CSF429 Projects List

- Project selection will be done in FCFS manner.
- Each project has two parts/tasks. However, second part will be available after midsem exams only if the first part is completed.
- Each project must be selected by at least one group.
- Project cannot be changed after September 15th (once checked/approved by IC).

1. Grammar Check for English Language (**maximum two groups**)
   (Preprocessing, Parts of Speech, Dependency Parsing, Language Modelling)

Part 1: Build a grammar checker using the concepts taught in class, you will be given a corpus with some paragraphs labeled correct and unknown; correct paragraphs have no grammatical errors, while unknown may/may not have sentences which have grammatical errors in them. The task is to find all the sentences that have these errors. For example: "In the country there lived a fox. The quick brown fox jump over the fence. Farmer Shaun was terrified." In this paragraph, the sentence "The quick brown fox jump over the fence" is grammatically incorrect.

Part 2: After finding the sentences which have these grammatical errors you have to suggest the correct alternatives. Bonus if your suggestions are also coherent with the context. For example in the previously given paragraph, the sentence: "The quick brown fox jump over the fence" Can be replaced with "The quick brown fox jumped over the fence", "The quick brown fox jumps over the fence" as well as "The quick brown dog jumped over the fence" however the first two alternatives are correct given the context that it was a fox and not a dog.

2. Literature Shelves and Relations (**maximum three groups. Datasets may be different**)
   (Preprocessing, Topic Modelling, Distributional Semantics, Information Extraction)

Part 1: Assume that you were given 10,000 Research papers to read in a bundle, completely un-assorted! Your first job would be to assort them in some hierarchical order, wherein these papers were separated in different sections/shelves. You are given some Computer Science Research papers with their metadata including metadata, The task is to assort them in shelves and sub-shelves. For example: You are given papers A,B,Z,1,2,-1,@,# At the level 1 the clusters could distinct based on character types, (A,B,Z), (1,2-1), (@,#), Further the cluster (1,2,-1) can be divided into ((1,2), (-1)) and so on.

Part 2: For the next step you need a citation graph of the given papers. The task is to find the citation graph, and possibly have a mechanism to find what papers would some new literature cite. The output should be an adjacency matrix. [Note: You can also use this concept to improve your literature shelves].

3. Syntax Analysis in Source Code (**maximum two groups**)
   (Parsing, Language Models, N-grams)

Part 1: All programming languages have a distinct grammar that has to be adhered to for compilation and execution. This arrangement is then used by the parser to compile a program.
IF '(' expression ')' statement _ ELSE statement
statement _ ELSE statement
For the above 2 statements the 2nd line is syntactically incorrect because of the absence of a preceding 'if' statement. Design a parser which is able to determine if a code snippet is syntactically correct.

Part 2: Intelligent code completion is a context-aware code completion feature in some programming environments that speeds up the process of coding applications by reducing typos and other common mistakes.

4. Stylometric Analysis (**maximum two groups**)
   (Preprocessing, Dependency parsing, Lexical and syntactic analysis.)

Part 1: Analyze the writing style of different authors, you will be given a corpus of books written by different authors, the idea is that different authors write in different manners. Using surface level, lexical, and syntactic features analyze the writing styles of different authors. Lexical features: formal word usage, vocabulary level, word concreteness, dialectic and industry words, synonyms, antonyms, etc. Sentence types variation Simple vs compound vs complex Loose vs periodic Declarative vs interrogative vs exclamatory Also investigate the usage of adjectives and phrases, and sentence structures/outlines

Part 2: Using the features analyzed in part 1, attribute authorship of the article/book

5. Aspect based sentiment analysis (**maximum two groups**)
   (Parsing, Information Extraction)

Part 1: Sentiment analysis is a widely used application of NLP. However pure classification still suffers from a lack of context and reasons for those particular sentiments. For example, a product may have a bad review because of the stock availability issues, but still, be a good product. The task is to analyze these sentiments based on certain aspects. Perform base sentiment analysis on sentences and look at the results, investigate the pitfalls of this format of classification.

Part 2: Extract aspect words and categorize them into aspects. And identify the sentiment of these aspects.

6. Cross POS (**maximum three groups**)
   (Morphology/Lexical Analysis, POS tagging)

Part 1: POS tagging is based on the grammar, semantic and syntactic structure of the language. Some languages have higher similarities in terms of semantic and lexical similarity. Compare the performance of semantic rule-based, HMM and RNN models for POS tagging on any dataset

Part 2: Cross lingual learning: Train your model on one language and evaluate on another dataset using the models trained above. (pre-trained multilingual embeddings can be used in this case.) For every cross-lingual scenario quantify the inter-language similarity using Lexical similarity and Semantic similarity

7. Query Formation (**maximum two groups**)
   (language models, preprocessing)

Part 1: Automatically correcting the queries by processing syntactic and semantic features of the words present in the query. Improving this correction by validating the search results.

Part 2: Automatically complete queries using n-gram linguistic models. Compare the quality of completion across various models.