## Group Members and Work Breakdown:

Kareem Dabbour - 300082990

* Handled vectorizing and ranking of documents
  + Wrote vectorizeDocs, getMaxFreq, normalize methods and Index.py class
* Wrote reRank method for BERT retrieval
* Wrote the “Tokens”, “Evaluation of Queries”, and sections in the report

Irtiza Hasan - 300074069

* Handled tokenization and preprocessing of queries and documents
  + Wrote tokenizeStr, preprocStr, processStopWords, tokenizeDocs, and tokenizeQuery methods
* Wrote queryExpand method for Query Expansion
* Wrote the, “Functionality”, and “Algorithm” sections in the report

Andrea Herscovich - 300057058

* Handled parsing of queries and documents, and saving results
  + Wrote makeQuery, bulkQuery, getDocs, getQueries, and saveToFile
* Wrote “Instructions” and “Discussion” section in the report

## Instructions on How to Run Program:

The program is written in python. The most updated version of python can be found [here](https://www.python.org/downloads/).

To make the installation process easier, we have provided a requirements.txt file that includes all the libraries used in this project. To use this, please ensure that a package manager for python is installed. We recommend pip, with the installation found in this [documentation](https://pip.pypa.io/en/stable/installation/#supported-methods). Executing this script will install all libraries used to run this program.

After changing the directory to the CSI4107\_A2\_MicroblogIR, execute the following command in the terminal window:

| pip install -r /path/to/requirements.txt |
| --- |

After installing all the libraries, the program can be run using the command in the specific model folder(base, bert, queryExpand):

| python3 system.py |
| --- |

**Note:**

Query Expansion: The “glove-twitter-200” corpus used for query expansion is quite large at 758 MB, and it will need to be downloaded from the internet to run the program. It takes about 8 minutes to download. The model will also need to be loaded whenever the program is executed. This process will take 2-3 minutes on each program execution.

BERT: Training the model on a Macbook Pro with the following specifications:

* 2.6 GHz 6-Core Intel Core i7
* 32 GB 2667 MHz DDR4
* AMD Radeon Pro 5500M 8 GB

Took ~120 minutes

Verify Results:Go to .results/Results.txt to view the best performing ranked results per query. In our program, the best performance was through query expansion.

## Functionality of Program:

This program is an information retrieval system for a collection of documents based on a query string. For this assignment there are 45,899 Twitter messages that are ranked based on its relevance to a given query. The ranking process includes stripping each document to its essential tokens, and then evaluating its relevance based on the provided query. The program uses term frequency–inverse document frequency (tf-idf) to evaluate the statistical significance of each query token compared to the content of each document Tf-idf is the ratio of the frequency of a given term compared to the number of documents in the dataset that contain the term. The terms that appear in both the query and document have their tf-idf score calculated. This program improves upon basic information retrieval, by using the neural methods of query expansion and vector re-ranking.

Query expansion is a process that uses semantic knowledge to find related words or concepts of a query. Related words can be synonyms, fixing spelling errors, various morphological forms, and semantically related words such as antonyms or meronyms. The query expansion in this program uses a model from Gensim, an open source semantic dataset. As the documents itself are a collection of Tweets on Twitter, the model used in this project is the glove-twitter-200, which includes 2 billion tweets, and 27 billion tokens.

Bidirectional Encoder Representations from Transformers (BERT), is a neural model that classifies queries into a multi-class classification problem by predicting the intent label for each query. Find the intent of a query can provide a relevant context to a word, which can then be encoded into a vector that represents that word. The BERT model used in our program was trained on the 1000 ranked documents returned, and the corresponding query. The model is trained to vectorize the documents and queries. The cosine similarity is found using these new vectors, and the documents are re-ranked.

