CSE508 Information Retrieval

Winter 2024

Assignment-1

Konam Akhil Vamshi

2020513

Question 1:

Approach and methodology:

1. Applied the below-mentioned pre-processing steps on the given dataset to narrow down the size of indexing in the coming parts.
    - Lowercasing text
    - Tokenization
    - Removing stopwords
    - Removing punctuations
    - Removing blank spaces

2. Preferred spaCy over NLTK for pre-processing as spaCy is more robust and contains a broader set of stopwords. This helps to further reduce the vocabulary size, reducing the size of indexes. This may help in faster retrieval of queries.

Outputs:
--- File: file1.txt ---
Original Text:
Loving these vintage springs on my vintage strat. They have a good tension and great stability. If you are floating your bridge and want the most out of your springs than these are the way to go.

After Lowercasing:
loving these vintage springs on my vintage strat. they have a good tension and great stability. if you are floating your bridge and want the most out of your springs than these are the way to go.

After Tokenization:
loving these vintage springs on my vintage strat . they have a good tension and great stability . if you are floating your bridge and want the most out of your springs than these are the way to go .

After Removing Stopwords:
loving vintage springs vintage strat . good tension great stability . floating bridge want springs way .

After Removing Punctuations:
loving vintage springs vintage strat good tension great stability floating bridge want springs way

After Removing Blank Spaces:
loving vintage springs vintage strat good tension great stability floating bridge want springs way

Final Processed Sentence:

loving vintage springs vintage strat good tension great stability floating bridge want springs way

--- File: file2.txt ---
Original Text:
Works great as a guitar bench mat. Not rugged enough for abuse but if you take care of it, it will take care of you. Makes organization of workspace much easier because screws won't roll around. Color is good too.

After Lowercasing:
works great as a guitar bench mat. not rugged enough for abuse but if you take care of it, it will take care of you. makes organization of workspace much easier because screws won't roll around. color is good too.

After Tokenization:
works great as a guitar bench mat . not rugged enough for abuse but if you take care of it , it will take care of you . makes organization of workspace much easier because screws wo n't roll around . color is good too .

After Removing Stopwords:
works great guitar bench mat . rugged abuse care , care . makes organization workspace easier screws wo roll . color good .

After Removing Punctuations:
works great guitar bench mat rugged abuse care care makes organization workspace easier screws wo roll color good

After Removing Blank Spaces:
works great guitar bench mat rugged abuse care care makes organization workspace easier screws wo roll color good

Final Processed Sentence:
works great guitar bench mat rugged abuse care care makes organization workspace easier screws wo roll color good

--- File: file4.txt ---
Original Text:
Great price and good quality. It didn't quite match the radius of my sound hole but it was close enough.

After Lowercasing:
great price and good quality. it didn't quite match the radius of my sound hole but it was close enough.

After Tokenization:

great price and good quality . it did n't quite match the radius of my sound hole but it was close enough .

After Removing Stopwords:
great price good quality . match radius sound hole close .

After Removing Punctuations:
great price good quality match radius sound hole close

After Removing Blank Spaces:
great price good quality match radius sound hole close

Final Processed Sentence:
great price good quality match radius sound hole close

--- File: file5.txt ---
Original Text:
I bought this bass to split time as my primary bass with my Dean Edge. This might be winning me over. The bass boost is outstanding. The active pickups really allow you to adjust to the sound you want. I recommend this for anyone. If you're a beginner like I was not too long ago, it's an excellent bass to start with. If you're on tour and/or music is making you money, this bass will be beatiful on stage. The color is a bit darker than in the picture. But, all around, this is a great buy.

After Lowercasing:
i bought this bass to split time as my primary bass with my dean edge. this might be winning me over. the bass boost is outstanding. the active pickups really allow you to adjust to the sound you want. i recommend this for anyone. if you're a beginner like i was not too long ago, it's an excellent bass to start with. if you're on tour and/or music is making you money, this bass will be beatiful on stage. the color is a bit darker than in the picture. but, all around, this is a great buy.

