Assignment - 2 Problem Statement 10 Points Possible 20

9/15/2024

Attempt 1 | In Progress | **NEXT UP: Submit Assignment** | Add Comment

Unlimited Attempts Allowed

8/31/2024 to 9/15/2024

Details

Probabilistic Context-Free Grammars and Parsing

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_2_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.
9. Please access the dataset using BITS Official email.

Link to the Dataset: SMS Spam Collection Dataset (https://drive.google.com/file/d/1kGfjj3jpHlqpLfvNC4yXe-jSDKtHJfZg/view?usp=sharing)

Description of Data: A collection of 5,574 SMS messages in English, labelled as either spam or ham (non-spam). The dataset is ideal for tasks involving language processing and probabilistic grammar analysis.

Task 1: Data Preprocessing and POS Tagging (2 Marks)

1. Download the dataset and load it as a DataFrame.
2. Remove punctuations, special characters, and stop words from the text.
3. Convert the text to lowercase and apply POS (Part of Speech) tagging to the first 5 rows.

Task 2: Probabilistic Context-Free Grammar (PCFG) Parsing (5 Marks)

1. Define a basic context-free grammar (CFG) for simple sentence structures using the

provided SMS data.
2. Convert the CFG into a Probabilistic CFG by calculating rule probabilities based on the dataset.
3. Parse the first 2 rows of text using the PCFG and visualize the resulting parse trees.

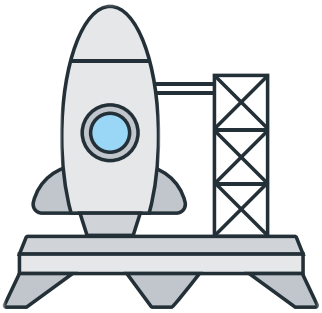
Task 3: Improving PCFGs (3 Marks)

1. Identify potential weaknesses of your PCFG based on the provided SMS data.
2. Propose improvements by splitting non-terminals or modifying rule probabilities. Demonstrate the impact of these changes by parsing a new sentence from the dataset.

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

Choose a submission type.

Upload | Arc



Drag a file here, or

Choose a file to upload

or

Webcam Photo

Canvas Files

Submit Assignment