## Algorithm:

In this program, the following functionality from external libraries are used:

* [Spacial Distance](https://docs.scipy.org/doc/scipy/reference/spatial.distance.html) from Scipy
  + Compute the Cosine distance between 1-D arrays used in BERT
* [Vectorizer](https://pypi.org/project/sent2vec/0.1.1/) from Sent2vec
  + Computes vectors in for BERT sentence embedding used for reranking
* [Downloader](https://radimrehurek.com/gensim/downloader.html) from gensim
  + The dataset model utilizing the GloVe Twitter Corpus used for query expansion
* [ElementTree](https://docs.python.org/3/library/xml.etree.elementtree.html) from XML
  + Parse and read the queries which are in an XML format

There are separate folders for each of the algorithms used: baseline, BERT, and query expansion. The structure of the code in each folder is organized into 3 files:

* Utils.py - Contains all util methods used to read and parse files, tokenize documents, calculate the document vector lengths, generate normalized tf-idf matrix, reranking algorithm, and query expansion
* Index.py - Calculates the the tf-idf for the documents and queries
* System.py - Main file that runs the program

1. **Parsing and Tokenizing Documents and Queries:**

* **Documents:**
  + Read each line of the Trec\_microblog11.txt file, and split the documentId and text
  + The text of the document was split on a space, and each individual word was tokenized with the following preprocesses:
    - Making all the letters lowercase
    - Removing leading and trailing white spaces.
    - Removing all URLs
    - Replacing punctuation with spaces
      * !"#$%&\'()\*+,-./:;<=>?@[\\]^\_`{|}~'...
    - Removing all numbers
    - Remove all empty tokens
    - We did not opt to use a stemmer similar to our Java program, as it adversely impacted performance
  + Store the preprocessed strings and document ids in a Dictionary
    - The parsed documents will be used throughout the program, so it is important to ensure that lookups are efficient. For this reason, a dictionary was used to ensure that documents can be retrieved in O(1) constant time.
* **Queries:**
  + As the queries are stored in an XML format, the library ElementTree is used to assist in the process, by creating individual ‘Element’ objects which can be parsed to make a Dictionary of queryNumber and queryText
  + The queries are then tokenized using the same process as described for documents

E.g. The query in the XML format:

<top>  
<num> Number: MB001 </num>  
<title> BBC World Service staff cuts </title>  
<querytime> Tue Feb 08 12:30:27 +0000 2011 </querytime>  
<querytweettime> 34952194402811904 </querytweettime>  
</top>

Is parsed and tokenized into the format: 1: ['bbc', 'world', 'service', 'staff', 'cuts']

1. **Index Matrix for Documents**

* The Index class is used to organize the logic for indexing the documents, and building the document:tf matrix for each token. It includes 2 instance variables:
  + Index: A dictionary of the tf-idf matrix for all unique tokens in the set of documents is constructed
  + numDocs: The number of documents that have been indexed, used to calculate the inverse document frequency (IDF)
* The algorithm for creating the dictionary can be described as follows:
  + Every document has each of its tokens counted and then the term frequency calculated for each unique token :
    - 1 + log(tf) if the term frequency of the token in the document is greater than 0, otherwise 0
  + The entries are stored in a map with the token as the key, and an array of all documents that contain the word alongside its tf value
* The Index matrix is a dictionary with the token as the key, and a list of tuples with the document ids, and the tf-idf value. The dictionary will ensure fast retrieval times, while the tuple helps organize the structure of the documents and their respective tf-idf values.

E.g. The tf- of the word “fascinating”, showing all document ids that contain the word alongside the tf-idf value:

[('34649950826532864', 1.0), ('29182077568032769', 1.0), ('31301306610089985', 1.0), ('32806556273737728', 1.0), ('30129249314676737', 1.0), ('29328130359427072', 1.0), ('33182072650604544', 1.0), ('32931913329020928', 1.0), ('31072676713865216', 1.0), ('30469363869941760', 1.0), ('31503942152101889', 1.0)]

1. **Vector Weightings**

* Each document has a computed vector that is generated by taking the summation of all tf-idf values per token
* The vector is then normalized by squaring all the weighted term frequencies, and then taking the square root
* The ddd.qqq weighting scheme ended up following *lnc.ltc*.
* The vector weights are stored in a dictionary with the document id as the key, and the vector length as the value
  + These weights will be used in the cosine normalization of the query-document dot product which is used to rank the documents from queries

