Information Retrieval
Sessional 7

A. Mustafi

October 12, 2023

1 Instructions

In today's sessional we shall work with the renowned Reuters corpus (often known as the Reuters 2178 corpus). The corpus is made up of news reports reported by the news agency and each report is annotated with many metadata fields. e.g. Title, Date etc.

2 Exercises

1. How many documents can you find which have the word Cocoa in the title.
2. Generate a new corpus containing only the contents in the BODY tag of each document.
3. Generate a tf-idf matrix representation for the new corpus. Consider only those terms which only contain alphabets or digits and has a minimum length of 5 characters. You can also remove stopwords using the nltk stopwords corpus.
4. Plot a distance matrix showing the cosine similarity between all the documents. You can used matplotlib's matsho function to plot the matrix.