

# Assignment\_1\_Problem\_Statement-0 10 Points Possible

## 6

7/14/2024

Attempt 1



In Progress

**NEXT UP: Submit Assignment**



Add Comment


### Unlimited Attempts Allowed

6/28/2024 to 7/14/2024

#### Details

##### General Instructions:

1. Follow the instructions in each question carefully.
2. Each group is expected to submit jupyter notebook (.ipynb) with output for each cell. The **name of the assignment** file must be the **Group ID**. For Example.  
NLP\_Assignment\_1\_Group\_123.ipynb
3. As it is a group assignment, only one group member needs to submit the assignment on behalf of the group.
4. Please mention the contribution of each group member (NAME, ID, CONTRIBUTION).
5. Submissions using other python IDEs will not be considered for grading.
6. In case the link to dataset is not useful, same dataset can be downloaded from any online resource.
7. A clear explanation for each output obtained is mandatory.
8. Justification of the output obtained for all the tasks is mandatory.

**Data set Link:** [NER\\_Dataset](https://docs.google.com/spreadsheets/d/1D0FSPqZ-w2ATzXxcx8C2LIwoWPba50Ve87yZQRsRyts/edit?usp=drive_link)  ([https://docs.google.com/spreadsheets/d/1D0FSPqZ-w2ATzXxcx8C2LIwoWPba50Ve87yZQRsRyts/edit?usp=drive\\_link](https://docs.google.com/spreadsheets/d/1D0FSPqZ-w2ATzXxcx8C2LIwoWPba50Ve87yZQRsRyts/edit?usp=drive_link))

The goal of this assignment is to focus on Word Embedding and Named Entity Recognition (NER). You will work with a dataset containing sentences tagged with POS and named entity labels. The assignment is divided into three parts: understanding and pre-processing the dataset, implementing a Word2Vec model for word embedding, and designing and evaluating a sequential model for NER.

#### Part I - Understanding the Dataset and Preprocessing

3 marks

You will become familiar with the dataset and perform the necessary pre-processing steps. This includes loading the dataset, explore its structure, and prepare for subsequent modelling tasks.

### **1. Loading the Dataset (1 mark)**

- Load the provided dataset and display the first few rows.

### **2. Exploratory Data Analysis (1 mark)**

- Count the number of sentences.
- Identify the unique POS tags and named entity tags.
- Visualize the distribution of sentence lengths.
- Visualize the frequency of different POS tags.

### **3. Data Preprocessing (1 mark)**

- Tokenize the sentences and map each word to its corresponding POS and named entity tags.
- Convert the words and tags into numerical representations suitable for modelling, such as word indices and tag indices.

## **Part II - Word Embedding using Word2Vec**

**3 marks**

You will implement a Word2Vec model to create word embeddings from the dataset and evaluate its effectiveness. This involves training the Word2Vec model on the dataset and visualizing the resulting word embeddings.

### **1. Word2Vec Model Implementation (1 mark)**

- Implement a Word2Vec model to generate word embeddings from the dataset.
- Ensure your model captures the contextual relationships between words effectively.

### **2. Word Embedding Visualization (1 mark)**

- Visualize the word embeddings using techniques such as t-SNE or PCA.
- Display how different words and their embeddings relate to each other in a lower-dimensional space.

### **3. Evaluation (1 mark)**

- Evaluate the quality of the word embeddings by examining their effectiveness in capturing semantic relationships.
- Present your evaluation results clearly and concisely.

## **Part III - NER using Sequential**

## Model

4 marks

You will design and implement a sequential model (such as RNN, LSTM, or GRU) to perform NER. You will train your model, evaluate its performance, and discuss the results.

### 1. Model Design and Training (2 marks)

- Choose a sequential model and justify your choice.
- Design the architecture of your model, specifying the input layer, hidden layers, and output layer.
- Train the model on the training dataset, ensuring proper handling of sequences and padding.

### 2. Model Evaluation (1 mark)

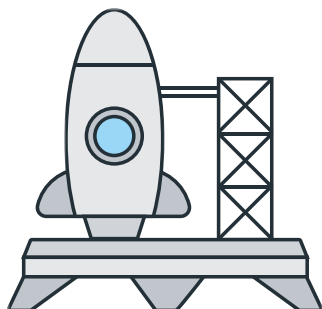
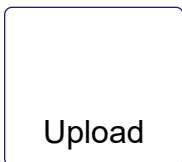
- Evaluate the performance of your trained model on the test dataset using metrics such as precision, recall, and F1-score for each named entity tag.
- Present the evaluation results clearly and concisely.

### 3. Analysis and Discussion (1 mark)

- Discuss the strengths and weaknesses of your sequential model based on the evaluation results. **Answer without justification will not be awarded marks.**
- Suggest potential improvements or alternative approaches to enhance performance. **Answer without justification will not be awarded marks.**

Keep in mind, this submission will count for everyone in your Assignment Groups group.

Choose a submission type.



or

 Webcam Photo

Drag a file here, or

Choose a file to upload

 Canvas Files

Submit Assignment