E.g. For the first document with id 34952194402811904, the unique tokens are ['bbc', 'world', 'service', 'staff', 'savage', 'cuts'], with each token having a tf-idf value of 1.0. The vector length is the square root of the sum of squares of each token’s weighted term frequency. This value is calculated as 2.44948974278

1. **Retrieving and Ranking Queries:**

* Each query has a list of ranked documents retrieved
  1. Each unique token in the query has its tf-idf calculated, and a normalized query vector generated representing the sum of all tokens’ tf-idf values
     + The query tf-idf is calculated by retrieving the idf value of each query term, and then multiplying it by the ratio of the occurrence of the term relative to the most frequent query term
  2. The index matrix for each unique query token is iterated on. For each entry in the list:
     + We build up the dot product by multiplying the query weight by the doc weight. This can be thought of as for document vector and query vector .
     + If this document has an entry in the return map already, simply add the two and update the map. This can be seen as doing the following: where the is the current token being evaluated.
  3. The dot product for each document is normalized using queryVecLen as well as the document vector length that is retrieved from the docVectLengths.
  4. The normalized dot products are sorted in descending order with the top 1000 ranked documents returned

1. **Query Expansion**

* Using the model from the “glove-twitter-200” corpus, each of the 50 queries were expanded by a maximum factor of 2:
  + For each word in the list of queries, the most similar word was appended to the query using the method: most\_similar
    - From the GenSim docs: “This method computes cosine similarity between a simple mean of the projection weight vectors of the given words and the vectors for each word in the model.”
    - This returns the n most similar words, each with a vector score between [0,1]. From our tests, we concluded that returning the single most relevant word, with a threshold of at least 0.55 resulted in the most precision

Query 1: “ BBC World Service staff cuts “

Tokenized words: ['bbc', 'world', 'service', 'staff', 'cuts']

After Query Expansion: ['bbc', 'world', 'service', 'staff', 'cuts', 'office', 'cut', 'services', 'itv']

Query 2: “ 2022 FIFA soccer“

Tokenized words: ['fifa', 'soccer']

After Query Expansion: ['fifa', 'soccer', 'xbox', 'football']

Query 3: “Haiti Aristide return“

Tokenized words: ['haiti', 'aristide', 'return']

After Query Expansion: ['haiti', 'aristide', 'return', 'president', ‘portauprince’]

**Note:** The “glove-twitter-200” corpus is quite large at 758 MB, and it will need to be downloaded from the internet to run the program. It takes about 8 minutes to download. The model will also need to be loaded whenever the program is executed. This process will take 2-3 minutes on each program execution.

1. **Rerank Original Results using BERT**

* Using the library sent2vec the BERT model is trained through the original results, queries and documents
  + Pass in each of the 49 queries, alongside the corresponding document as sentences in the BERT model, creating a set of vectors
  + The model tokenizes and vectorizes the queries as well as the documents which are then returned
  + Replace all negative vector values with 0 (compare the first quadrant)
  + The cosine similarity for each new document vector is recalculated with the query vector
  + The new ranked results are returned

1. **Saving to File:** The ranked documents for each query are saved to the results file

## Tokenization

After the tokenization process, punctuation, numbers, stopwords, URLs, leading and trailing white spaces, were removed. The following is an example of 5 documents before and after the tokenization process, as well as 100 tokens from the vocabulary set.

After preprocessing the documents, the number of unique terms in our vocabulary set was 62774. The number of total words in the set of documents was 711508 so this tokenization process marked a 1033.44% decrease.

