Assignment 2 - CSI4107

Students

Samuel Daly - 8173488
Ryan Matte - 300027432
Michel Moore - 300063096

Task Division

Experiment 1: Samuel Daly & Michel Moore

Experiment 2: Samuel Daly & Ryan Matte

Folder Layout

Experiment 1

```
 ├── Modules
 │   ├── data
 │   │   ├── results.txt
 │   │   ├── topics_MB1-49.xml
 │   │   ├── trec-microblog11-qrels.txt
 │   │   ├── trec-microblog11.txt
 │   │   └── tweets_embeddings_dict.json (Not included in folder since file is too big)
 │   ├── processor.py
 │   └── query.py
 ├── universal-sentence-encoder_4 (Not included in folder since file is too big)
 ├── requirements.txt
 ├── embed.py
 └── main.py
```

Experiment 2

```
 ├── Modules
 │   ├── data
 │   │   ├── document_word_dict.json
 │   │   ├── frequency_dict.json
 │   │   ├── results.txt
 │   │   ├── stopwords.txt
 │   │   ├── topics_MB1-49.xml
 │   │   ├── trec-microblog11-qrels.txt
 │   │   ├── trec-microblog11.txt
 │   │   └── weighted_dict.json
 │   ├── word2Vec
 │   │   ├── word2vec_twitter_model.bin (Not included in folder since file is too big)
 │   │   ├── word2vecReader.py
 │   │   └── word2vecReaderUtils.py
 │   ├── indexer.py
 │   ├── preprocessor.py
 │   └── query.py
 ├── requirements.txt
 ├── create_index.py
 └── create_results.py
```

Links to download the missing files

- Universal Sentence Encoder
- Word2vec Twitter Model

How to run the program

Experiment 1

topics_MB1-49.txt must be converted to XML

1. Rename topics_MB1-49.txt to topics_MB1-49.xml
2. Add to first line: `<data>`
3. Add to last line: `</data>`

Windows

1. To install necessary packages, run ***pip install -r requirements.txt***
2. To run, run ***python main.py***

MacOS

1. To install necessary packages, run ***pip3 install -r requirements.txt***
2. To run, run ***python3 main.py***

Experiment 2

topics_MB1-49.txt must be converted to XML

1. Rename topics_MB1-49.txt to topics_MB1-49.xml
2. Add to first line: `<data>`
3. Add to last line: `</data>`

Windows

1. To install necessary packages, run ***pip install -r requirements.txt***
2. To create the inverted index, run ***python create_index.py***
3. To create the results, run ***python create_results.py***

MacOS

1. To install necessary packages, run ***pip3 install -r requirements.txt***
2. To create the inverted index, run ***python3 create_index.py***
3. To create the results, run ***python3 create_results.py***

Functionality of the program

For experiment 1, a lot of code was changed. We decided to use the Universal Sentence Encoder found on TensorFlow-hub. On the website, there was a pre-trained model which we used for this experiment. To start off, we removed some of the pre-processing we did in the original assignment, since we did not really need to process the tweets much. The only processing that we did was removing links and removing words that contained numbers. Next, we embedded each tweet using the Universal Sentence Encoder. Once embedded, we save the vector of each tweet in a dictionary and attached it to the tweet number. This allowed us to know which vector belonged to which tweet. The following step was to find the similarity score between the query and documents. For this, we loop through the queries, take the text of the query, do the same processing as earlier on said query and then run the query through a loop where it finds the similarity score between the query and each document. For this we use ***np.inner()*** which returns the inner product of 2 vectors and gives us the similarity between the document and the query. We then sort all the scores and save the top 1000 documents in the file called ***results.txt***.

For experiment 2, not much was needed to be changed in terms of the code. The goal of this experiment was to do some query vector modification or query expansion based on pretrained word embeddings. We ended up using a Word2Vec model built on a Twitter Corpus and some code built by Loreto Parisi which we found on GitHub. (https://github.com/loretoparisi/word2vec-twitter) This code allowed us to run our previous code without many changes. The code we found on GitHub allowed us to find similar words for the words that appeared in each query. By finding those similar words or synonyms, it allowed us to expand the query.

