Lab 1 - Indexing and Term Weighting

Instructor
Parth Mehta (parth_mehta@daiict.ac.in)

Teaching Assistants
Adarsh Gupta (202411083@daiict.ac.in),
Bhavesh Baraiya (202101241@daiict.ac.in)

August 2024

Lab Manual

Topics Covered from the Introduction to IR book (Manning et. al): Pre-processing (Section 2.2), Boolean Indexing (Section 1.2), Term Weighting(Section 6.2), TF-Idf Indexing (Section 6.3).

You are given a dataset consisting of approx 32,000 news articles. The data is structured in JSON format; each article has four fields: id, title, summary and text. For this lab session, we will only use the id and text fields. You are expected to complete the list of tasks mentioned below *during the lab hours*. For Step 2 you can use existing python libraries (NLTK or Spacy). For all other problems, you are expected to write a solution from scratch.

Note: The use of GPT for such trivial tasks is generally frowned upon and highlights your lack of interest or/and ability. Also since the instructor uses it almost daily in his other life (to solve real problems, not tf-idf) he can easily detect it with a few simple questions. Save yourself some embarrassment.

1. Data loading
2. Data Preprocessing
    - Case normalization
    - Stop word removal
    - Stemming
    - Removing numbers and non-ascii characters
    - Word Tokenization
3. Creating a list of vocabulary
4. Creating a boolean term-document index
5. Computing Tf and Idf scores for each term
6. Creating a Tf-Idf Inverted index
7. Create compressed inverted index by eliminating some of the terms using one of the following methods:
    - Terms with very high document frequency
    - Terms with a low total count in the corpus. Think why this helps, although it may seem counter-intuitive given the previous point.
    - Terms-document entries with low tf-idf scores

Advanced Exploratory Topics

This lab is designed as a warm-up exercise and some of you might find it too easy. In that case you can look ahead and experiment with the following problems, which we will cover in a future lab session.

1. Explore Pyterrier
2. Implement preprocessing and tf-idf indexing pipeline in PyTerrier
3. Compare the vocabulary size, index size and tf-idf values from the index created by PyTerrier with the one created in the previous excercise.