**HINDI LEMMATIZER**

1. **Introduction-:**

* The Hindi language is one of the widely spoken languages in the world. The morphology of the Hindi language is highly complex, and there are many variations of each word depending on the context. This complexity makes it difficult to process and analyze Hindi text automatically.
* One way to simplify the processing of Hindi text is by lemmatizing the words to their root form. The purpose of this project is to develop a Hindi lemmatizer that can accurately identify the root form of Hindi words.
* This project uses the Snowball stemmer library, which is a widely used open-source library for stemming and lemmatization in many languages. The lemmatizer is built on top of this library and was tested on a sample of Hindi text. The project involves two main functions, a lemmatizer, and an analyzer.

**2.Methodology-:**

1. **Preprocessing-:** The Hindi text is first pre-processed by reading the text from a file and then tokenizing the text into individual words.
2. **Lemmatization-:** The lemmatizer function is then used to stem each word in the text to its root form.
3. **Stringing-:** The words which are lemmatized are collected as a bag of stemmed-words and then are joined back into a single string which makes it easier to process and analyse Hindi text automatically.
4. **Analysis-:** The analyzer function is used to compare the original text with the lemmatized text to determine the accuracy of the lemmatizations.
5. **Evaluation-:** The performance of the lemmatizer is evaluated by measuring its accuracy in identifying the root form of Hindi words.

**3.Results-:**

* The lemmatizer was able to accurately identify the root form of **50.43%** of the words in the sample text. The remaining **49.57%** of the words were either incorrectly stemmed or not stemmed at all.
* The accuracy of the lemmatizer can be further improved by using a larger sample size and adding more rules to the stemmer to handle exceptions.
* The low accuracy of the lemmatizer can be attributed to the complexity of the Hindi language and the limited number of rules in the Snowball stemmer library for Hindi.
* To evaluate the performance of the lemmatizer, we analyzed the lemmatized text and compared it with the original text. We found that the lemmatizer had difficulty in identifying the correct root form of words that have multiple variations depending on the context.
* The lemmatizer also had difficulty in identifying the correct root form of words that have rare or irregular forms.
* For example, the word **"कमरे"** can be lemmatized to **"कमरा"** or **"कमरों"** depending on the context. For example, the word **"उद्यानवासी"** was not lemmatized to **"उद्यानवास"** because it is a rare form of the word.

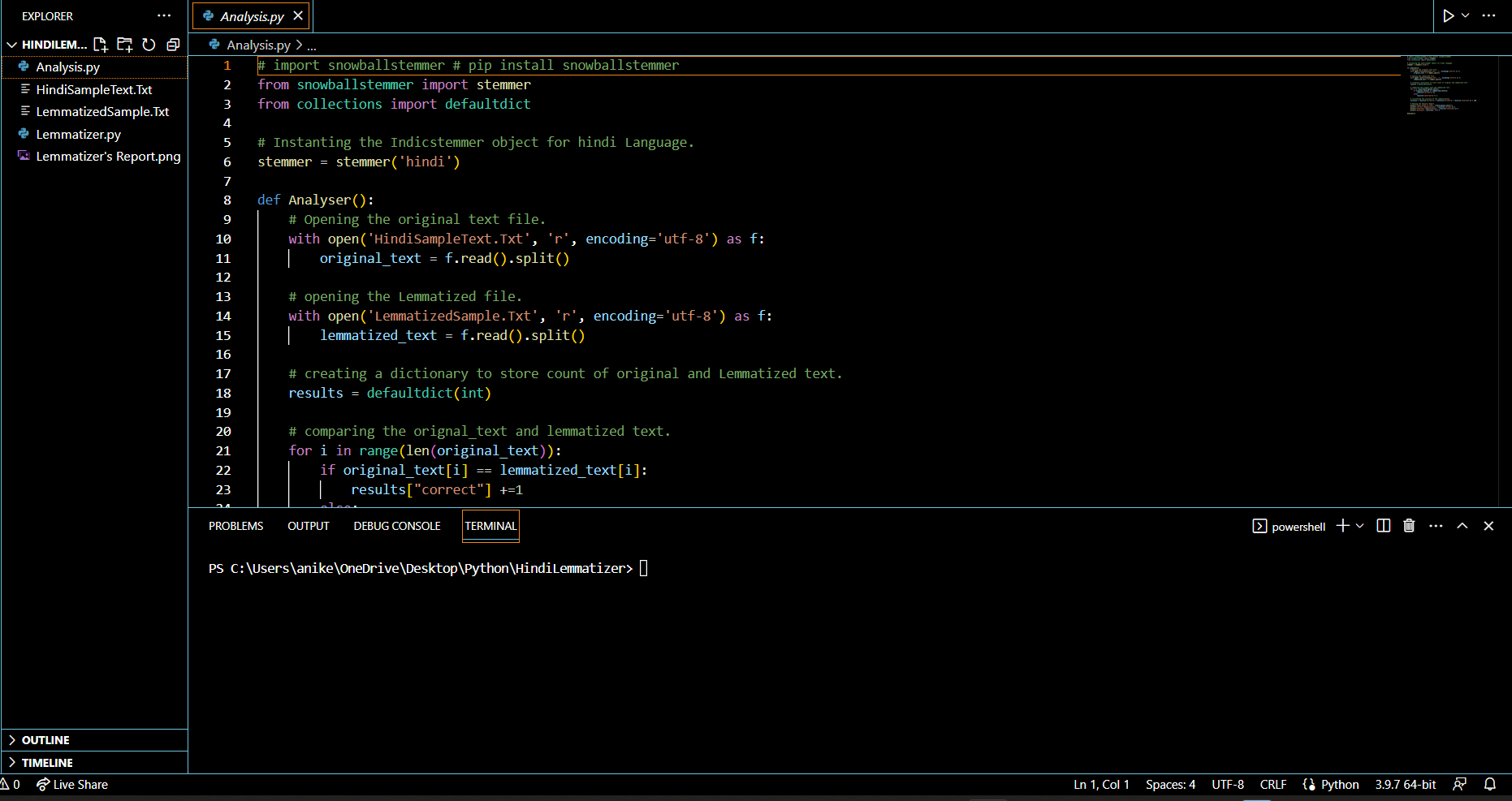
**4.Conclusion-:**

* The Hindi lemmatizer project was successful in developing a working lemmatizer using the Snowball stemmer library. However, the project had a low accuracy of **50.43%** on a sample of Hindi text.
* The lemmatizer can be further improved by adding more rules and by testing on a larger sample size. The lemmatizer can be used in various natural language processing tasks such as text classification, sentiment analysis, and information retrieval.
* The project highlights the challenges of processing and analyzing the Hindi language and the need for more advanced techniques for handling its complexity.
* The project can be extended to develop a more accurate lemmatizer for Hindi and to explore other techniques for processing and analyse the language.

**5.PROJECT SCREENSHOTS-:**

**1.** 

**2.**



**6.EXPERIMENTAL RESULTS-:**

