

**EEC 525 / CIS 660 DATA MINING**

**LAB 2**

**Part 1: Document Vectorization to Measure Similarity of Web Pages for the Given Topics of a User Query,**

**Part 2: Building Similarity Matrix for Webpage/Document Categorization for the Given Topics for Text Analysis of Documents (Webpages)  
&   
Part3: Analysis and Discussion of Problems**

Submitted By

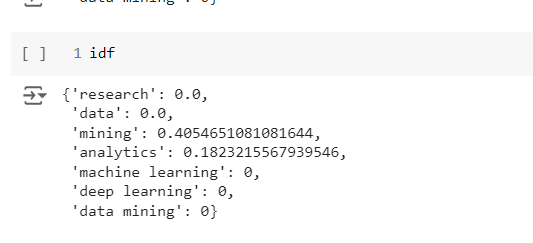
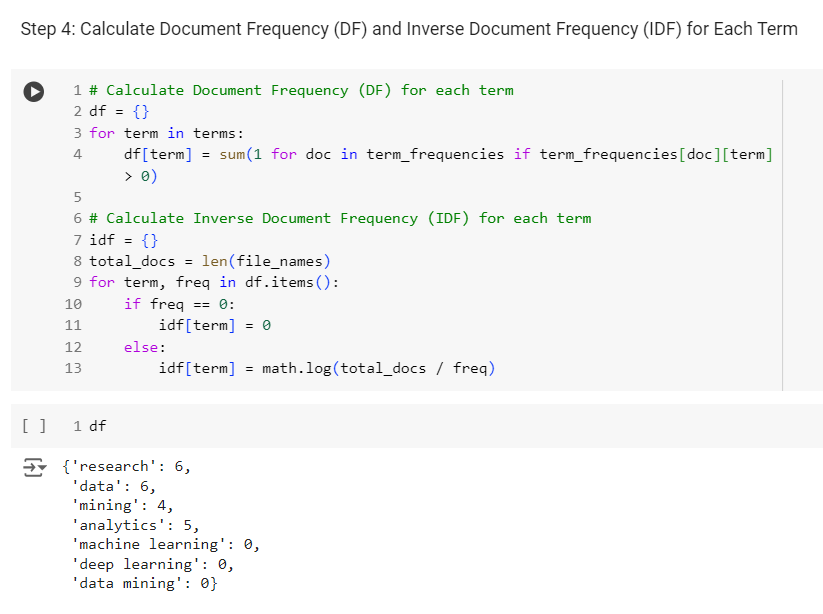
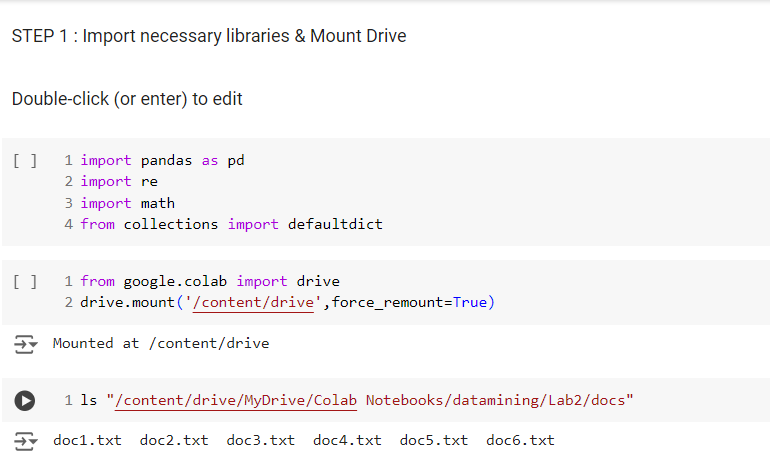
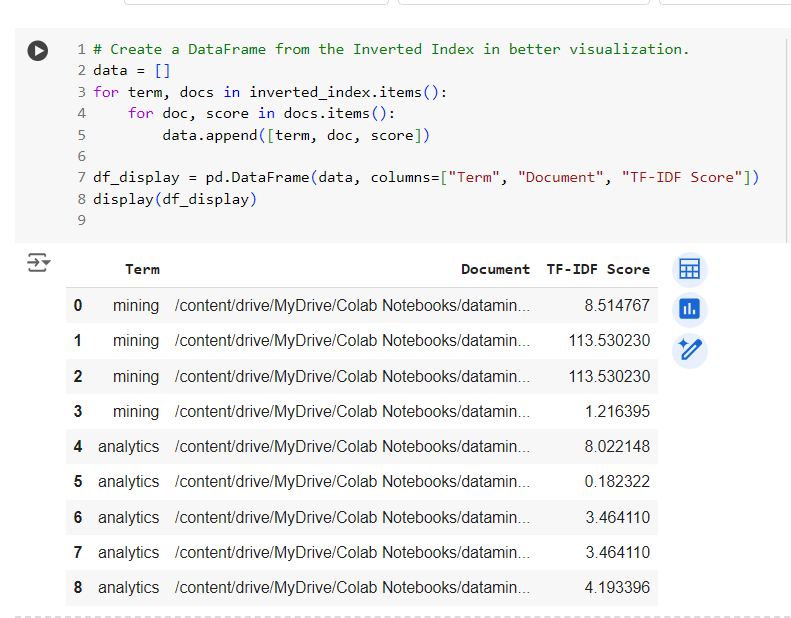
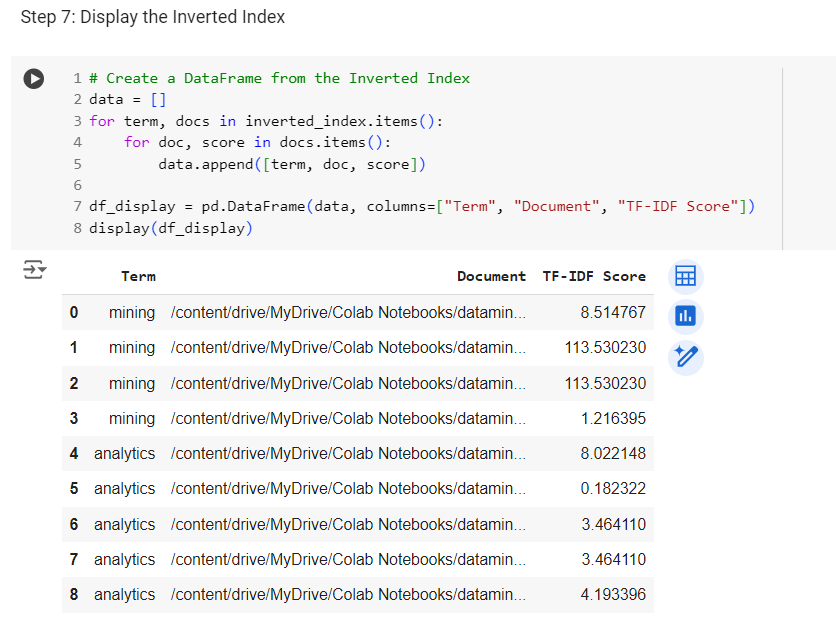
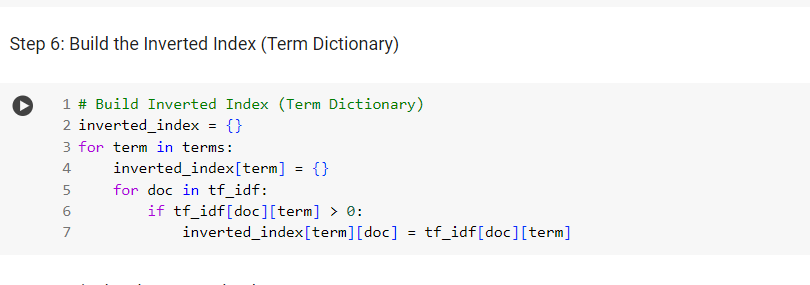
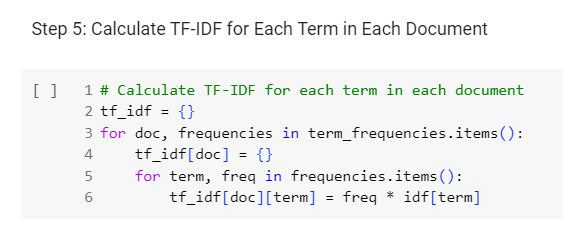
**Dinky Mishra**

