CS 600 Project

Name: Hrishikesh Dinkar Kanade

Introduction

This project implements a search engine using a compressed Trie and an Inverted index.
The code is written in Python using a Jupyter Notebook.

Files

1) Inputs are in the “HTML_files” folder.
2) “Project_CS_600.ipynb” contains the Python code with outputs.
3) “Video_Demo.mp4” contains the video demonstration.
4) “Readme” file.

Implementation

1) Input files: The HTML files are saved locally. Six Wikipedia pages related to bears have been used.
2) Reading the files: The files are read in using BeautifulSoup. The text for each webpage is stored in a dictionary using the web pages as keys.
3) Tokenization: The data is tokenized using regex.
4) Filtering: The stopwords are filtered using the stopword corpus from the NLTK library.
5) Creation of inverted index: The inverted index is implemented using a dictionary where words are keys and the values are the webpages they occur in along with the number of times the word appears in the page.
6) Creation of the Compressed Trie:
    a) Trie nodes: Each node contains a dictionary of child nodes and a boolean representing whether the child is the end of the word.
    b) Compressed trie methods:
        i) Init: initializes the root node.
        ii) Longest_common_prefix: finds and returns the longest common prefix given two strings.
        iii) Print_trie: prints out the contents of the tree.
        iv) Insert: For inserting items into the tree.
        v) Search: searches for a word in the trie and returns True if the word is present in the Trie, otherwise, it returns False.
7) Search and ranking:
    a) Search:
        i) Each word in the input string is searched in the trie.
        ii) If found in the Trie, the word is searched in the inverted index.

iii) The inverted index provides the occurrence list and number of appearances.

b) Ranking:
   i) The pages are sorted by the number of appearances and printed out.
   ii) If there are more than multiple words then the intersection of the sets of pages is printed out.

c) If the word is not present in the trie then a message that the word is not in the vocabulary is printed out.

Sample outputs

Sample 1

```
searchword='pandas'
multiple_search_webpages(searchword, vocab_trie, inverted_indices,filter_words)
```

```
input: pandas

'pandas' found in these pages:

C:/Users/kanad/Documents/CS 600/Project/HTML_files\Giant_panda _Wikipedia.html 123 instances
C:/Users/kanad/Documents/CS 600/Project/HTML_files\Red_panda_Wikipedia.html 45 instances
C:/Users/kanad/Documents/CS 600/Project/HTML_files\Bear_Wikipedia.html 12 instances
```

```
searchword='potato'
multiple_search_webpages(searchword, vocab_trie, inverted_indices,filter_words)
```

```
input: potato

Word 'potato' is not in vocabulary.
```

Sample 2

```
searchword='extinction of pandas'
multiple_search_webpages(searchword, vocab_trie, inverted_indices,filter_words)
```

```
input: extinction of pandas

'extinction' found in these pages:

C:/Users/kanad/Documents/CS 600/Project/HTML_files\Brown_bear_ Wikipedia.html 5 instances
C:/Users/kanad/Documents/CS 600/Project/HTML_files\Giant_panda _Wikipedia.html 1 instances
C:/Users/kanad/Documents/CS 600/Project/HTML_files\American_black_bear_Wikipedia.html 1 instances


'pandas' found in these pages:

C:/Users/kanad/Documents/CS 600/Project/HTML_files\Giant_panda _Wikipedia.html 123 instances
C:/Users/kanad/Documents/CS 600/Project/HTML_files\Red_panda_Wikipedia.html 45 instances
C:/Users/kanad/Documents/CS 600/Project/HTML_files\Bear_Wikipedia.html 12 instances



Intersection of results for all words:

C:/Users/kanad/Documents/CS 600/Project/HTML_files\Giant_panda _Wikipedia.html
```

Sample 3

```
input: bear eating trash

'bear' found in these pages:

C:/Users/kanad/Documents/CS 600/Project/HTML_files\Brown_bear_ Wikipedia.html 324 instances
C:/Users/kanad/Documents/CS 600/Project/HTML_files\American_black_bear_Wikipedia.html 262 instances
C:/Users/kanad/Documents/CS 600/Project/HTML_files\Polar_bear_Wikipedia.html 233 instances
C:/Users/kanad/Documents/CS 600/Project/HTML_files\Bear_Wikipedia.html 214 instances
C:/Users/kanad/Documents/CS 600/Project/HTML_files\Giant_panda _Wikipedia.html 38 instances
C:/Users/kanad/Documents/CS 600/Project/HTML_files\Red_panda_Wikipedia.html 9 instances


'eating' found in these pages:

C:/Users/kanad/Documents/CS 600/Project/HTML_files\Giant_panda _Wikipedia.html 7 instances
C:/Users/kanad/Documents/CS 600/Project/HTML_files\American_black_bear_Wikipedia.html 6 instances
C:/Users/kanad/Documents/CS 600/Project/HTML_files\Bear_Wikipedia.html 6 instances
C:/Users/kanad/Documents/CS 600/Project/HTML_files\Red_panda_Wikipedia.html 5 instances
C:/Users/kanad/Documents/CS 600/Project/HTML_files\Polar_bear_Wikipedia.html 5 instances
C:/Users/kanad/Documents/CS 600/Project/HTML_files\Brown_bear_ Wikipedia.html 4 instances


'trash' found in these pages:

C:/Users/kanad/Documents/CS 600/Project/HTML_files\Brown_bear_ Wikipedia.html 1 instances



Intersection of results for all words:

C:/Users/kanad/Documents/CS 600/Project/HTML_files\Brown_bear_ Wikipedia.html
```

Sample 4

```
input: which bear stays at north pole?

'bear' found in these pages:

C:/Users/kanad/Documents/CS 600/Project/HTML_files\Brown_bear_ Wikipedia.html 324 instances
C:/Users/kanad/Documents/CS 600/Project/HTML_files\American_black_bear_Wikipedia.html 262 instances
C:/Users/kanad/Documents/CS 600/Project/HTML_files\Polar_bear_Wikipedia.html 233 instances
C:/Users/kanad/Documents/CS 600/Project/HTML_files\Bear_Wikipedia.html 214 instances
C:/Users/kanad/Documents/CS 600/Project/HTML_files\Giant_panda _Wikipedia.html 38 instances
C:/Users/kanad/Documents/CS 600/Project/HTML_files\Red_panda_Wikipedia.html 9 instances


'stays' found in these pages:

C:/Users/kanad/Documents/CS 600/Project/HTML_files\American_black_bear_Wikipedia.html 1 instances
C:/Users/kanad/Documents/CS 600/Project/HTML_files\Polar_bear_Wikipedia.html 1 instances


'north' found in these pages:

C:/Users/kanad/Documents/CS 600/Project/HTML_files\American_black_bear_Wikipedia.html 49 instances
C:/Users/kanad/Documents/CS 600/Project/HTML_files\Brown_bear_ Wikipedia.html 39 instances
C:/Users/kanad/Documents/CS 600/Project/HTML_files\Bear_Wikipedia.html 30 instances
C:/Users/kanad/Documents/CS 600/Project/HTML_files\Polar_bear_Wikipedia.html 13 instances
C:/Users/kanad/Documents/CS 600/Project/HTML_files\Red_panda_Wikipedia.html 12 instances
C:/Users/kanad/Documents/CS 600/Project/HTML_files\Giant_panda _Wikipedia.html 1 instances


'pole' found in these pages:

C:/Users/kanad/Documents/CS 600/Project/HTML_files\Polar_bear_Wikipedia.html 3 instances
```