**Project Report**

**Name:** Yashvardhan Udia **Netid**: yudia2

**T**he project was a detailed analysis of various aspects of Natural Language processing and delved deep into the comprehensive analysis of several critical components.

Throughout the project, I successfully implemented and experimented with several NLP techniques and tools. I was able to:

* Implement Text Preprocessing: I effectively preprocessed raw text data, including tasks like tokenization, stop word removal, and stemming/lemmatization. This preprocessing step significantly enhanced the quality of the data for further analysis.
* Conduct Syntactic Analysis: My code successfully analyzed the syntactic structure of sentences, identifying various patterns such as question sentences, verb placements, Modal verbs usage, noun usage etc. This analysis provided valuable insights into the grammatical correctness of a sentence.
* Utilize Word Embedding: I integrated word embedding techniques, such as spacy vector in this project for semantic analysis. These techniques enabled me to represent words/sentences/essays in a continuous vector space, capturing semantic relationships between them using cosine similarity.

What didn’t work? While many milestones were achieved, few things caused problems.

* Complexity: In general the code could be optimized to run even faster but the complexity of the NLP tasks kept me from addressing these issues (due to bad time-management skills).
* Certain NLP tasks, like syntax analysis, sentence well-formedness were very intellectually intensive tasks and no matter how much research I did to improve these, I always fell short.
* I noticed that after part 1 my codes accuracy was in high 80s but it fell down to high 70s due to the usage of additional parameters which goes to show the need for a more complex overview of the parameters in c3 and d1.
* There was some issue involving the usage of cosine similarity based on high/low essays with their prompt and with a different prompt. All these 4 combinations resulted in almost similar values. This was slightly overcome using the averages and scaling them to 1 to 5, but there is a huge room for improvement here.

To improve the results, a more nuanced approach towards sentence classification, syntax analysis, sentence well-formedness, identification of grammatical errors, and pragmatic/semantic similarity towards the prompt is required. Even though many ideas were used to achieve this, there’s still room for a more complex approach.

To conclude, the various implementations required for us were very helpful in pushing our boundaries beyond what we learnt in the class and explore for ourselves various sub-domains like a sentence could have multiple sentences and that not all sentences should be counted as a single unit.