Search Engine

Daoyuan Chen

10371489

Input files are 5 html pages, each of which contains a news article. The search engine program is written in Python.

1. For each page, the program stores all contents into a string.
2. Use BeautifulSoup module to remove all tags and leave useful text in the string.
3. Replace punctuation with white space from the string.
4. Convert all characters to lowercase and put each word from the string into a list.
5. Filter out stop words from the list using stopwords module from nltk package.
6. Use addWord() method in class Trie to add every word into a trie. Use isEnd property in class TrieNode to indicate whether this node is an external node of a word. This way, we can find a word in the trie which is the prefix of another word, let’s say “bit” and “bite”. Meanwhile, for each newly added word, generate an array [0, 0, 0, 0, 0] and add 1 to the corresponding page index (‘cnn’: 0, ‘fox\_news’: 1, ‘msnbc’: 2, ‘ntd’: 3, ‘voa’: 4), then push to the occurrence\_lists array and set the index\_OL property of the external node to the index in occurrence\_lists.

For example, “fake” is the first word in ‘cnn.html’ => trie, e.isEnd = True, occurrence\_lists.push(occurrence list of “fake”), e.index\_OL =0, occurrence list of “fake” = [1,0,0,0,0].

1. Use searchWord() method in class Trie to search every word in trie. First, it will filter out all punctuation and stop words and put each word in the string into a list. Then use the index\_OL in the external node as a pointer to fetch the corresponding occurrence list in array occurrence\_lists.
2. Use the return values (word and its occurrence list), sort the occurrence list in descending order and print out the search result.