Intelligent Semantic Web Search

Jayaty Jayaty
North Carolina State University
jjayaty

Aishwarya Seth
North Carolina State University
aseth

I. Introduction

A vast majority of information has shifted online, web search has become one of the most important tools these days. However, just the possibility of having access to a humongous amount of data present online is not enough, if you are not able to extract relevant information. The ability to search across multitudes of web pages has advanced a lot but the ability to search intelligently and semantically is still relatively new. Users may not necessarily be well aware of the accurate keywords required to be used for giving them correct results. Therefore, a more lenient approach is required to provide relevant results to the users instead of a strict keyword search approach. In this project, we try to obtain the results of a web search query in a way that is not restricted to a strict keyword search. Thus, we title this type of search as an *Intelligent Semantic Web Search*.

Fig. 1. Keyword based web search

II. Usage Scenario

The intelligent and semantic web search is the next step towards making the user experience more seamless. This application can be integrated wherever a feature for user search is needed. For example, search engines, word meaning lookup on a website. This application can also be modified to integrate intelligence to semantic code search, for finding a specific piece of code in a huge project or code base or even online.

III. Special Features Of Application

The most special feature of the application is what we emphasize as the semantic search. The semantic search feature will empower the user to search what they want without exactly remembering the words or keywords for it. The feature will help in displaying the relevant results even in the absence of precise keywords as well. The feature is relatively recent and unorthodox from traditional tools. Search becomes easier by identifying entities and mapping unstructured data thereby removing the reliance on specific terminologies or precise memory/recall.

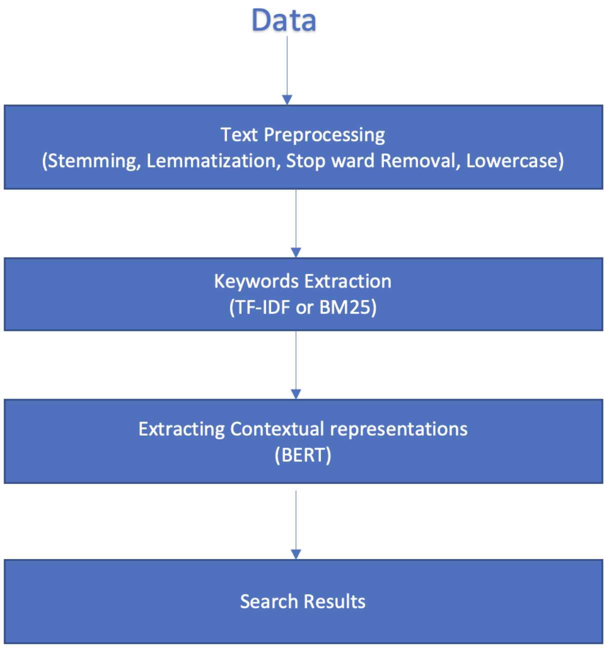

Data
Text Preprocessing
(Stemming, Lemmatization, Stop ward Removal, Lowercase)
Keywords Extraction
(TF-IDF or BM25)
Extracting Contextual representations
(BERT)
Search Results

Fig. 2. Intelligent semantic web search

IV. Methodology

Our intelligent semantic web search consists of three phases: Text Pre-processing, Keywords Extraction and Contextual Representation Extraction.

A. Text Pre-processing

The user query needs to be pre-processed first before proceeding with the task of searching. Our text-pre-processing techniques will include converting the query text into lowercase, stemming, lemmatization and finally stop words and punctuations removal(using the list of most common stop words available).

B. Keywords Extraction

The next step is using keywords extraction. Although we argue we need something more advanced than simple keyword extraction, we still agree that keyword extraction helps in development of initial relationships and contexts. We will be using either Tf-IDF or BM25 for keywords extraction.

