MSc in Artificial Intelligence (2020-2021)

Natural Language Processing

Exercise 3: Sentiment Analysis

In this exercise you are asked to classify tweets according to the sentiment they convey (i.e. positive, negative or neutral). For this purpose you are given a subset taken from the dataset available for SemEval-2017, Task 4, "Sentiment Analysis in Twitter" (http://alt.qcri.org/semeval2017/task4/index.php?id=results).

You can find the required test data in the following location:

- https://drive.google.com/file/d/1TP55ZimN7g9ckW7pMFk0dKLCG7OPEFt9/view?usp=drivesdk

The data have already been cleaned and ready to use.

You are required to find the sentiment conveyed in the tweets found in the test datasets.

For this task you will need to use the TexBlob Python library.

The documentation for the library can be found here.

In order to calculate the sentiment for the tweets you need to follow the steps below in order:

1. Read the test dataset from the file provided in the link above
2. For each tweet in the dataset you need to create a TextBlob object
3. Read the documentation and find which attribute contains the polarity.
4. Return the polarity of each tweet in terms of positive, negative or neutral (make sure you document your assumptions)
5. Calculate the accuracy, precision, recall and f-score of the classification (*HINT: you will need to use the scikit-learn package for this step. Once installed make sure to check the documentation for sklearn.metrics. All the information you need is there*)
6. Return the confusion matrix for the results.
7. *OPTIONAL: Show the confusion matrix graphically*