Simple Search Engine

Dataset:

- We downloaded the dataset from Kaggle.
- This dataset is derived from Wikipedia.
- The text file we are using consists of 7.8 million sentences in it.
- Each sentence is presented on a separate line and originates from the opening text of content pages.
- **Link:** [Wikipedia Sentences (kaggle.com)]

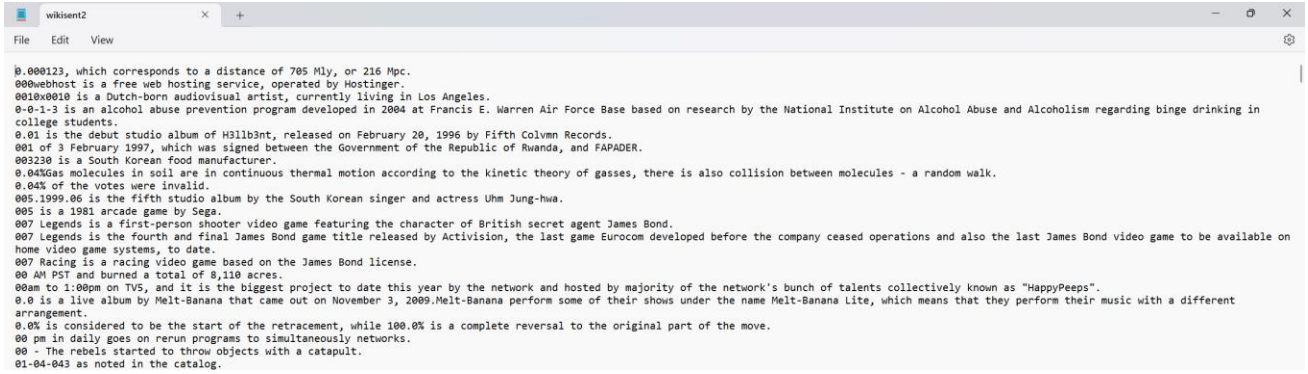
Rationale of this Project:

The goal of this project is to implement a simple search engine on Wikipedia Sentences dataset. This dataset consists of 7.8 million sentences. We start with creating an inverted index and then implement a ranking algorithm like TF-IDF. This project aims to demonstrate the fundamental principles behind building a simple search engine that can process and retrieve documents based on user queries.

Text Processing:

- This step involves cleaning and normalizing our dataset.
- Here we have cleaned the text file by removing punctuation and stops words, performed stemming and lowercasing.

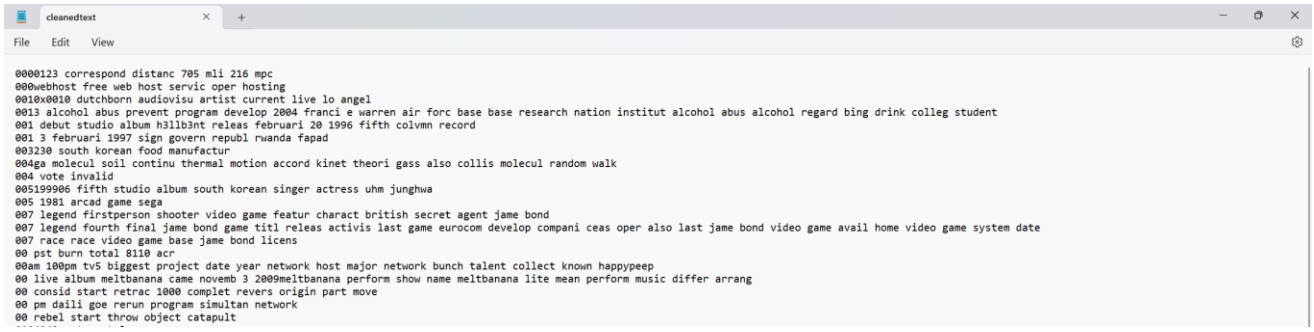
Snapshot of the dataset before Text Processing:


wikisent2

File Edit View

0.000123, which corresponds to a distance of 705 Mly, or 216 Mpc.
000webhost is a free web hosting service, operated by Hostinger.
0010x0010 is a Dutch-born audiovisual artist, currently living in Los Angeles.
0-0-1-3 is an alcohol abuse prevention program developed in 2004 at Francis E. Warren Air Force Base based on research by the National Institute on Alcohol Abuse and Alcoholism regarding binge drinking in college students.
0.01 is the debut studio album of H3llb3nt, released on February 20, 1996 by Fifth Colvmn Records.
001 of 3 February 1997, which was signed between the Government of the Republic of Rwanda, and FAPADER.
003230 is a South Korean food manufacturer.
0.04%Gas molecules in soil are in continuous thermal motion according to the kinetic theory of gasses, there is also collision between molecules - a random walk.
0.04% of the votes were invalid.
005.1999.06 is the fifth studio album by the South Korean singer and actress Uhm Jung-hwa.
005 is a 1981 arcade game by Sega.
007 Legends is a first-person shooter video game featuring the character of British secret agent James Bond.
007 Legends is the fourth and final James Bond game title released by Activision, the last game Eurocom developed before the company ceased operations and also the last James Bond video game to be available on home video game systems, to date.
007 Racing is a racing video game based on the James Bond license.
00 AM PST and burned a total of 8,110 acres.
00am to 1:00pm on TV5, and it is the biggest project to date this year by the network and hosted by majority of the network's bunch of talents collectively known as "HappyPeeps".
0.0 is a live album by Melt-Banana that came out on November 3, 2009.Melt-Banana perform some of their shows under the name Melt-Banana Lite, which means that they perform their music with a different arrangement.
0.0% is considered to be the start of the retracement, while 100.0% is a complete reversal to the original part of the move.
00 pm in daily goes on rerun programs to simultaneously networks.
00 - The rebels started to throw objects with a catapult.
01-04-043 as noted in the catalog.

Snapshot of the dataset after Text Processing:

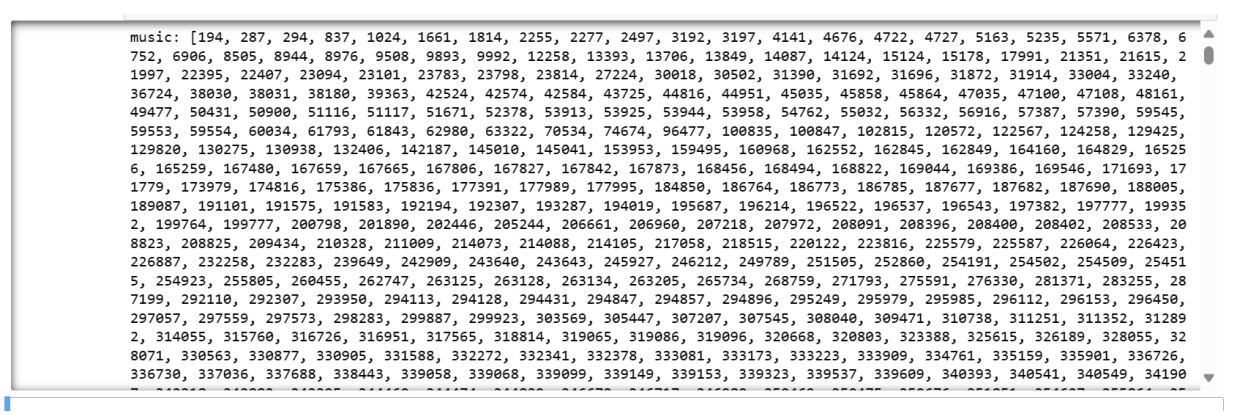


```
0000123 correspond distanc 705 mli 216 mpc
000webhost free web host servic oper hosting
0010x0010 dutchborn audiovisu artist current live lo angel
0013 alcohol abus prevent program develop 2004 franci e warren air forc base base research nation institut alcohol abus alcohol regard bing drink colleg student
001 debut studio album h3llb3nt releas februari 20 1996 fifth colvmn record
001 3 februari 1997 sign govern republ rwanda fapad
003230 south korean food manufactur
004ga molecul soil continu thermal motion accord kinet theori gass also collis molecul random walk
004 vote invalid
005199906 fifth studio album south korean singer actress uhm junghwa
005 1981 arcad game sega
007 legend firstperson shooter video game featur charact british secret agent jame bond
007 legend fourth final jame bond game titl releas activis last game eurocom develop compani ceas oper also last jame bond video game avail home video game system date
007 race race video game base jame bond licens
00 pst burn total 8110 acr
00am 100pm tv5 biggest project date year network host major network bunch talent collect known happypeep
00 live album meltbanana came novemb 3 2009meltbanana perform show name meltbanana lite mean perform music differ arrang
00 consid start retrac 1000 complet revers origin part move
00 pm daili goe rerun program simultan network
00 rebel start throw object catapult
```

