NLP Assignment-3 Report

Name: Saransh Sharma
Roll Number: 20CS30065

Part-1:

Word2Vec

i) Text Preprocessing Steps:

- Lowercasing: All text data was converted to lowercase to ensure consistency.
- Remove Punctuation: Non-alphanumeric characters and whitespace were retained.
- Replace Backslashes: Backslashes were substituted with a space character.
- Remove Stopwords: Common English stopwords were eliminated using NLTK's stopwords corpus.
- Stemming: Porter Stemming algorithm was applied to reduce words to their root or base form. It was found to provide superior results compared to lemmatization.

ii) Vocabulary Parameters:

- Min_count: 10
- Max_count: 80
- Number of words in vocab: 1005

iii) Neural Network Parameters:

- Mean Sentence Length: 26.0216
- Median Sentence Length: 26.0
- Min Sentence Length: 9
- Max Sentence Length: 97
- Embedding Dimensions: 100
- Number of Iterations: 300 (determined based on the development set)

iv) Evaluation:

- Test Accuracy: 0.746

- Macro F1 Score: 0.744390254253956
- Confusion Matrix:

	Predicted Class 0	**Predicted Class 1**	**Predicted Class 2**	**Predicted Class 3**
True Class 0	**95**	12	12	6
True Class 1	7	**108**	5	5
True Class 2	10	7	**81**	27
True Class 3	9	12	15	**89**

- Classification Report:

	precision	**recall**	**f1-score**	**support**
0	0.79	0.76	0.77	125
1	0.78	0.86	0.82	125
2	0.72	0.65	0.68	125
3	0.70	0.71	0.71	125
accuracy			0.75	500
macro avg	0.74	0.75	0.74	500
weighted avg	0.74	0.75	0.74	500
