**CSCE 5200 – Information Retrieval and Web Search**

**Assignment – 01**

1. Search tool for Amazon reviews.
2. Given, we have 100 text files: each one containing a review about a product.
3. I have used Google Colab, An interactive python interpreter.
4. I have segmented the assignment into three parts:

* Loading and reading the files.
* Removing the given stop words and creating an Inverted index
* Prompt to ask the user to enter the search word and retrieve the Document IDs in which the search word is present.

1. **Loading and reading the files** –

Graphical user interface, application

Description automatically generated

* Created a folder in the sample data folder of google colab and uploaded all the files.
* Also uploaded the stop words file to the colab.

Graphical user interface, text, application, email

Description automatically generated

* Imported the required libraries.

Text, letter

Description automatically generated

* Created a for loop to iterate through the text in the files and convert the corpus of text into lowercase.
* Read all the text files and created a corpus out of all the text.

1. **Removing the given stop words and creating an Inverted index**Graphical user interface, text, application, email

   Description automatically generated

* Tokenized the corpus and removed the punctuation symbols.
* Then dropped the stop words.
* Then created an inverted index, which contains the words and positing’s list.

Text, letter

Description automatically generated

* Sample of the inverted index.

1. **Prompt to ask the user to enter the search word and retrieve the Document IDs in which the search word is present.**

Graphical user interface, text, application

Description automatically generated

* Prompts the user to insert a search word, then it tries to find a match in the inverted index, if it finds a match it will return the Document ID’s.

Graphical user interface, text, application, email

Description automatically generated

* If the search word does not exist in the inverted index, it displays the search word does not exist.

Graphical user interface, text, application, email

Description automatically generated



* Example for searching a stop word.



Graphical user interface, text, application, email

Description automatically generated

**References**:

1. <https://github.com/wesdoyle/Javelin/blob/master/Notebooks/Inverted%20Index.ipynb>
2. <https://www.geeksforgeeks.org/inverted-index/>