

Coding Challenge

2022 Internship

Write a program in Java or Python to count words in web page content.

Inputs:

1. List of URLs (select your own URLs for testing) provided in a text file, e.g. urls.txt
2. List of words (select your own words for testing) provided in a text file, e.g. words.txt

Outputs:

1. For each URL, print:
 - a. the top 3 words on the given list
 - b. the number of occurrences of each wordList the words sorted by the number of occurrences in descending order.
2. For each word on the given list, print the total number of occurrences across all the specified URLs. List the words sorted by the total number in descending order.

E.g.

Output #1
=====

<https://www.example1.com>

cat - 12

dog - 9

bear - 7

<https://www.example2.com>

lion - 14

tiger - 8

bear - 3

=====

Output #2

=====

lion - 14

cat - 12

bear - 10

dog - 9

tiger - 8

Notes:

- Be sure to format the code for readability
- Add short comments to explain the code
- Be prepared to show how you handle errors, e.g. if the URL is not accessible, or if the input files are not available, etc.
- You may use open source libraries for your program to speed up your implementation. Please be sure to include that in the build/run scripts, and list them in the README file (see below).
- You may use online references or books to design and code. Please be sure to list those references in the README file (see below).
- How to submit the code:
 - Package the code, test data files, and a README file with build/run scripts/instructions into a zip file.
 - Upload the zip file to Google Drive or any other cloud-based storage and email the link to me at vsomayaji@noknok.com.
- README file should include:
 - A brief explanation of the program.
 - Information about how to run it in a terminal on the command-line, e.g:

```
$ java -jar <your jar file> ...
```
 - List of open source libraries used
 - List of references
- Once I receive the code, I'll review it and schedule a code-walkthrough (approx. 1 hour).