An example of the first 5 documents, and their tokenized words:

| DocId: ﻿34952194402811904 Original Doc: save bbc world service from savage cuts  Tokens: ['bbc', 'world', 'service', 'savage', 'cuts']  DocId: 34952186328784896 Original Doc: a lot of people always make fun about the end of the world but the question is are u ready for it  Tokens: ['lot', 'people', 'fun', 'world', 'question']  DocId: 34952041415581696 Original Doc: rethink group positive in outlook technology staffing specialist the rethink group expects revenues to be marg  Tokens: ['rethink', 'group', 'positive', 'outlook', 'technology', 'staffing', 'specialist', 'rethink', 'group', 'expects', 'revenues', 'marg']  DocId: 34952018120409088 Original Doc: zombie fund manager phoenix appoints new ceo phoenix buys up funds that have been closed to new business and  Tokens: ['zombie', 'fund', 'manager', 'phoenix', 'appoints', 'ceo', 'phoenix', 'buys', 'funds', 'closed', 'business']  DocId: 34952008683229185 Original Doc: latest top world releases  Tokens: ['latest', 'top', 'world', 'releases'] |
| --- |

The first 100 Tokens from the list of Documents: ['world', 'cuts', 'bbc', 'service', 'savage', 'question', 'people', 'lot', 'fun', 'outlook', 'group', 'positive', 'marg', 'specialist', 'revenues', 'technology', 'expects', 'staffing', 'rethink', 'funds', 'business', 'buys', 'fund', 'appoints', 'phoenix', 'ceo', 'zombie', 'manager', 'closed', 'top', 'releases', 'latest', 'wonderland', 'presents', 'alice', 'dinner', 'posted', 'cdt', 'catonsville', 'territory', 'category', 'canada', 'calgary', 'jobs', 'alberta', 'job', 'location', 'bu', 'today', 'funding', 'school', 'transparency', 'lack', 'plans', 'news', 'free', 'depressing', 'deprived', 'plan', 'saving', 'manchester', 'hit', 'city', 'hardest', 'details', 'th', 'council', 'global', 'interested', 'professional', 'services', 'translation', 'dead', 'float', 'fitness', 'model', 'full', 'beautiful', 'smile', 'mostest', 'cook', 'david', 'lick', 'stand', 'piss', 'cnt', 'asses', 'meanies', 'blue', 'beware', 'thebluemeanies', 'como', 'alisson', 'warcraft', 'perde', 'dentes', 'os', 'exciting', 'gearing', 'hello']

## Evaluation for Queries:

For each query, the top 1000 results are returned in descending order. For this evaluation, only the top 10 matching documents are shown. Also included is the original document’s content.