After expanding the query, we needed to re-vectorize the query. To do that, we just used the code we previously built for assignment 1, but with some slight modifications given that we were passing a list of in the function instead of a full query. With the query re-vectorized, we were able to build new results and test them using the TREC evaluation method. Those results can be found in the ***Results*** section of this file.

Results

Previous Results

```
 map                    all 0.2771
P_10                    all 0.3020
```

Experiment 1

```
 map                    all 0.2304
P_10                    all 0.2878
```

Experiment 2

```
 map                     all 0.2076
P_10                    all 0.1735
```

Discussion

Looking at theses results, we can see that they are not as good as the results we achieved in our first assignment. The results for experiment 1 are lower due to the fact that we take the whole sentence instead of just taking keywords like we did in A1. Not as much pre-processing is required for the universal sentence encoder and this means we will get a lower score given that it includes stopwords and it does not stem the words. For experiment 2, we can see an even lower score because we used query expansion. Query expansion itself is used to increase the quality of the search results but that comes at the expense of precision. This is why we are seeing a lower MAP and a lower Precision for the first 10 documents.

Sample queries

Experiment 1

These queries are the result of running experiment number 1.

Query 3: "Haiti Aristide return"

Results

1. 33254598118473728 Haiti/ Presidenzi http://www.worldonlinereview.com/italia/2011/02/03/haiti-presidenziali-ballottaggio-tra-ex-first-lady-e-cantante-virgilio/
2. 34410414846517248 ARISTIDE SERAIT DE RETOUR EN HAITI
3. 32439513519230976 Haiti Noir http://bit.ly/hLMKjG #haiti
4. 33325579583365120 Haiti poll revised http://bit.ly/gQJU0s #Haiti
5. 35032969643175936 Haiti will raise again
6. 29296574815272960 Haiti – Aristide : His return, an international affair… – Haitilibre.com http://bit.ly/gzyLXG #haiti
7. 34682906718908416 Haiti: One Year Later, Acupuncturists Return http://www.acupuncturetoday.com/mpacms/at/article.php?id=32366
8. 34518512273719296 Haiti overwhelmed by dead, devastation http://bit.ly/gcj56c #Haiti
9. 35088534306033665 Haiti concede passaporte a Aristide.
10. 34896269163896832 Haiti: Verjagter Ex-Präsident Aristide will heimkehren http://bit.ly/h7EHvx

Query 20: "Taco Bell filling lawsuit"

Results

1. 30727342653444098 Taco Bell buzz on the 'beef' class-action lawsuit http://www.latimes.com/health/boostershots/la-taco-bell-beef-buzz-20110126,0,6128287.story
2. 32865855435968512 Taco Bell Has Beef with Lawsuit’s Claims http://zoo.mn/euh1Oe
3. 31101259872215040 Taco Bell issues response to lawsuits: suck it. http://tinyurl.com/4qqwhlz
4. 31948737517453312 Taco Bell answers lawsuit: "Yes, that is beef" - http://on.msnbc.com/fmYKNX
5. 30004107020337152 Taco Tuesday indeed... Taco Bell sued: Lawsuit filed in beef over Taco Bell 'meat' - Sun-Sentinel.com http://bit.ly/hnk6vo // CC: @XeL13
6. 30310229434437632 GROSS! Taco Bell Sued for Bogus Beef http://adage.com/u/R2p5wa
7. 30962906484969472 Taco Bell Fights Back On Beef Lawsuit http://dlvr.it/FH1pt
8. 29356462186696704 Free taco bell ! #forthewin
9. 31085134354583552 Taco Bell has a beef with meat lawsuit .. http://tinyurl.com/669maut
10. 31052864197496832 Taco BELL* sued for serving beef that is 35% beef..I lied.