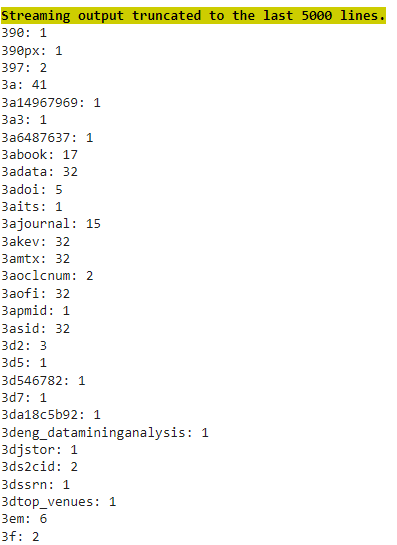
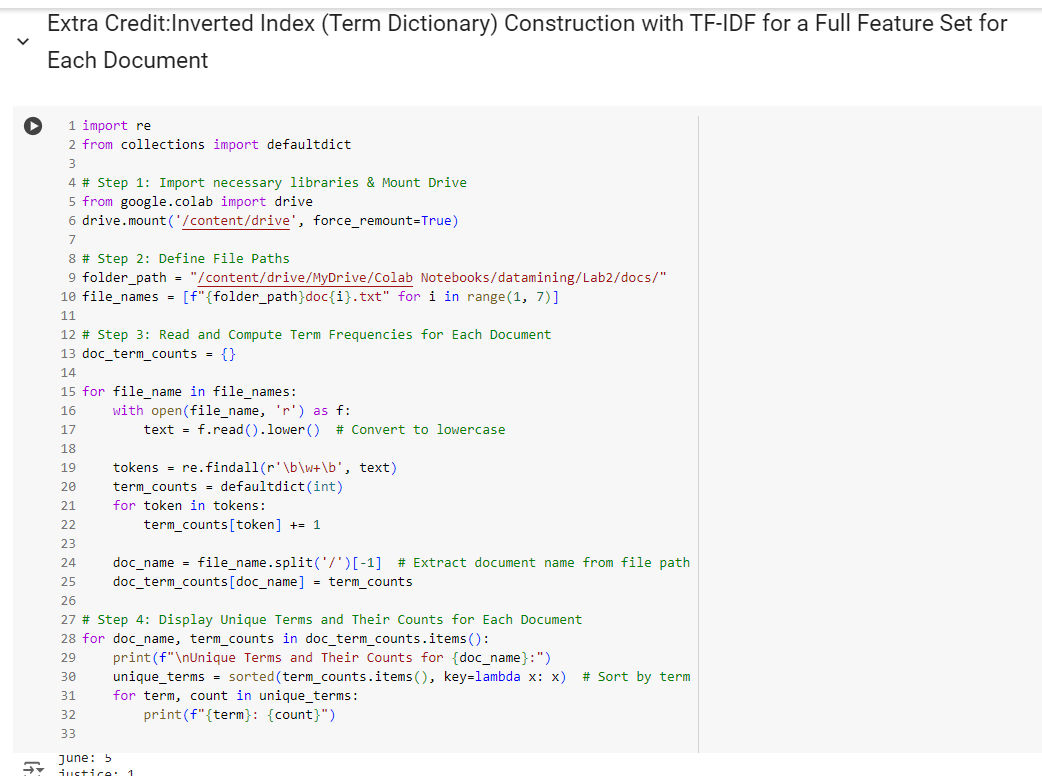
**CSU ID: 2864923**

