Basic Search Engine

Phase 1: Specification:

1. Create a search engine that accepts query input and displays URL results.
2. Read the Sample data file and perform tokenization.
3. Clean the tokens by removing punctuation and checking if the char is alpha and return the clean token or empty if it is not alpha.
4. Create the forward index after reading the data file and return dictionary with a key-value pair, where the key is the URL and the value is a set of tokenized words.
5. Create the inverted index by reversing the key value pairs received from the forward index, where the key is a collection of tokenized words and the values are URLs.
6. Search the inverted index keys for the entered query, and then show the values of the corresponding index keys' URLs.

Phase 2: Design:

Modules and basic structures:

1. **mySearchEngine(dbfile):** This method takes dbfile as input and performs necessary operations and asks for the user input for query to search and displays the matched URLs as output.

   readdocs: This is a data member that holds the forward index after reading the data file and in a dictionary with a key-value pair, where the key is the URL and the value is a set of tokenized words.

   inverted_index: This data member stores the inverted index after reversing the key value pairs received from the forward index, where the key is a collection of tokenized words and the values are URLs.

   len_url: This variable holds the number of URLs indexed by the search engine.

   readDocs(dbfile): This is a method responsible for reading documents from the database file and returning them in a dictionary with a key-value pair, where the key is the URL and the value is a set of tokenized words.

   buildInvertedIndex(readdocs): This is a method that builds and returns the inverted index by reversing the key value pairs received from the forward index, where the key is a collection of tokenized words and the values are URLs.

findQueryMatches(inverted_index, query): This method is responsible for finding query matches in the inverted index based on a given search query. It should return a list of URLs that match the query.

2. **readDocs(dbfile):** This is a method responsible for reading the database file and returning forward index in a dictionary with a key-value pair, where the key is the URL and the value is a set of tokenized words.

   readdocs: This is a dictionary which stores the result of URL and tokenized words in a key-value pair.

   url: Local variable to store the URL.

   content: Local variable to store the content of the dbfile.

   cleaned_tokens: This is a set() which holds the unique tokens after cleaning the punctuations in tokens.

   pagebodyflag: Local variable to perform reading operations.

3. **cleanToken(token)**: This method returns the clean tokens after removing the string punctuations from the beginning and end of a token and also checks if the token is alpha and return the clean token or empty if it is not alpha.

4. **buildInvertedIndex(docs):** This method builds the inverted index by reversing the key value pairs received from the forward index, where the key is a collection of tokenized words and the values are URLs.

5. **findQueryMatches(index, query):** This method searches the inverted index keys for the entered query, and then return the values of the corresponding index keys' URLs in an **output_url** set.

Phase 3: Pseudocode:

1. Function cleanToken(token):

   Remove Punctuations from beginning and end of token

   For each character in token:

   If character is not in set(punctuation characters):

Add character to cleaned_token

is_alpha = False

For each character in cleaned_token:

If character is an alphabetic character:

Set is_alpha to True

Break

If is_alpha is True and length of alphabetic character more than 1:

Return cleaned_token converted to lowercase

Else:

Return an empty string

2. Function readDocs(dbfile):

readdocs = {} # Initialize an empty dictionary

url = None

content = ""

cleaned_tokens = set()

pagebodyflag = False

Open dbfile for reading

For each line in dbfile:

Strip leading and trailing whitespace from the line

If line starts with "<pageBody>":

Set pagebodyflag to True

Else if line starts with "<endPageBody>":

Set pagebodyflag to False

Set url to None

Else if line starts with "http" and pagebodyflag is False:

If url is not None:

Add cleaned_tokens to readdocs dictionary with url as the key

Set cleaned_tokens to an empty set

Set url to the current line

Set content to an empty string

Else:

If url is not None:

```
            Append line to content
            Tokenize content into tokens
            For each token in tokens:
                Clean the token using cleanToken function
                Add the cleaned token to cleaned_tokens set
    Close dbfile
    Return readdocs dictionary

3. Function buildInvertedIndex(docs):
    inverted_index = {}  # Initialize an empty dictionary
    For each url, words in docs.items():
        For each word in words:
            If word is not in inverted_index:
                Add word to inverted_index with an empty set as the value
            Add url to inverted_index[word]
    Return inverted_index dictionary

4. Function findQueryMatches(index, query):
    querysplit = Split query into a list of terms
    output_url = an empty set
    For each value in querysplit:
        operator = ""
        If value starts with '+':
            Set operator to '+'
            Remove the '+' from value
        Else if value starts with '-':
            Set operator to '-'
            Remove the '-' from value
        clean_value = Clean value using cleanToken function
        If clean_value is not empty and clean_value is in index:
            result = Get the set of URLs associated with clean_value from index
            If operator is '+':
                Intersect output_url with result
```