C. *Contextual Representation Extraction*

We will be extracting contextual representation in order to get the core meaning of the query which user intends to search for. Getting the core meaning helps in getting similar words which might mean the same thing as the query and lead to actual results which the user is searching for. Contextual representation extraction have been achieved by obtaining embeddings and finally extracting similarity using advanced models like BERT. We use a variant of the BERT model called Sentence-BERT or SBERT for this task as it does not require training from scratch and can produce results faster.

V. Implementation Progress

The first phase and the third phase of the intelligent semantic web search have been completed successfully with preliminary results. The implementation described below has been done with the third phase in mind and will change in the next stage as outlined in the ’Planned Enhancements’ section. The implementation details can be summarized as follows:

A. *Dataset*

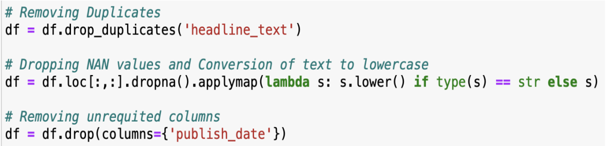
50k news headlines were used from ’A Million News Headlines’ dataset[11] published on Kaggle. The original dataset contains news headlines sourced from the reputable Australian news source ABC and published over a period of eighteen years. It is in the form of a csv file and has 2 columns namely, ’published date’ and ’headline text’.

B. *Data Pre-processing*

Since the project is still at a preliminary stage, only basic preprocessing has been done till now. We intend to develop further on this and refine our dataset as detailed in the ’Planned Enhancements’ section.
The 50k news headlines were loaded into a Pandas dataframe and following forms of text pre-processing techniques were used to polish the dataset:

1) Removing duplicates
2) Dropping NAN values
3) Conversion of text to lowercase
4) Removal of unrequited column: ’published date’

The last step was done on the basis of general observation that the published date of a news article would not lead to any significant improvement in the extraction of semantic similarity. Hence, it was decided to be dropped from the dataset.

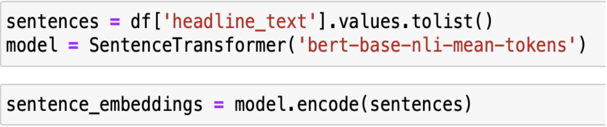


```
# Removing Duplicates
df = df.drop_duplicates('headline_text')

# Dropping NAN values and Conversion of text to lowercase
df = df.loc[:,:].dropna().applymap(lambda s: s.lower() if type(s) == str else s)

# Removing unrequited columns
df = df.drop(columns={'publish_date'})
```

Fig. 3. Text Pre-processing

C. *Model*

We have used a variation of the BERT model, Sentence Bert aka SBERT model, for extracting inferences and semantic meaning from the texts of the news articles. The process requires to first generate the word embeddings of the texts. The embeddings are then used for determining the semantic similarity between the news articles and the searched query. We have used cosine similarity for the task of determining semantic similarity between news articles and the search query provided by the user.



```
sentences = df['headline_text'].values.tolist()
model = SentenceTransformer('bert-base-nli-mean-tokens')
```

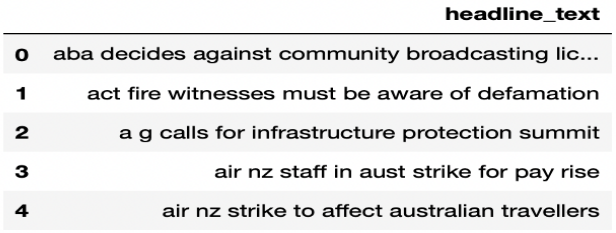
```
sentence_embeddings = model.encode(sentences)
```

Fig. 4. Model execution

VI. Preliminary Results

A. *Processed Data*

After executing the data pre-processing steps mentioned above, the resultant dataset is as shown below. The SBERT model chosen to generate embeddings and contextual representations is executed on the dataset shown.

