**Mavens Coding and Data Exercise**

**Coding Exercise**

This exercise involves counting word sequences within text. For example, in the text:

*The fat cat sat on the mat.*

the 5-word sequences are:

the fat cat sat on

fat cat sat on the

cat sat on the mat

Please produce a Python script which takes as input a file containing short pieces of text, one per line, UTF-8 encoded. Your code should read the texts from the file and compute all the 5-word sequences which occur in any line of text. The output, written to a UTF-8-encoded CSV file, should be the ten most commonly-occurring sequences and how often each occurs.

A word sequence is considered the same as another if the words it contains match, regardless of case or punctuation that occurs between words. For example:

*the cat in the hat*

is the same word sequence as:

*The cat. In the hat!*

A line break terminates a word sequence.

Your source code should include either comments, or a doc string, which explains any known limitations of the code, or any assumptions that you have made.

A short file of test data should also have been sent to you, if you wish to use it.

Please spend up to two hours on this exercise. If you do not have a complete solution, send it in with comments explaining any missing sections of code, or sections which don’t work to your satisfaction.

**Data Exercise**

Consider again word sequences within texts, as described above. Outline a method for using word sequences to detect positive and negative sentiment in texts.

For example, the texts *This is a low quality product* and *The weather is awful* have negative sentiment, while the texts *The weather is great* and *This product helps me* have positive sentiment.

You do not need to implement the method you outline, and your answer should be no longer than around 700 words. You can assume that you have a representative sample of negative and positive sentiment texts available to you.

If you have any questions, then please do not hesitate to contact me on email: luke.russon@mavens.co.uk