- Here we have saved the cleaned text to new file.

Indexing:

- In this step we created an index that maps each word in the text file to list in the inverted index.
- Here we iterate over each word in the cleaned text file and keep track of its position.
- Then we add the position of each word to its corresponding list in the inverted index.
- If a word does not exist in the inverted index yet, then the function automatically creates a key with an empty list and appends the position.
- We have included an example usage of Indexing, here the code reads the cleaned text file, generates the inverted index, and prints the position of a sample word (music) if it exists in the inverted index.



```
music: [194, 287, 294, 837, 1024, 1661, 1814, 2255, 2277, 2497, 3192, 3197, 4141, 4676, 4722, 4727, 5163, 5235, 5571, 6378, 6
752, 6906, 8505, 8944, 8976, 9508, 9893, 9992, 12258, 13393, 13706, 13849, 14087, 14124, 15124, 15178, 17991, 21351, 21615, 2
1997, 22395, 22407, 23094, 23101, 23783, 23798, 23814, 27224, 30018, 30502, 31390, 31692, 31696, 31872, 31914, 33004, 33240,
36724, 38030, 38031, 38180, 39363, 42524, 42574, 42584, 43725, 44816, 44951, 45035, 45858, 45864, 47035, 47100, 47108, 48161,
49477, 50431, 50900, 51116, 51117, 51671, 52378, 53913, 53925, 53944, 53958, 54762, 55032, 56332, 56916, 57387, 57390, 59545,
59553, 59554, 60034, 61793, 61843, 62980, 63322, 70534, 74674, 96477, 100835, 100847, 102815, 120572, 122567, 124258, 129425,
129820, 130275, 130938, 132406, 142187, 145010, 145041, 153953, 159495, 160968, 162552, 162845, 162849, 164160, 164829, 16525
6, 165259, 167480, 167659, 167665, 167806, 167827, 167842, 167873, 168456, 168494, 168822, 169044, 169386, 169546, 171693, 17
1779, 173979, 174816, 175386, 175836, 177391, 177989, 177995, 184850, 186764, 186773, 186785, 187677, 187682, 187690, 188005,
189087, 191101, 191575, 191583, 192194, 192307, 193287, 194019, 195687, 196214, 196522, 196537, 196543, 197382, 197777, 19935
2, 199764, 199777, 200798, 201890, 202446, 205244, 206661, 206960, 207218, 207972, 208091, 208396, 208400, 208402, 208533, 20
8823, 208825, 209434, 210328, 211009, 214073, 214088, 214105, 217058, 218515, 220122, 223816, 225579, 225587, 226064, 226423,
226887, 232258, 232283, 239649, 242909, 243640, 243643, 245927, 246212, 249789, 251505, 252860, 254191, 254502, 254509, 25451
5, 254923, 255805, 260455, 262747, 263125, 263128, 263134, 263205, 265734, 268759, 271793, 275591, 276330, 281371, 283255, 28
7199, 292110, 292307, 293950, 294113, 294128, 294431, 294847, 294857, 294896, 295249, 295979, 295985, 296112, 296153, 296450,
297057, 297559, 297573, 298283, 299887, 299923, 303569, 305447, 307207, 307545, 308040, 309471, 310738, 311251, 311352, 31289
2, 314055, 315760, 316726, 316951, 317565, 318814, 319065, 319086, 319096, 320668, 320803, 323388, 325615, 326189, 328055, 32
8071, 330563, 330877, 330905, 331588, 332272, 332341, 332378, 333081, 333173, 333223, 333909, 334761, 335159, 335901, 336726,
336730, 337036, 337688, 338443, 339058, 339068, 339099, 339149, 339153, 339323, 339537, 339609, 340393, 340541, 340549, 34190
```