Else if operator is '-':

Subtract result from output_url

Else:

Union output_url with result

Return output_url

5. Function mySearchEngine(dbfile):

readdocs = Call readDocs(dbfile)

inverted_index = Call buildInvertedIndex(readdocs)

While True:

Query = Input "Enter a search query (or empty string to quit): "

If Query is empty:

Break

matches = Call findQueryMatches(inverted_index, Query)

Print "Found", len(matches), "Matching Pages"

If matches is not empty:

Print matches

Phase 4: Testing and Output:

Stand while building index...

Indexed 5 pages containing 1869 unique terms.

Enter a search query (or empty string to quit):

1. **Search the query**: Kalamazoo

   Output:

   Enter a search query (or empty string to quit): Kalamazoo

   Found 1 Matching Pages

   {'https://wmich.edu/you.html'}

```
PS W:\WMU Assignments\Program for Grad\Assignment 2> & C:/Users/rohan/AppData/Local/Microsoft/WindowsApps/python3.11.exe "w:
/WMU Assignments/Program for Grad/Assignment 2/search.py"
Stand while building index...
Indexed 5 pages containing  1869  unique terms.
Enter a search query (or empty string to quit): Kalamazoo
Found  1  Matching Pages
{'https://wmich.edu/you.html'}
Enter a search query (or empty string to quit):
```

2. **Search the query**: syllabus

   Search the query: ajay

   Search the query: syllabus +ajay

   Search the query: syllabus -ajay

   Output:

   Enter a search query (or empty string to quit): syllabus

   Found 2 Matching Pages

   {'https://www.cs.wmich.edu/gupta/teaching/cs5950/5950F23PGSweb/TopicsCovered%20ProgGradStu.html', 'https://cs.wmich.edu/elise/courses/cs531/assignments-SI19.html'}

   Enter a search query (or empty string to quit): ajay

   Found 2 Matching Pages

   {'https://www.cs.wmich.edu/gupta/teaching/cs5950/5950F23PGSweb/TopicsCovered%20ProgGradStu.html',
   'https://www.cs.wmich.edu/~gupta/teaching/cs603/wsnSp04/ClassPolicies.html'}

   Enter a search query (or empty string to quit): syllabus +ajay

   Found 1 Matching Pages

   {'https://www.cs.wmich.edu/gupta/teaching/cs5950/5950F23PGSweb/TopicsCovered%20ProgGradStu.html'}

   Enter a search query (or empty string to quit): syllabus -ajay

   Found 1 Matching Pages

   {'https://cs.wmich.edu/elise/courses/cs531/assignments-SI19.html'}

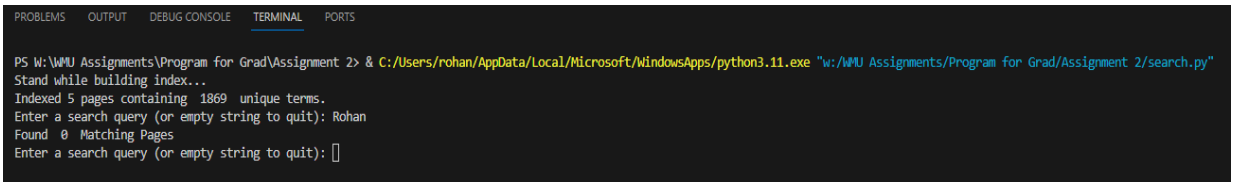
```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS

PS W:\WMU Assignments\Program for Grad\Assignment 2> & C:/Users/rohan/AppData/Local/Microsoft/WindowsApps/python3.11.exe "w:/WMU Assignments/Program for Grad/Assignment 2/search.py"
Stand while building index...
Indexed 5 pages containing 1869 unique terms.
Enter a search query (or empty string to quit): syllabus
Found 2 Matching Pages
Enter a search query (or empty string to quit): ajay
Found 2 Matching Pages
{'https://www.cs.wmich.edu/gupta/teaching/cs5950/5950F23PGSweb/TopicsCovered%20ProgGradStu.html', 'https://www.cs.wmich.edu/~gupta/teaching/cs603/wsnSp04/ClassPolicies.html'}
Enter a search query (or empty string to quit): syllabus +ajay
Found 1 Matching Pages
{'https://www.cs.wmich.edu/gupta/teaching/cs5950/5950F23PGSweb/TopicsCovered%20ProgGradStu.html'}
Enter a search query (or empty string to quit): syllabus -ajay
Found 1 Matching Pages
{'https://cs.wmich.edu/elise/courses/cs531/assignments-SI19.html'}
Enter a search query (or empty string to quit):
```