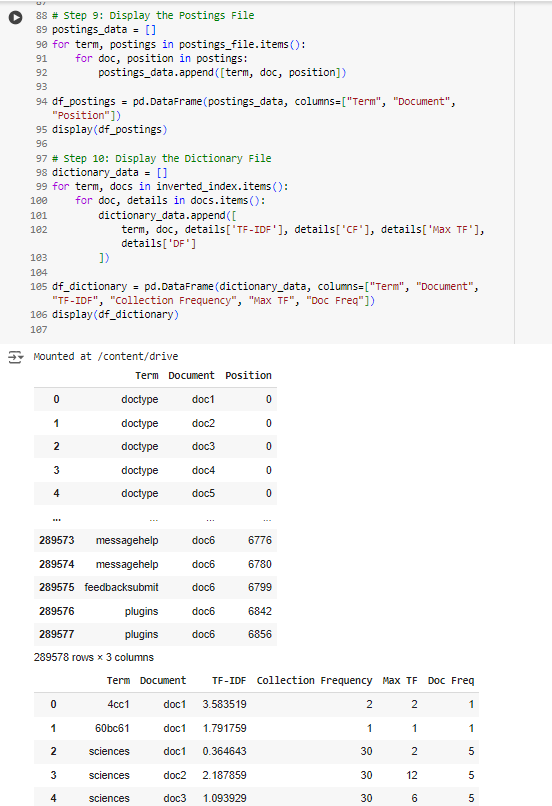
**Table of contents**

|  |  |  |
| --- | --- | --- |
| **S.NO** | **Content** | **Page No** |
| 1 | **Part 1: Document Vectorization to Measure Similarity of Web Pages for the Given Topics of a User Query** | 3-7 |
| 2 | **EXTRA CREDIT: Inverted Index (Term Dictionary) Construction with TF-IDF for a Full Feature Set for Each Document** | 8-9 |
| 3 | **Part 2: Building Similarity Matrix for Webpage/Document Categorization for the Given Topics for Text Analysis of Documents (Webpages)** | 10-13 |
| 4 | **Part3: Analysis and Discussion of Problems** | 14-15 |

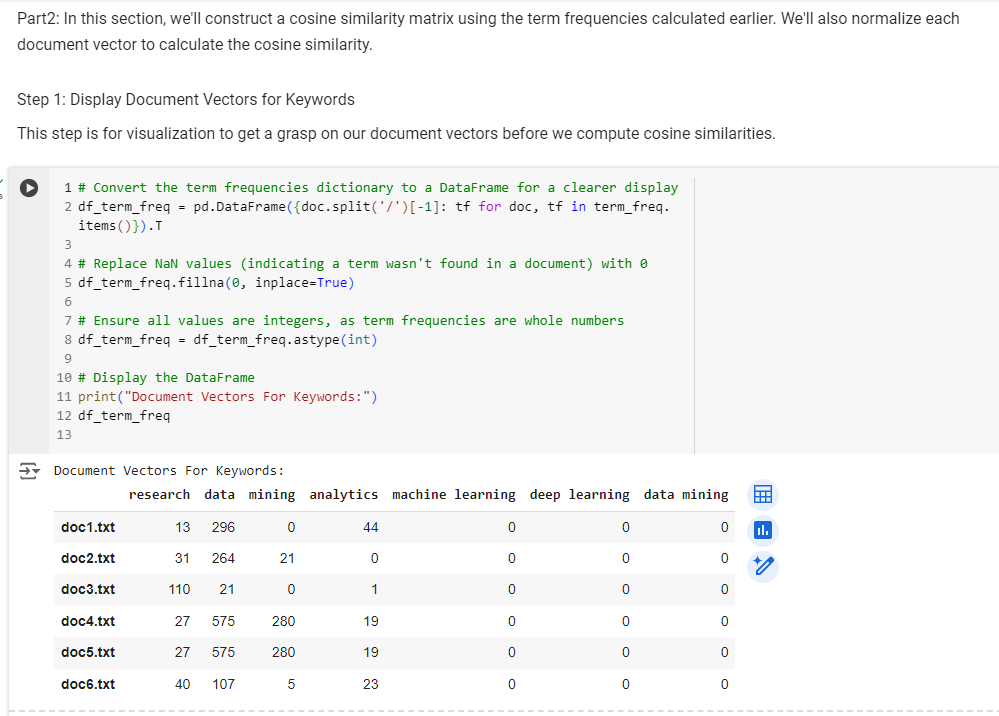
**Overview:**  
For this assignment, you will construct document vectors for six webpages based on the frequency of seven specified topics: research, data, mining, analytics, data mining, machine learning, and deep learning. You will save the HTML content of the webpages as .txt files, process these files to count the frequency of each topic word and bi-gram phrase, and then create vectors representing these frequencies. This analysis helps in understanding the content distribution across the documents, which is useful for tasks like search engine optimization and content categorization.  
Tools used :Google Colab, Python Libraries: Pandas,re,math  
  