Search Algorithm:

- Here we have implemented a search algorithm TF-IDF (Term Frequency – Inverse Document Frequency) to match the input user query with the document in the index.
- Here we have processed the cleaned text document and calculated TF-IDF scores.
- Then a Data Frame is created for easy lookup of important terms and queries based on the relevance to the TF-IDF to the document.

	Word	TF-IDF Score	Frequency
0	00	1.147590e-04	264
1	000	1.934384e-04	445
2	0000	1.999588e-05	46
3	00000	8.693862e-07	2
4	000000	3.477545e-06	8
...	...	...	...
1631945	zzz3	4.346931e-07	1
1631946	zzzap	1.304079e-06	3
1631947	zzzax	8.693862e-07	2
1631948	zzzz	1.738772e-06	4
1631949	zzzzz	4.346931e-07	1

[1631950 rows x 3 columns]

- The above snapshot is the output of TF-IDF scores of each word in the document, combines them with word frequencies.
- The output is presented in tabular format for easier analysis and visualization.

Ranking:

- Here we prioritize words in the search results based on the TF-IDF score, to make sure that most relevant results appear on top.
- We have calculated the TF-IDF scores for words in the text file and stored the scores in a data frame for easy lookup.

- The data frame is sorted by the TF-IDF scores in descending order, so the highest scores are the top.
- We have also calculated the frequency of each word in the text file and returns the most occurred words in the text file.
  Snapshot of top 5 words with the highest TF-IDF scores.

```
Top 5 words with the highest TF-IDF scores:
            TF-IDF
born      0.212212
state     0.186211
new       0.164416
unit      0.164037
american  0.163867
```

Snapshot of top 5 frequently words occurred in the document:

```
Top 5 most frequently occurring words:
born: 385332
state: 338120
also: 312653
new: 298545
unit: 297856
```

Relevant sections to the syllabus:

Stemming:

- Stemming is a text normalization technique to reduce words to their base root form by removing affixes such as prefixes, suffixes, and pluralization.
- In this project stemming is used to map different variations of words to the same root, which helps in information retrieval.

Indexing:

- It is the process of organizing and structuring a collection of documents or words in the documents to perform a fast retrieval.
- In this project we have created an inverted index that maps words to the sentences where they occurred. This mapping allows us to efficiently retrieval specific words in the dataset.

TF-IDF:

- TF-IDF (Term Frequency and Inverse Document Frequency), is used to measure the importance of words in a document.
- TF is used to measure how frequently a word appears in the document.
- IDF is used to measure how important a word is across the document.

Page Rank:

- Page ranking is aspect of search engine that evaluates the importance of web pages based on the factors like page rank scores of the webpages.
- In this project we performed ranking by measuring the TF-IDF score and number of occurrences of a word in the document.

Conclusion:

This project acts as the hands-on experience in understanding and implementing search engine functionalities. Through this project we have gained practical experience in text processing like cleaning, stemming, and creating an inverted index for efficient document retrieval. We also learned about the TF-IDF algorithm and its role in ranking documents based on term importance and frequency.