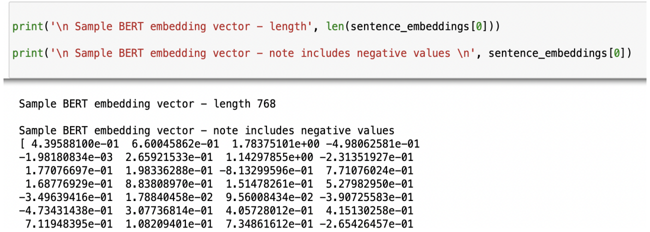


	headline_text
0	aba decides against community broadcasting lic...
1	act fire witnesses must be aware of defamation
2	a g calls for infrastructure protection summit
3	air nz staff in aust strike for pay rise
4	air nz strike to affect australian travellers

Fig. 5. Processed Data

B. *Embedding Vectors generated by SBERT model*

After running the SBERT model, the embedding vectors generated from news articles are as shown below in the snapshot.



```
print('\n Sample BERT embedding vector - length', len(sentence_embeddings[0]))

print('\n Sample BERT embedding vector - note includes negative values \n', sentence_embeddings[0])
```

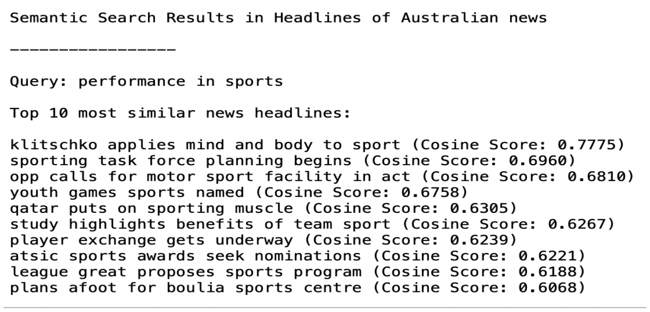
```
Sample BERT embedding vector - length 768

Sample BERT embedding vector - note includes negative values
[ 4.39588100e-01  6.60045862e-01  1.78375101e+00 -4.98062581e-01
-1.98180834e-03  2.65921533e-01  1.14297855e+00 -2.31351927e-01
 1.77076697e-01  1.98336288e-01 -8.13299596e-01  7.71076024e-01
 1.68776929e-01  8.83808970e-01  1.51478261e-01  5.27982950e-01
-3.49639416e-01  1.78840458e-02  9.56008434e-02 -3.90725583e-01
-4.73431438e-01  3.07736814e-01  4.05728012e-01  4.15130258e-01
 7.11948395e-01  1.08209401e-01  7.34861612e-01 -2.65426457e-01
```

Fig. 6. Embedding Vectors

C. Preliminary output

To verify our semantic search results, we input four search queries to find the most semantically related headlines using the cosine similarity score. The outputs after searching the respective queries have been shown below along with the respective queries and the snapshots:

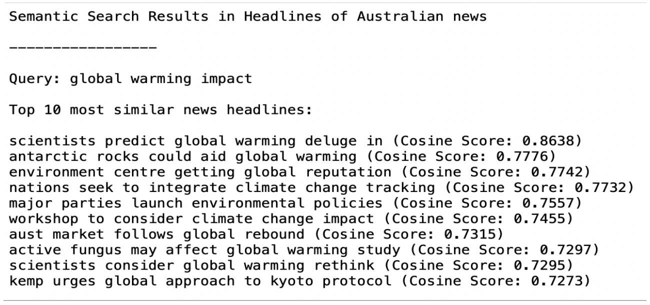
1) ’performance in sports’



```
Semantic Search Results in Headlines of Australian news

-----------------

Query: performance in sports

Top 10 most similar news headlines:

klitschko applies mind and body to sport (Cosine Score: 0.7775)
sporting task force planning begins (Cosine Score: 0.6960)
opp calls for motor sport facility in act (Cosine Score: 0.6810)
youth games sports named (Cosine Score: 0.6758)
qatar puts on sporting muscle (Cosine Score: 0.6305)
study highlights benefits of team sport (Cosine Score: 0.6267)
player exchange gets underway (Cosine Score: 0.6239)
atsic sports awards seek nominations (Cosine Score: 0.6221)
league great proposes sports program (Cosine Score: 0.6188)
plans afoot for boulia sports centre (Cosine Score: 0.6068)
```