This Lab2 has two Parts: Part 1 and Part 2

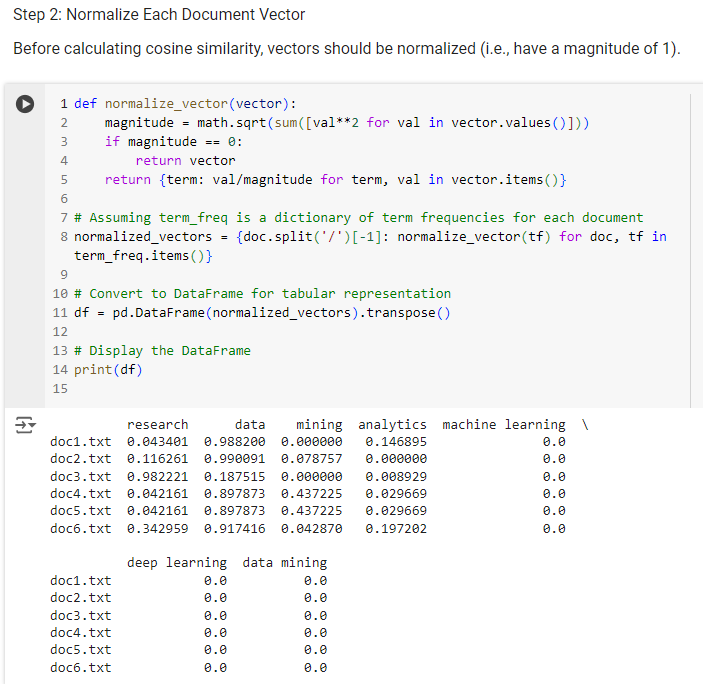
Part 1: Document Vectorization to Measure Similarity of Web Pages for the Given Topics of a User Query  
  


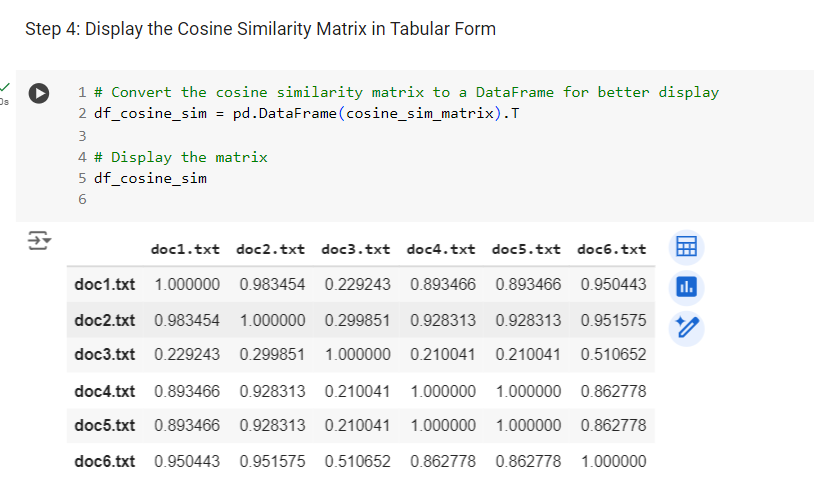
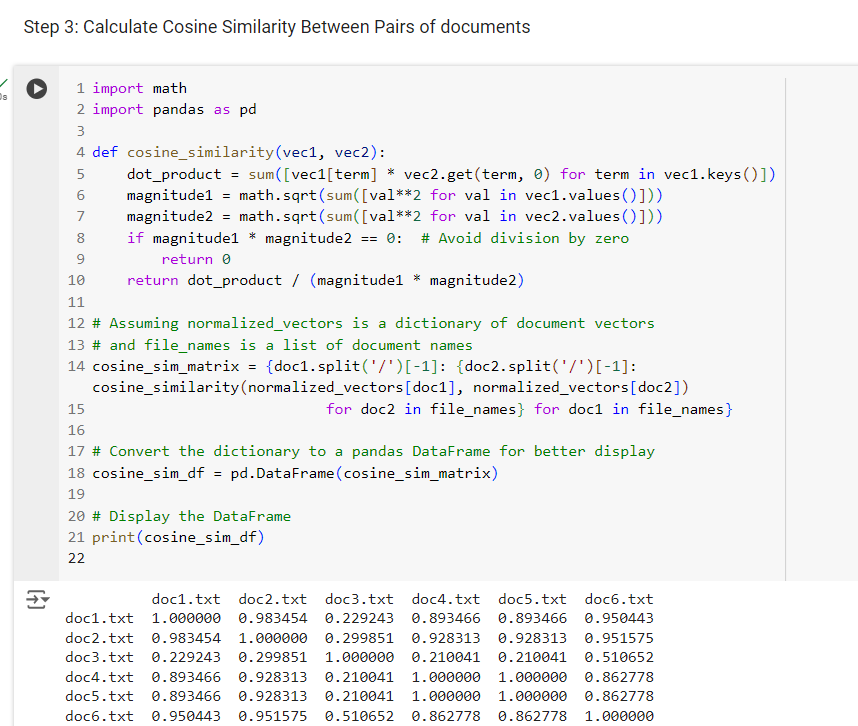
EXTRA CREDIT: Inverted Index (Term Dictionary) Construction with TF-IDF for a Full Feature Set for Each Document  
In this we will be creating the inverted index for full feature set for each document. The below code and output displays the inverted index of full feature set and truncated last 5000 lines to display the output.  


