Week 6 Research Notebook

From the reading:
The importance of linguistic features, and application of machine learning algorithms by converting them and doing a quantitative analysis on it is highlighted by the paper.

From the research topic "Universities":

There are multiple techniques to find out document similarity, and ranking of documents based on their similarity. The LSA approach, which is **Latent Semantic Analysis** seems most appropriate to be used in the evaluation of university brochure or sop documents and find out similarity between them to contrast it against the rankings that these universities hold in a worldwide ranking context. (Other approaches are LDA, Word Embeddings, Topic Modelling, Hierarchical clustering and these techniques have different strengths and weaknesses, and their suitability depends on the specific research question and data available)

QS world ranking of universities can be referred. And the ranking can be compared using correlation metrics such as **Spearman's rank correlation coefficient or Kendall's tau-b**. (Other techniques are Pearson correlation, Point-biserial correlation, Phi coefficient, Cramer's V and Biserial correlation)

Based on the reading the following approach should be employed:

1) Obtain a list of universities with their statements of purpose and their rankings in the QS World University Rankings.
2) Calculate the document similarity metric as described in the previous answer for each pair of university statements of purpose.
3) Aggregate the similarity scores for each university by taking the average or maximum similarity score with all other universities.
4) Rank the universities based on their similarity scores.
5) Rank the universities based on their QS World University Rankings.
6) Compare the rankings using correlation metrics such as Spearman's rank correlation coefficient or Kendall's tau-b.