**Baseline:**

*Query 3: “ Haiti Aristide return”*

| 3 Q0 29278582916251649 1 0.7657420886846849 myRun  Haiti - Aristide : His return, an international affair... - http://haitilibre.com/fben.php?id=2193 3 Q0 32273316047757312 2 0.7123073567365776 myRun  Haiti to give Aristide passport - http://www.bbc.co.uk/news/world-latin-america-12330414 3 Q0 32204788955357184 3 0.6990236920062056 myRun  Haiti opens door for return of ex-president Aristide http://tf.to/fJDt  3 Q0 32333726654398464 4 0.6874673881420821 myRun  #Aristide!! 3 Q0 29613127372898304 5 0.6656528399716845 myRun  If Duvalier Can Return to Haiti, Why Can’t Aristide? – New America Media http://bit.ly/eCWStk #haiti 3 Q0 31861291236724738 6 0.6639721067030084 myRun  Haitian Politics - Rev Jeremiah Wright Wants ARISTIDE To Return To Haiti: Yes, It will be good for Aristide to r... http://bit.ly/f2hFSi 3 Q0 29615296666931200 7 0.6471701842765625 myRun  If Duvalier Can Return to Haiti, Why Can't Aristide? - New America Media http://goo.gl/fb/USKRk 3 Q0 29296574815272960 8 0.6382109033929411 myRun  Haiti – Aristide : His return, an international affair… – Haitilibre.com http://bit.ly/gzyLXG #haiti 3 Q0 35088534306033665 9 0.6168762662364208 myRun  Haiti concede passaporte a Aristide. 3 Q0 34410414846517248 10 0.6168762662364208 myRun  ARISTIDE SERAIT DE RETOUR EN HAITI  *Query 10: “Egyptian protesters attack museum”*  10 Q0 31396671703220224 1 0.5513841114195462 myRun  museum http://yfrog.com/h25y0iij http://yfrog.com/hsu6aezj http://yfrog.com/h0arrqj http://yfrog.com/h8ijfrj http://yfrog.com/h067787496j 10 Q0 31518116253007874 2 0.5450264608126611 myRun  Rosicrucian Egyptian Museum: Rosicrucian Egyptian Museum http://bit.ly/fZhNnq 10 Q0 31499543946207232 3 0.5001254735475791 myRun  A collection of tweets on the the Egyptian Museum in Cairo: Is the Egyptian Museum Under Threat? http://bit.ly/eOgAHp #Jan25 #egypt #museum 10 Q0 31410583270068225 4 0.4720067607984883 myRun  Looters storm Egyptian Museum... http://drudge.tw/i7YrNH 10 Q0 31915245379256320 5 0.4635586057498040 myRun  attack attack! - smokahontas 10 Q0 31177737502724096 6 0.4635586057498040 myRun  Jammin attack attack! (: 10 Q0 29220224423174145 7 0.4635586057498040 myRun  who on twitter is going to attack attack? :D 10 Q0 31075323059638272 8 0.4572597989349993 myRun  Protesters forming teams to protect the Egyptian Museum from thieves. #Egypt #Jan25 #SidiBouzid 10 Q0 31245512300568577 9 0.4572597989349993 myRun  Confirmed: Egyptian protesters & activists successfully protected the national museum from looters #Egypt #Jan25 10 Q0 30637812046897152 10 0.4472991678228437 myRun  Egyptian protesters take to the streets for a third day http://www.thestar.com/news/world/article/928913--egyptian-protesters-take-to-the-streets-for-a-third-day?bn=1&sms\_ss=twitter&at\_xt=4d4185027c2217d3,0 … |
| --- |
|  |

**Query Expansion:**

*Query 3: “ Haiti Aristide return”*

| 3 Q0 32211683082502144 1 0.5647824434702122 myRun #int'l #news: Haiti opens door for return of ex-president Aristide: PORT-AU-PRINCE (Reuters) - Haiti'... http://bit.ly/gSIFwd #singapore 3 Q0 34962786534559745 2 0.4613799414419731 myRun Haiti Issues New Passport to Ex-Leader Aristide: PORT-AU-PRINCE - The Haitian government said i... http://bit.ly/dOCkkF commondreams.org 3 Q0 29278582916251649 3 0.45876980127082634 myRun Haiti - Aristide : His return, an international affair... - http://haitilibre.com/fben.php?id=2193 3 Q0 34283448172544000 4 0.44814575010977314 myRun The former pop singer who could be Haiti's president: BY TRENTON DANIEL PORT-AU-PRINCE -- Hours after Haiti's el... http://bit.ly/gW7rUX 3 Q0 32273316047757312 5 0.42569416276966804 myRun Haiti to give Aristide passport - http://www.bbc.co.uk/news/world-latin-america-12330414 3 Q0 29613127372898304 6 0.4202729654703586 myRun If Duvalier Can Return to Haiti, Why Can't Aristide? - New America Media http://bit.ly/eCWStk #haiti 3 Q0 29615296666931200 7 0.4187976147636564 myRun If Duvalier Can Return to Haiti, Why Can't Aristide? - New America Media http://goo.gl/fb/USKRk 3 Q0 29035242371158017 8 0.4118458371939881 myRun Breaking #news #tcot Haiti braces for new votes protests: PORT-AU-PRINCE: Haiti braced for more unrest, with p... http://twurl.nl/8evblo 3 Q0 32333726654398464 9 0.41183766029991553 myRun #Aristide!! 3 Q0 31861291236724738 10 0.39778993555331754 myRun Haitian Politics - Rev Jeremiah Wright Wants ARISTIDE To Return To Haiti: Yes, It will be good for Aristide to r... http://bit.ly/f2hFSi |
| --- |
| *Query 10: “ Egyptian protesters attack museum”* |

