ASSIGNMENT 3 – TRIES ARTICLES

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.