Experiment 2

These queries are the result of running experiment number 2.

Query 3: "Haiti Aristide return"

Results

1. 34950800157450240 John Baer: Who didn't see this coming?: TO THOSE who know Ed and Midge Rendell - heck, to the Philly world at la... http://bit.ly/ii6WEO
2. 29555143699603456 1-2-3 Spa! Relax and Unwind at Half Moon, A RockResort. Book 3 nights or more at Half Moon Jamaica, and receive a...

http://fb.me/NXexbkC6

3. 33274580210556928 http://go-jamaica.com/news/read_article.php?id=26130 #Jamaica #Prime Mnister to launch #climatechange programme
4. 30702782788927489 feednews: Jamaica: “Dog-Paw”: Written by Janine Mendes-Franco “The cliche that truth is stranger than fiction is... http://bit.ly/gKF45u
5. 30701708191465472 Jamaica: “Dog-Paw”: Written by Janine Mendes-Franco “The cliche that truth is stranger than fiction is true”: Ac... http://bit.ly/hPf8SD
6. 34768755347169280 LIRR service is experiencing eastbound delays between 5 to 10 minutes through Jamaica due to an earlier train with equipment problems.
7. 28984571475271680 RT @Joe_Taxi: RT @mediahacker: S. Africa and Cuba negotiating to facilitate return of Aristide http://www.timeslive.co.za/sundaytimes/article866448.ece/I-want-to-go-home-says-Haitis-Aristide … #Haiti
8. 29296574815272960 Haiti – Aristide : His return, an international affair… – Haitilibre.com http://bit.ly/gzyLXG #haiti
9. 32387196078006272 Haiti allows ex-president's return: Jean-Bertrand Aristide, who was Haiti's first democratically elected leader,... http://aje.me/fQ4j4T
10. 32211683082502144 #int'l #news: Haiti opens door for return of ex-president Aristide: PORT-AU-PRINCE (Reuters) - Haiti'... http://bit.ly/gSIFwd #singapore

Query 20: "Taco Bell filling lawsuit"

Results

1. 29906116062220290 Lawsuit: Taco Bell Ground Beef Is Really Just "Meat Filling" - @consumerist http://consumerist.com/2011/01/lawsuit-says-taco-bell-ground-beef-is-really-just-taco-meat-filling.html?utm_source=streamsend&utm_medium=email&utm_content=13297631&utm_campaign=Fo
2. 31082136219947008 Taco Bell Counters 'Meat Filling' Charges in Lawsuit With Print, Web Effort http://goo.gl/fb/H9wcB
3. 31043176684851200 Oxymoron alert: "The lawsuit is bogus & filled with completely inaccurate facts" Taco Bell President said. Inaccurate facts? Freudian slip?
4. 29853985930219520 That ain't necessarily "beef" in your Taco Bell burrito...a new lawsuit wants the chain to label it "taco meat filling" instead.
5. 30344959127191552 Obama: Let's rein in frivolous lawsuits. Kucinich then sues House cafeteria over olive pit in sandwich. #badtiming http://bit.ly/fgFk1j
6. 30986819369697281 @nflnetwork "The Rudy Rule" Sounds Like Something To Prevent Lawsuits. Black Players,coaches,& GM's Should All The Opportunity As White Ones
7. 32443636847218688 The Ins And Out Of Small Claims: Filling a civil lawsuit against an organization or person in hopes of col... http://tinyurl.com/4s8h474
8. 29158340424634368 Lawsuit Loans – Filling the Need For Funding http://bit.ly/hYuw1C
9. 30700675742572545 I wonder what makes all the people who want taco bell to go vegetarian think that they will put real vegetables in the filling.
10. 30012275423186944 Hey, @H0TMessBarbie, you hear Taco Bell is being sued because their beef filling is only 35% beef? I told you it makes a good enema!