Project Round - 2

Take two books downloaded from the previous round + one new book to be downloaded from the same link. Let B1, B2 and B3 are the books.

For B1 and B2 (after doing the needed pre-processing, done in the previous round):

First Part:

1. Find the nouns and verbs in both the novels. Get the immediate categories (parent) that these words fall under in the WordNet.
2. Get the frequency of each category for each noun and verb in their corresponding hierarchies and plot a histogram for the same for each novels.

Second Part:

1. Recognise all Persons, Location, Organisation (Types given in Fig 22.1) in book. For this you have to do two steps: (1) First recognise all the entity and then (2) recognise all entity types. Use performance measures to measure the performance of the method used - For evaluation you take a considerable amount of random passages from the Novel, do a manual labelling and then compare your result with it. Present the accuracy with F1 score here.

Use B1, B2 and B3 for the following:

Third Part:

1. Create TF-IDF vectors for all books and find the cosine similarity between each of them and find which two books are more similar.
2. Do lemmatization of the books and recreate the TF-IDF vectors for all the books and find the cosine similarity of each pair of books.

Existing NLTK libraries can be used for the above. When preparing the report please state all the packages, libraries used with brief description about the algorithm it uses.

When you are submitting submit the report of Round - 1 and Round - 2 together as a single report. **This report (PDF only) should be submitted in Moodle ONLY.**

Last date for submission of reports: 11.59pm December 21, 2021.