0.0354	0.2594	0.2651	0.7793	0.5387
cran	fasttext-wiki-news-subwords-300	sif	0.4585	0.2431	0.1338	0.0529	0.2785	0.2831	0.8619	0.595
cran	fasttext-wiki-news-subwords-300	usif	0.4594	0.2428	0.1355	0.0525	0.2792	0.2837	0.8618	0.5959
cacm	doc2vec	-	0.1515	0.0758	0.0401	0.01	0.0891	0.1013	0.7135	0.3746
cisi	doc2vec	-	0.1677	0.1021	0.0668	0.0385	0.1122	0.1148	0.744	0.45
med	doc2vec	-	0.2551	0.1661	0.1146	0.0339	0.1786	0.1724	0.7251	0.4818
cran	doc2vec	-	0.2423	0.0883	0.0412	0.017	0.1239	0.1423	0.764	0.4367