Trie Articles

Due Apr 13 by 11:59pm **Points** 100 **Submitting** a file upload **Available** Mar 30 at 9am - May 6 at 11:59pm about 1 month

This assignment was locked May 6 at 11:59pm.

A file, companies.dat, contains a list of company names on each line. Company names may have multiple words in the name. Sometimes, a company might have multiple names associated with it. The first name is the primary name, and the remainder are synonyms. In this case, the company names will be separated by a tab on the same line. (Create a sample version of this file for your testing. The final file used for grading is not published.)

Write a program that can read a news article from standard input. Keep reading until you get a line in the article that consists entirely of a period symbol (.).

Identify each company name in the article, and display each company name on the screen, one line at a time. Always display the primary name of the company identified, not the synonym you found in the text. On the same line, display the "relevance" of the company name hit. Relevance is defined as frequency of the company name appearing in the article divided by the number of words in the article." For example, Microsoft in "Microsoft released new products today." should result in a relevance of 1/5, or 20%. If two names for the same company match, they count as matches for the same one company. Display the relevance in percentage. You should ignore the following words in the article (but not the company name) when considering relevance: a, an, the, and, or, but

You must normalize the company names for the search. Punctuation and other symbols should not impact the search. So the appearance of Microsoft Corporation, Inc. in the companies.dat file should match with Microsoft Corporation Inc in the article. However, the search *must* be case sensitive.

Output:

Company	Hit Count	Relevance
Microsoft	6	4.38889%
Apple Inc.	4	3.08333%
Verizon Wireless	2	2.38889%
Total	**12**	**10%**
Total Words	**120**	

Output should consist of

- Each Company Name, Hit Count, and the Relevance (Relevance = HitCount / Total Number of Words).
- The second to last row of your output should read Total, Total Hit Count, and Total Relevance.
- The last row should simply output the total number of words in the file.

Note: You must not submit your "node_modules" folder if you are working on NodeJs/JavaScript. (Just submit your JavaScript source code and package.json file)

Trie Articles

Criteria	Ratings		Pts
Input: Prompt user for a news article.	5.0 pts Full Marks	0.0 pts No Marks	5.0 pts
Input: Read data from file named "company.dat". (No points if either filename is incorrect or used absolute path)	5.0 pts Full Marks	0.0 pts No Marks	5.0 pts
Calculate: Company's hit count (includes synonym)	5.0 pts Full Marks	0.0 pts No Marks	5.0 pts
Calculate: Company's Relevance (Must be decimal a value up to 4 digits. Ex: 6.000%)	5.0 pts Full Marks	0.0 pts No Marks	5.0 pts
Stopwords: Ignore words "a", "an", "the", "and", "or", and "but". (-8 points if these words in company names are ignored)	10.0 pts Full Marks	0.0 pts No Marks	10.0 pts
Output: Every line should have Company Name, Hit Count, and the Relevance	5.0 pts Full Marks	0.0 pts No Marks	5.0 pts
Output: Second last row should have Total, Total Hit Count, and Total Relevance.	5.0 pts Full Marks	0.0 pts No Marks	5.0 pts
Output: The last row should have the total number of words in the file.	5.0 pts Full Marks	0.0 pts No Marks	5.0 pts
Data Structure: Implementation of Tries	30.0 pts Full Marks	0.0 pts No Marks	30.0 pts
Search: Normalize company name	5.0 pts Full Marks	0.0 pts No Marks	5.0 pts
Search: No impact of punctuation and other symbols	5.0 pts Full Marks	0.0 pts No Marks	5.0 pts
Search: Case sensitive	5.0 pts Full Marks	0.0 pts No Marks	5.0 pts
Coding Style and Test Cases	10.0 pts Full Marks	0.0 pts No Marks	10.0 pts
Note: (a) Late submission penalty per policy (b) 5 points penalty if the output for improper format and indentation.	0.0 pts Full Marks	0.0 pts No Marks	0.0 pts
			Total Points: 100.0