CMPE493 Homework 1 Report – Zühtü Anıl Tutar

1. Steps of data pre-processing.
   1. NewID, Title and Body parts of the news in Reuters data set was extracted using a function called text\_extractor. This function was used to put the extracted parts into a list.
   2. A function called dissector was used to remove the unwanted parts of the extracted text. This function also assigned the extracted parts to the variables of News object. All of the created News objects were then put in a list called news.
   3. Strings extracted from the dataset were all transformed into lowercase and punctuations in the strings were removed.
   4. A function called news\_tokenizer is used to tokenize the extracted strings. This function removes all substrings that include numbers and removes some of the stop words. This function also removes leftover unwanted data that starts with “lt” extracted from the dataset.
   5. After tokenization, a Trie object for the dictionary and a Linked List object for the postings list was created. The postings lists were stored in a python dictionary.
   6. The python dictionary that holds the postings lists was transformed into a list to store the postings without getting a stack overflow error. This list, along with the dictionary is written in the computer using pickle library.
   7. The final inverted index contains 34715 unique tokens.
2. Data structures
   1. A Trie object was used to store the dictionary.
      1. This object has a function that returns True if a word is in the dictionary.
   2. Linked Lists were used to store the postings lists.
      1. When adding a new node to the Linked List, it checks for the occurrence of the same word in the same document. If the word occurs in the same document, instead of adding a new node, this function adds the position of the word to the position of the same node, storing both positions in the same node.
3. Text

   Description automatically generated
4. 