Fig. 7. Top 10 most similar news headlines for the query: ’performance in sports’

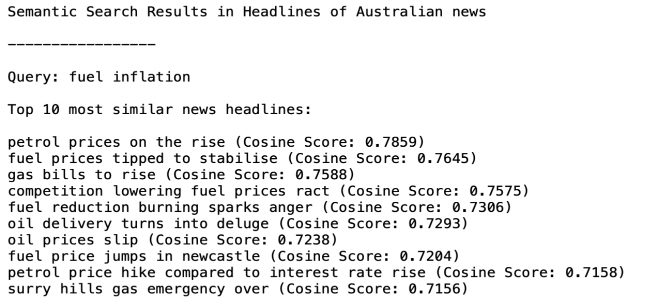
2) ’Global Warming Impact’



```
Semantic Search Results in Headlines of Australian news

----------------

Query: global warming impact

Top 10 most similar news headlines:

scientists predict global warming deluge in (Cosine Score: 0.8638)
antarctic rocks could aid global warming (Cosine Score: 0.7776)
environment centre getting global reputation (Cosine Score: 0.7742)
nations seek to integrate climate change tracking (Cosine Score: 0.7732)
major parties launch environmental policies (Cosine Score: 0.7557)
workshop to consider climate change impact (Cosine Score: 0.7455)
aust market follows global rebound (Cosine Score: 0.7315)
active fungus may affect global warming study (Cosine Score: 0.7297)
scientists consider global warming rethink (Cosine Score: 0.7295)
kemp urges global approach to kyoto protocol (Cosine Score: 0.7273)
```

Fig. 8. Top 10 most similar news headlines for the query: ’Global Warming Impact’

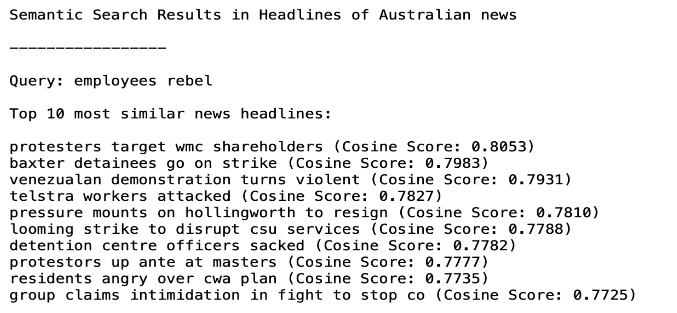
3) ’Fuel Inflation’



```
Semantic Search Results in Headlines of Australian news

----------------

Query: fuel inflation

Top 10 most similar news headlines:

petrol prices on the rise (Cosine Score: 0.7859)
fuel prices tipped to stabilise (Cosine Score: 0.7645)
gas bills to rise (Cosine Score: 0.7588)
competition lowering fuel prices ract (Cosine Score: 0.7575)
fuel reduction burning sparks anger (Cosine Score: 0.7306)
oil delivery turns into deluge (Cosine Score: 0.7293)
oil prices slip (Cosine Score: 0.7238)
fuel price jumps in newcastle (Cosine Score: 0.7204)
petrol price hike compared to interest rate rise (Cosine Score: 0.7158)
surry hills gas emergency over (Cosine Score: 0.7156)
```

Fig. 9. Top 10 most similar news headlines for the query: ’Fuel Inflation’

4) ’employees rebel’



```
Semantic Search Results in Headlines of Australian news

-----------------

Query: employees rebel

Top 10 most similar news headlines:

protesters target wmc shareholders (Cosine Score: 0.8053)
baxter detainees go on strike (Cosine Score: 0.7983)
venezualan demonstration turns violent (Cosine Score: 0.7931)
telstra workers attacked (Cosine Score: 0.7827)
pressure mounts on hollingworth to resign (Cosine Score: 0.7810)
looming strike to disrupt csu services (Cosine Score: 0.7788)
detention centre officers sacked (Cosine Score: 0.7782)
protestors up ante at masters (Cosine Score: 0.7777)
residents angry over cwa plan (Cosine Score: 0.7735)
group claims intimidation in fight to stop co (Cosine Score: 0.7725)
```

