Technical Test

In the scope of our work, we may be called upon to apply both custom and standard machine learning techniques and algorithms.

**When responding to the following questions please don’t forget to include your code and results for the data enrichment!**

## Question 1

For this exercise, we will ask you to write a toy algorithm for the purposes of enrichment of data. The purpose of this algorithm is to take a given text dataset and a set of sentiment terms with scores, and with this assign to nouns within each record a sentiment based on nearby sentiment terms (in an extremely naive way).

The following is provided:

1. CSV of text, where each row represents a single customer review.
2. CSV of labels, where each row represents the classification of that review
3. CSV of sentiment terms, with score of either 1 or -1 representing this positive or negative sentiment.

The algorithm is as described:

1. For each record in customer reviews:
   1. Find each of the given sentiment terms.
   2. For each term found, locate the nearest noun in the sentence.
   3. Assign to this noun the same sentiment value as the sentiment term. In the case of the same noun appearing multiple times (regardless of position):
      1. Find the sum of all sentiment scores assigned to that noun
      2. Assign a final sentiment value of 1 to the noun if positive, -1 if negative, and 0 otherwise.

Please code the above algorithm in Python and apply it to the test dataset provided.

## Question 2

Now we want to use the sentiment information associated with the nouns as calculated via the above algorithm. Produce a second dataset enriched in some way with the sentiment information found via the algorithm above -- your choice as to how to do so.

## Question 3

Using the supplied labels, perform a classification of both the old and enriched datasets using your choice of classification techniques. Discuss briefly the difference (or similarity) in performance between the original and enriched dataset, both overall and per class.

## Question 4

Realistically, this algorithm (in question 1) will probably not provide meaningful enrichment to the features in real-world data. Is there anything you’d suggest to improve it? What else would you do to improve signal?