T26 - Natural Language Processing with SpaCy

Practical Task 1

To plan for this script, I have broken down the required steps necessary below.

Task plan breakdown:

1. **Download the dataset of product reviews.**
    a. The dataset is essentially a csv file and we are mainly interested in the review strings in the review.txt column.
2. **Implementing a sentiment analysis model using spaCytextblob.**
    a. We will use spacytextblob to check each review to determine its polarity. Polarity of 1 indicates a very positive sentiment, polarity of -1 indicates a very negative sentiment whereas a polarity of 0 indicates a neutral sentiment.
3. **Preprocess the text data**
    a. Dropping the NA values.
    b. Extracting the review text per row.
    c. Removing the stop words leaving only the more important words for sentiment analysis.
4. **Creating a dataframe for sentiment analysis.**
    a. Creating a data frame with 3 columns (reviews, sentiment, sentiment score)
    b. Running the scikit packages to make sentiment predictions on the product reviews.

Model output based after training and testing the data:

Accuracy: 0.89

	precision	recall	f1-score	support
Negative	0.83	0.19	0.31	53
Neutral	1.00	0.06	0.11	69
Positive	0.89	1.00	0.94	878
accuracy			0.89	1000
macro avg	0.91	0.41	0.45	1000
weighted avg	0.89	0.89	0.85	1000

Precision is generally quite high for negative, neutral and positive reviews.

The model has a perfect recall for positive reviews but scores quite low for neutral reviews and also quite poorly on the negative reviews.