Fig. 10. Top 10 most similar news headlines for the query: ’employees rebel’

VII. Planned Enhancements

Future enhancements to the current implementation can be categorized into data augmentation, keyword extraction, as described in the second phase, and transfer learning. The enhancements have been elaborated as follows:

A. Data Augmentation

We plan to increase our data because of two reasons. First reason is to simply increase the dataset size in order to improve the performance of our model. Second reason is to be more inclusive of the various categories of search queries on the web or, in other words, generalize better. For example, users can query anything from sports to news to medical information or general trends. In order to achieve this, we have planned to merge various categories of datasets to get an aggregate dataset comprising of news articles, medical articles, technical articles and general articles.

B. Keyword Extraction

As described in the ’Methodology’ section, we intend to incorporate keyword extraction in order to improve the accuracy of our intelligent semantic web search. Although our project is more focused on contextual representations, keyword extraction is a powerful technique which we can integrate for improving the performance of our model significantly. This will be achieved by either using TF-IDF or BM-25 on the texts of the articles and extracting top keywords from it after further text pre-processing techniques such as stop words removal.

C. Transfer Learning

Our SBERT model has been fine tuned on NLI dataset originally. Since we plan to increase our dataset multifold, we intend to split our dataset and perform transfer learning on our variation of BERT model, SBERT or Sentence BERT. This will lead to increase in the expanse of topics covered which has not been seen by our model before and significantly improve the model performance.

References

[1] Jui-Ting Huang, Ashish Sharma, Shuying Sun, Li Xia, David Zhang, Philip Pronin, Janani Padmanabhan, Giuseppe Ottaviano, Linjun Yan, ”Embedding-based Retrieval in Facebook Search”, CoRR abs/2006.11632, 2020 https://arxiv.org/abs/2006.11632

[2] Pascal Hitzler, ”A Review of the Semantic Web Field”, Communications of the ACM Volume 64Issue 2February 2021 pp 76–83 https://doi.org/10.1145/3397512

[3] Hamel Husain, ”How To Create Natural Language Semantic Search For Arbitrary Objects With Deep Learning”, https://towardsdatascience.com/semantic-code-search-3cd6d244a39c

[4] Yusuf Sermet, Ibrahim Demir, ”A Semantic Web Framework for Automated Smart Assistants: COVID-19 Case Study”, arXiv:2007.00747, https://arxiv.org/abs/2007.00747

[5] ”A decade of Semantic Web research through the lenses of a mixed methods approach”, Semantic Web, vol. 11, no. 6, pp. 979-1005, 2020, https://content.iospress.com/articles/semantic-web/sw200371

[6] ”The Semantic Web identity crisis: In search of the trivialities that never were”, Semantic Web, vol. 11, no. 1, pp. 19-27, 2020, https://content.iospress.com/articles/semantic-web/sw190372

[7] ”Information extraction meets the Semantic Web: A survey”, Semantic Web, vol. 11, no. 2, pp. 255-335, 2020, https://content.iospress.com/articles/semantic-web/sw180333

[8] Nils Reimers, Iryna Gurevych,”Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 3982–3992, Hong Kong, China, November 3–7, 2019, https://aclanthology.org/D19-1410.pdf

[9] Kexin Wang, Nils Reimers, Iryna Gurevych, ”TSDAE: Using Transformer-based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning”,arXiv:2104.06979v3 [cs.CL] 10 Sep 2021, https://arxiv.org/pdf/2104.06979.pdf

[10] Nils Reimers and Iryna Gurevych, ”Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation””, arXiv:2004.09813v2 [cs.CL] 5 Oct 2020, https://arxiv.org/pdf/2004.09813.pdf

[11] https://www.kaggle.com/godeep48/a-million-news-wordcloud-on-kmeans/data