| 10 Q0 31414860923277312 1 0.35200821849939723 myRun Al Jazeera: Protestors create a human shield around the Egyptian Museum to protect it. #egypt #egipto #jan25 10 Q0 31499543946207232 2 0.33947856586902153 myRun A collection of tweets on the the Egyptian Museum in Cairo: Is the Egyptian Museum Under Threat? http://bit.ly/eOgAHp #Jan25 #egypt #museum 10 Q0 31075323059638272 3 0.3224144748616637 myRun Protesters forming teams to protect the Egyptian Museum from thieves. #Egypt #Jan25 #SidiBouzid 10 Q0 31245512300568577 4 0.3224144748616637 myRun Confirmed: Egyptian protesters & activists successfully protected the national museum from looters #Egypt #Jan25 10 Q0 31396671703220224 5 0.31848212476299875 myRun museum http://yfrog.com/h25y0iij http://yfrog.com/hsu6aezj http://yfrog.com/h0arrqj http://yfrog.com/h8ijfrj http://yfrog.com/h067787496j 10 Q0 31518116253007874 6 0.31491998253766257 myRun Rosicrucian Egyptian Museum: Rosicrucian Egyptian Museum http://bit.ly/fZhNnq 10 Q0 31961451212046337 7 0.311496126452153 myRun 5 QUESTIONS for Lisa Becker, organizer of CU Art Museum's 2011 Faculty Exhibition http://ow.ly/1b6D8L 10 Q0 31071191917662208 8 0.3114625052299619 myRun Egyptian Government Attacks Egypt's Internet #libertarian http://bit.ly/gvJ0u7 10 Q0 31204939556454400 9 0.3088321377980836 myRun Army, Protestors protect imperiled Cairo Museum http://bit.ly/hwFKyH 10 Q0 31076061462659072 10 0.29886680370330654 myRun al Jazeera: Thousands of Egyptian youth form human shields to protect the Egyptian Museum. #Museum #Egypt #Jan25 |
| --- |
|  |

**BERT:**

*Query 3: “ Haiti Aristide return”*

| 3 Q0 29278582916251649 1 0.6097804772855405 myRun Haiti - Aristide : His return, an international affair... - http://haitilibre.com/fben.php?id=2193 3 Q0 32204788955357184 2 0.5566508708925954 myRun Haiti opens door for return of ex-president Aristide http://tf.to/fJDt 3 Q0 32333726654398464 3 0.5475118207973322 myRun #Aristide!! 3 Q0 29296574815272960 4 0.5082347368068743 myRun Haiti - Aristide : His return, an international affair... - Haitilibre.com http://bit.ly/gzyLXG #haiti 3 Q0 29613127372898304 5 0.505109031062121 myRun If Duvalier Can Return to Haiti, Why Can't Aristide? - New America Media http://bit.ly/eCWStk #haiti 3 Q0 32273316047757312 6 0.49129173770197654 myRun Haiti to give Aristide passport - http://www.bbc.co.uk/news/world-latin-america-12330414 3 Q0 35088534306033665 7 0.49129173770197654 myRun Haiti concede passaporte a Aristide. 3 Q0 29615296666931200 8 0.4820737952317191 myRun If Duvalier Can Return to Haiti, Why Can't Aristide? - New America Media http://goo.gl/fb/USKRk 3 Q0 31861291236724738 9 0.4680737536316128 myRun Haitian Politics - Rev Jeremiah Wright Wants ARISTIDE To Return To Haiti: Yes, It will be good for Aristide to r... http://bit.ly/f2hFSi 3 Q0 32383831071793152 10 0.4641950787686086 myRun Yah Haiti: Haiti allows ex-president's return: Jean-Bertrand Aristide, who was Haiti's first democratically ele.... http://bit.ly/hLAgwO |
| --- |
|  |
| *Query 10: “Egyptian protesters attack museum”* |