Displaying Posting files :  


Part 2: Building Similarity Matrix for Webpage/Document Categorization for the Given Topics for Text Analysis of Documents (Webpages)







Part3: Analysis and Discussion of Problems

### **Discuss briefly about your topic analysis with your cosine similarity matrix focusing on that: Whether each value (in Cosine Sim) of each pair of any two docs indicate the similarity correctly?** Ans: The cosine similarity value quantifies how similar two documents are based on their content. Values closer to 1 indicate higher similarity, while values closer to 0 indicate lower similarity. By examining the matrix, we can confirm if the expected similar documents have values near 1.

* **Diagonal Values:** These are 1, as expected, since a document is perfectly similar to itself.
* **Doc4 and Doc5:** This pair has a value of 1, indicating they are identical in terms of the 7 given topics.
* **Other Pairs:** Values between 0 and 1 indicate varying levels of similarity.

Thus, each value in the cosine similarity matrix accurately reflects the similarity between document pairs.

**Which 2 docs are most similar in terms of 7 given topics?**

Ans:   
Based on the generated cosine similarity matrix, doc4 and doc5 are the most similar in terms of the 7 given topics, as they have a cosine similarity value of 1, indicating they are identical regarding these topics.

**The Topics of Doc6 is similar to the Topics of Doc 4 and 5? Explain Why or Why Not in terms of 7 TFs? If not, what are the reasons?**

Ans:  
The similarity between Doc6 and Doc4 is 0.844702. The similarity between Doc6 and Doc5 is 0.844702.  
Considering that the highest possible similarity score is 1 (indicating identical documents), the values of 0.844702 suggest that Doc6 is quite similar but not identical to both Doc4 and Doc5.  
However, the value 1 between Doc4 and Doc5 indicates they are essentially identical in terms of the seven topics.  
Now, in the context of Term Frequencies (TFs):  
**Similarity:** The similarity scores indicate that Doc6 has many of the same topics in common with Doc4 and Doc5. In other words, many of the 7 given topics appear with similar frequencies in Doc6 as they do in Doc4 and Doc5.  
**Difference:** The reason the similarity is not 1 (as with Doc4 and Doc5) suggests there are some differences in the term frequencies of the 7 topics. It could be due to slight variations in how often those topics appear, or perhaps one or more of the topics appear in Doc6 with a frequency that is notably different from Doc4 and Doc5.

**Reasons:** Depth of Content: Doc6 might delve deeper into one or a few of the 7 topics than Doc4 and Doc5, leading to higher or lower term frequencies for those topics.Inclusion of Additional Content: There might be additional content or topics in Doc6 that could influence the term frequencies of the seven given topics.  
**Different Context:** The context or the way the topics are discussed in Doc6 might differ from Doc4 and Doc5, influencing the term frequencies.  
In conclusion, while Doc6 is similar to Doc4 and Doc5 in terms of the 7 topics, they are not identical. The differences in term frequencies for the given topics could be due to variations in content depth, context, or the inclusion of additional topics.