3. **Search the query**: Rohan

   Output:

   Enter a search query (or empty string to quit): Rohan

   Found 0 Matching Pages



4. **Search the query**: temptations

   Search the query: binary

   Search the query: temptations +binary

   Search the query: temptations +binary -wireless

   Search the query: temptations +binary +wireless

   Search the query: temptations +wireless

   Search the query: temptations +binary wireless

   Search the query: temptations +binary wireless -architecture

   Output:

   Enter a search query (or empty string to quit): temptations

   Found 1 Matching Pages

   {'https://www.cs.wmich.edu/~gupta/teaching/cs603/wsnSp04/ClassPolicies.html'}

   Enter a search query (or empty string to quit): binary

   Found 1 Matching Pages

   {'https://www.cs.wmich.edu/gupta/teaching/cs5950/5950F23PGSweb/TopicsCovered%20ProgGradStu.html'}

   Enter a search query (or empty string to quit): temptations +binary

   Found 0 Matching Pages

   Enter a search query (or empty string to quit): temptations +binary -wireless

   Found 0 Matching Pages

   Enter a search query (or empty string to quit): temptations +binary +wireless

   Found 0 Matching Pages

   Enter a search query (or empty string to quit): temptations +wireless

   Found 1 Matching Pages

   {'https://www.cs.wmich.edu/~gupta/teaching/cs603/wsnSp04/ClassPolicies.html'}

   Enter a search query (or empty string to quit): temptations +binary wireless

   Found 2 Matching Pages

{'https://cs.wmich.edu/~alfuqaha/Spring06/cs5550/projects.html',

'https://www.cs.wmich.edu/~gupta/teaching/cs603/wsnSp04/ClassPolicies.html'}

Enter a search query (or empty string to quit): temptations +binary wireless -architecture

Found 1 Matching Pages

{'https://www.cs.wmich.edu/~gupta/teaching/cs603/wsnSp04/ClassPolicies.html'}

Enter a search query (or empty string to quit):

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS

PS W:\WMU Assignments\Program for Grad\Assignment 2> & C:/Users/rohan/AppData/Local/Microsoft/WindowsApps/python3.11.exe "w:/WMU Assignments/Program for Grad/Assignment 2/search.py"
Stand while building index...
Indexed 5 pages containing  1869  unique terms.
Enter a search query (or empty string to quit): temptations
Found  1  Matching Pages
Enter a search query (or empty string to quit): binary
Found  1  Matching Pages
{'https://www.cs.wmich.edu/gupta/teaching/cs5950/5950F23PGSweb/TopicsCovered%20ProgGradStu.html'}
Enter a search query (or empty string to quit): temptations +binary
Found  0  Matching Pages
Enter a search query (or empty string to quit): temptations +binary -wireless
Found  0  Matching Pages
Enter a search query (or empty string to quit): temptations +binary +wireless
Found  0  Matching Pages
Enter a search query (or empty string to quit): temptations +wireless
Found  1  Matching Pages
{'https://www.cs.wmich.edu/~gupta/teaching/cs603/wsnSp04/ClassPolicies.html'}
Enter a search query (or empty string to quit): temptations +binary wireless
Found  2  Matching Pages
{'https://www.cs.wmich.edu/~gupta/teaching/cs603/wsnSp04/ClassPolicies.html', 'https://cs.wmich.edu/~alfuqaha/Spring06/cs5550/projects.html'}
Enter a search query (or empty string to quit): temptations +binary wireless -architecture
Found  1  Matching Pages
{'https://www.cs.wmich.edu/~gupta/teaching/cs603/wsnSp04/ClassPolicies.html'}
```

5. Search the query:

   Output:

   Program will exit

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS

PS W:\WMU Assignments\Program for Grad\Assignment 2> & C:/Users/rohan/AppData/Local/Microsoft/WindowsApps/python3.11.exe "w:/WMU Assignments/Program for Grad/Assignment 2/search.py"
Stand while building index...
Indexed 5 pages containing  1869  unique terms.
Enter a search query (or empty string to quit):
PS W:\WMU Assignments\Program for Grad\Assignment 2>
```

Notes

Reference

https://www.geeksforgeeks.org/inverted-index/