Welcome to the assessment for Data Scientist role: Sentiment Analysis

We at Novo Nordisk, deal with lot of structured and unstructured text data. Understanding and analysing textual data, and inferring context is an important aspect of natural language processing. Your role will require you to have a strong working knowledge for

- pre-processing textual data,
- different NLP techniques, and the appropriate situation for each technique
- applied statistics
- exploratory analysis

A total of 2 questions on each of the above topics are listed below, and you have a total of 2 days to work on these questions. Please email your responses and outputs back to us.

There are no right or wrong answers; we would like to use this as a starting point for a discussion on how you might approach a task like this: Workflow, tools, metrics and data considerations.

We would of course like you to use good coding practises and show your exploratory skills as well.

You can use tools like **Python/ R / Jupyter notebook** to share your solutions.

All the best!

Exercise 1:

Attached along with this document is a data file (data.csv). It contains approximately 15k tweets which have been classified as positive, neutral and negative. As part of this exercise, we are going to evaluate approach to train a model to identify sentiment.

Using attached data as training set, develop both a **machine learning model and a deep learning model** to predict sentiment of text. Please note, the data engineer who was extracting this data on weekend was not so happy doing it.

Outputs expected:

1. Create and share your Jupyter notebook files for model training. You can choose to use Google Collab.
    a. Explain your choice of algorithms and approach you used for your models.
    b. Add comments wherever appropriate.

Exercise 2:

Find all sentence which are similar tweet, with tweet_id= 8233. Explain the reason behind the similarity and methodology which you used to find it.

Good Luck!!!