1 Inverted Index (2 + 4 pts)

Consider the following document collection \( C = \{C_1, C_2, C_3\} \) (given as one document per line):

- new york times
- new york post
- london times

Explain briefly all the relevant data structures constructed (including their characteristics such as sortedness) for implementing (uncompressed) inverted index structure for Boolean Retrieval, and their instantiation for the above example.

2 Ranking using Vector Space Model (3 + 8 + 4 pts)

Consider the following document collection \( D = \{D_1, D_2, D_3\} \) (given as one document per line):

- clear sky blue sky
- the blue car
- sky is nice

Assume that the stopword list contains words the and is, and words are not stemmed. For the given example, show all the relevant data structures constructed and all the relevant statistics computed (such as tf-idf values shown explicitly as ‘(tf,idf)’ with the documents in the postings list) for implementing (uncompressed) inverted index structure for Vector Space Ranked Retrieval. Assume that term frequency factor is just the count of the number of term occurrences in a document (rather than the normalized value) and the inverse document frequency is the reciprocal of the fraction of documents that contain the term.

What are the “relative” relevance scores and the ranking of the documents for the query : sky?
3. Short Answer Questions (5 + 3 + 3 + 8 pts)

1. What are the variable byte codes encoding for the following postings list: 1, 26, 282?

2. What is the fundamental difference between classification and clustering?

3. State and explain clearly all the independence assumptions made to ensure tractability of Naive Bayes Multinomial model.

4. Discuss the robustness of (i) centroid-based classifier, (ii) kNN classifier, and (iii) SVM classifier in terms of the subsets of data points in the training set that determine the classification boundaries.