After Tokenization:
i bought this bass to split time as my primary bass with my dean edge . this might be winning me over . the bass boost is outstanding . the active pickups really allow you to adjust to the sound you want . i recommend this for anyone . if you 're a beginner like i was not too long ago , it 's an excellent bass to start with . if you 're on tour and/or music is making you money , this bass will be beatiful on stage . the color is a bit darker than in the picture . but , all around , this is a great buy .

After Removing Stopwords:
bought bass split time primary bass dean edge . winning . bass boost outstanding . active pickups allow adjust sound want . recommend . beginner like long ago , excellent bass start . tour and/or music making money , bass beatiful stage . color bit darker picture . , , great buy .

After Removing Punctuations:
bought bass split time primary bass dean edge winning bass boost outstanding active pickups allow adjust sound want recommend beginner like long ago excellent bass start tour and/or music making money bass beatiful stage color bit darker picture great buy

After Removing Blank Spaces:
bought bass split time primary bass dean edge winning bass boost outstanding active pickups allow adjust sound want recommend beginner like long ago excellent bass start tour and/or music making money bass beatiful stage color bit darker picture great buy

Final Processed Sentence:
 bought bass split time primary bass dean edge winning bass boost outstanding active pickups allow adjust sound want recommend beginner like long ago excellent bass start tour and/or music making money bass beatiful stage color bit darker picture great buy

--- File: file7.txt ---
Original Text:
Absolute BEST guitar hangers on the market... You will not beat this price! Buy them while you still can for this cheap

After Lowercasing:
absolute best guitar hangers on the market... you will not beat this price! buy them while you still can for this cheap

After Tokenization:
absolute best guitar hangers on the market ... you will not beat this price ! buy them while you still can for this cheap

After Removing Stopwords:
absolute best guitar hangers market ... beat price ! buy cheap

After Removing Punctuations:
absolute best guitar hangers market beat price buy cheap

After Removing Blank Spaces:
absolute best guitar hangers market beat price buy cheap

Final Processed Sentence:
absolute best guitar hangers market beat price buy cheap

Question 2:

Methodology

- **Preprocessing with spaCy:** Lowercasing text, tokenisation, removal of stopwords, punctuation and spaces using spaCy.
- **Inverted Index Creation:** Python function inv_index that iterates over preprocessed text files in a directory, creating a mapping of unique words to the files they appear in.
- **Index Serialization:** Using pickle to serialise the inverted index to save it and use it later.
- **Boolean Query Processing:** process_query function to support AND, OR, AND NOT, OR NOT operations by manipulating sets of documents based on the presence or absence of query terms in the inverted index.
- **Query and Operation Input:** Taking query and operations input from the user.
- **Preprocessing User Queries:** Preprocessing the above input using spaCy.
- **Result Display:** Displaying the results of retrieved documents as per instructions.

Sample input and output:

Input:
1
Car bag in a canister
OR, AND NOT

Output:
Query 1: car OR bag AND NOT canister
Number of documents retrieved for query 1: 31
Names of the documents retrieved for query 1: file863.txt, file363.txt, file174.txt, file930.txt, file746.txt, file981.txt, file699.txt, file956.txt, file892.txt, file118.txt, file686.txt, file698.txt, file73.txt, file860.txt, file797.txt, file864.txt, file264.txt, file780.txt, file886.txt, file166.txt, file466.txt, file313.txt, file459.txt, file942.txt, file404.txt, file665.txt, file682.txt, file542.txt, file573.txt, file3.txt, file738.txt

Question 3:

Methodology:

- **Positional Index Creation:** Created a positional index mapping each unique word to its positions within documents.
- **Serialisation with Pickle:** Using pickle to serialise the inverted index to save it and use it later.
- **Handling Phrase Queries:** ‘process’ function to process input phrase queries by checking for the exact sequence of words in the positional index.
- **Query and Operation Input:** Taking query and operations input from the user.
- **Preprocessing User Queries:** Preprocessing the above input using spaCy.
- **Result Display:** Displaying the results of retrieved documents as per instructions.

Sample input and output:

Input:
1
fun play

Output:
Number of documents retrieved for query 1 using positional index: 3
Names of documents retrieved for query 1 using positional index: file279.txt, file6.txt, file854.txt