| 10 Q0 31499543946207232 1 0.34240518569529316 myRun A collection of tweets on the the Egyptian Museum in Cairo: Is the Egyptian Museum Under Threat? http://bit.ly/eOgAHp #Jan25 #egypt #museum 10 Q0 31414860923277312 2 0.3259786306979002 myRun Al Jazeera: Protestors create a human shield around the Egyptian Museum to protect it. #egypt #egipto #jan25 10 Q0 31075323059638272 3 0.3253292572788611 myRun Protesters forming teams to protect the Egyptian Museum from thieves. #Egypt #Jan25 #SidiBouzid 10 Q0 31396671703220224 4 0.32056491807542914 myRun museum http://yfrog.com/h25y0iij http://yfrog.com/hsu6aezj http://yfrog.com/h0arrqj http://yfrog.com/h8ijfrj http://yfrog.com/h067787496j 10 Q0 31518116253007874 5 0.3169249932635923 myRun Rosicrucian Egyptian Museum: Rosicrucian Egyptian Museum http://bit.ly/fZhNnq 10 Q0 31245512300568577 6 0.3101892760024484 myRun Confirmed: Egyptian protesters & activists successfully protected the national museum from looters #Egypt #Jan25 10 Q0 31204939556454400 7 0.3098635612646741 myRun Army, Protestors protect imperiled Cairo Museum http://bit.ly/hwFKyH 10 Q0 30579268534669312 8 0.2982055894811259 myRun Egyptian protestors plan major demonstration http://bit.ly/dOxlwq 10 Q0 30361148733784064 9 0.2965512620538577 myRun Vintage clothing exhibition at Hill-Stead Museum: FARMINGTON, CT - Hill-Stead Museum is pleased to present a spe... http://bit.ly/esJec9 10 Q0 31406326609350657 10 0.2935968714742236 myRun NO NO NO !!!! THEY ENTERED THE MUSEUM IN EGYPT !!!!! |
| --- |
|  |

## Discussion

**Baseline:**

The baseline approach that simply used tf-idf and cosine similarity produced the worst performance with a MAP of 19.98% and a P@10 of 26.94%. Coincidentally, it produced the highest vectorized scores for queries 3 and 10. This makes sense as our weighting scheme likely contributed to this. We’re using lnc.ltc so we ignore the IDF for document vectors when calculating their length. Therefore, long document strings that are actually very relevant to the query are ranked lower than they should be. Since the baseline approach does not have its queries expanded, it has fewer words in the queries producing a higher value.

The baseline model works best for shorter queries (<4 words), that contain unique tokens.



*Figure 1: Results of the baseline approach from the Trec Eval Script*

**BERT:**

BERT performed marginally better than the baseline approach. This could be due to the fact that the BERT tokenization was only trained on the documents that were retrieved as results - which isn’t a large enough sample size. If, however, there was a pre-trained BERT model that was trained on a twitter corpus, it would have probably yielded better results. 

*Figure 2: Results of the BERT algorithm from the Trec Eval Script*

**Query Expansion:**

Query expansion performed the best among the 3 approaches, and showed a 10.75 % improvement compared to the baseline model, while MAP improved by 13.43%. As discussed in class, query expansion relies on semantics to mitigate the conventional problems found in basic IR systems. Through Gensim, the most similar word for each unique query token was appended to the original query. Gensim’s most\_similar method computes cosine similarity between a simple mean of the projection weight vectors of the given words and the vectors for each word in the model. While the internal weights are unknown, Gensim uses synonyms, meronyms, and relationships to assign different weights and determines the semantic similarity. As the overall length of the query was now larger, the lnc.ltc score was lower. However, the newly added tokens were useful in finding more relevant documents, and this was evident in the highest MAP score and P@10.

The query expansion model works best for longer queries (>6 words) that contain ambiguous tokens, that have many synonyms and are polysemous. This was the case for most of the queries so it makes sense why this approach performed the best.



*Figure 3: Results of the query expansion from the Trec Eval Script*