University College Dublin
MIS41110 Programming for Analytics
Practical Sheet 4

Miguel Nicolau

1. Download the file "car_opinions.csv" from BrightSpace. Write a program that asks the user for the name of the reviews file. Then, for each word in each review in the file, add a score of 1 each time the word appears in a positive review, and a score of -1 every time it appears in a negative review. Save the resulting list of words and their associated scores to a new file, whose name will be given by the user.

2. Use the resulting list with the program "analyser.py", discussed in class and available in BrightSpace. What is its score? Can it beat the hand-written file "car_sentiment.csv"?

3. Tune your program. Save only words of a certain minimum length (parameter L), and only if they occur a mimimum number of times (parameter O). Can you find good values of L and O that will increase the score of your list of words?

4. Implement other ways to increase the predicting power of your list. Things to consider:

   - Score words based on whether they appear in a review or not, instead of how often they appear per review (i.e. 1 point if a word appears in a positive review, regardless of how often it appears in that review).
   - Only store words which appear consistently in the same type of review (positive or negative). Ex: only store a word if it appears 66% or more in positive/negative reviews.
   - Take into account the positive/negative appearance ratio, discussed in the previous point, when scoring a word.
   - Modify the scoring system, such that each word ends up with a score between -1 and 1.

Do not forget to use good programming skills, including good variable names, code structure, comments, and a good compromise between readable code vs. effective code.