# NLP ASSIGNMENT 2 SUBMITTED BY: MUHAMMAD KAMRAN

**SUBMITTED TO: SIR ABDUL AHAD ROLL NO: 8837**

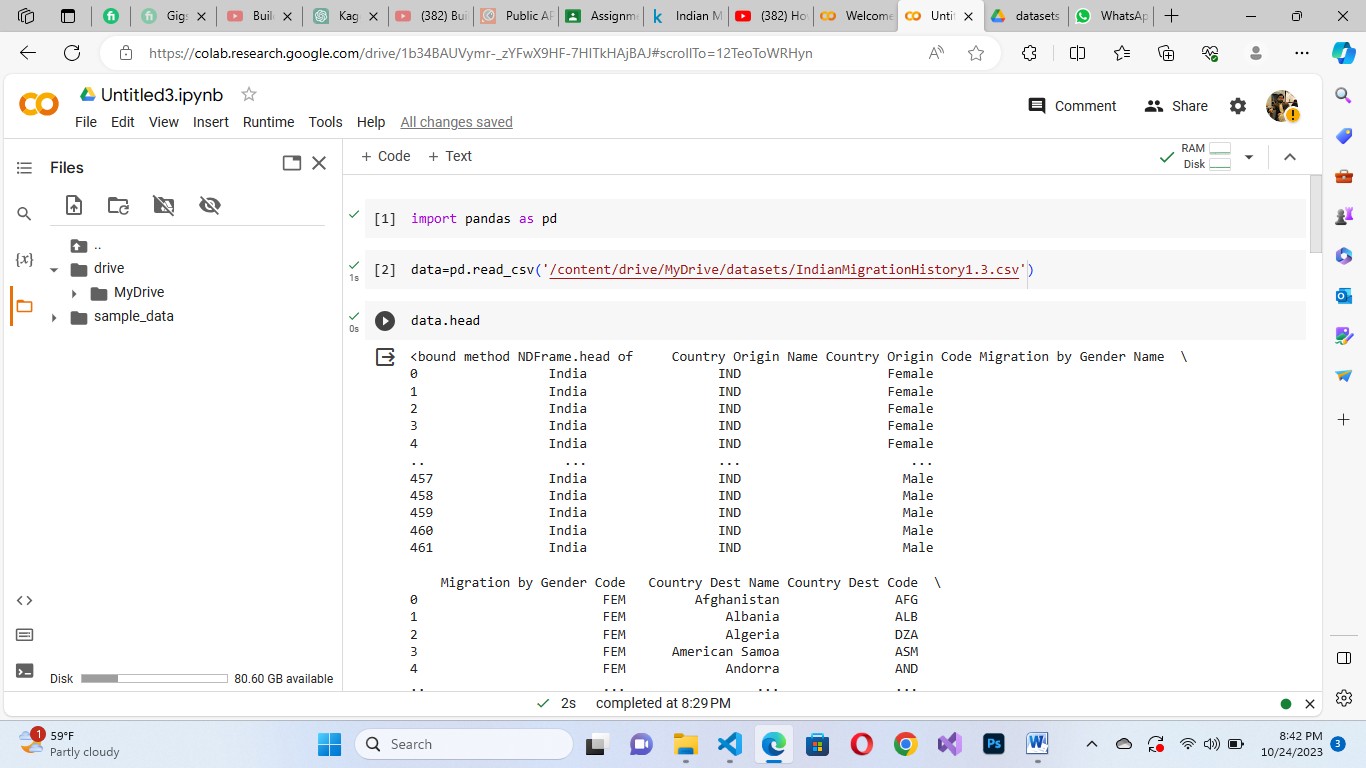
# DATE: 24/10/2023 BSSE-7TH

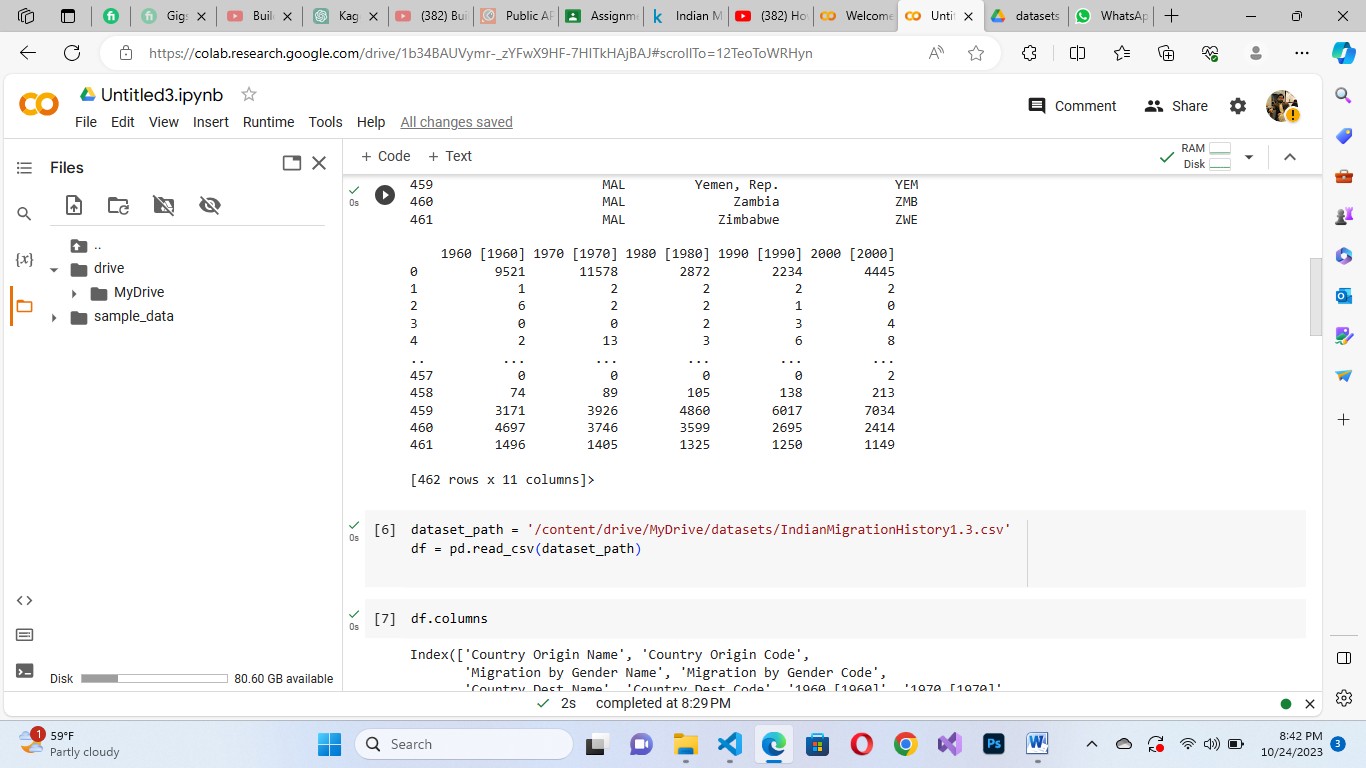
QUESTION NO 1

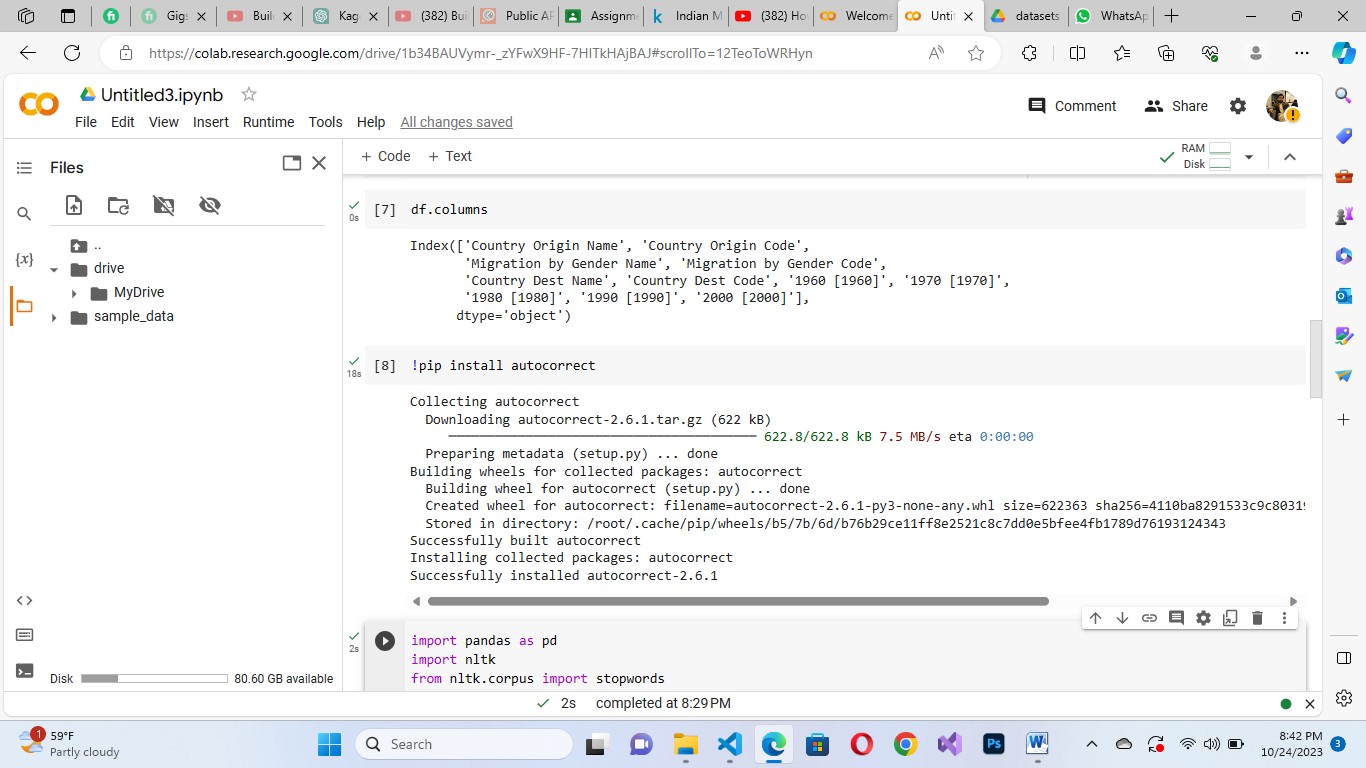
Choose a dataset of your choice and apply preprocessing steps on it as was discussed in class. i.e Text Lowercase, Tokenization, Removal of (Punctuations, stop words, numeric tokens, html/xml tags, white spaces ), stemming and Lemmatization, spell correction, handling special characters etc.

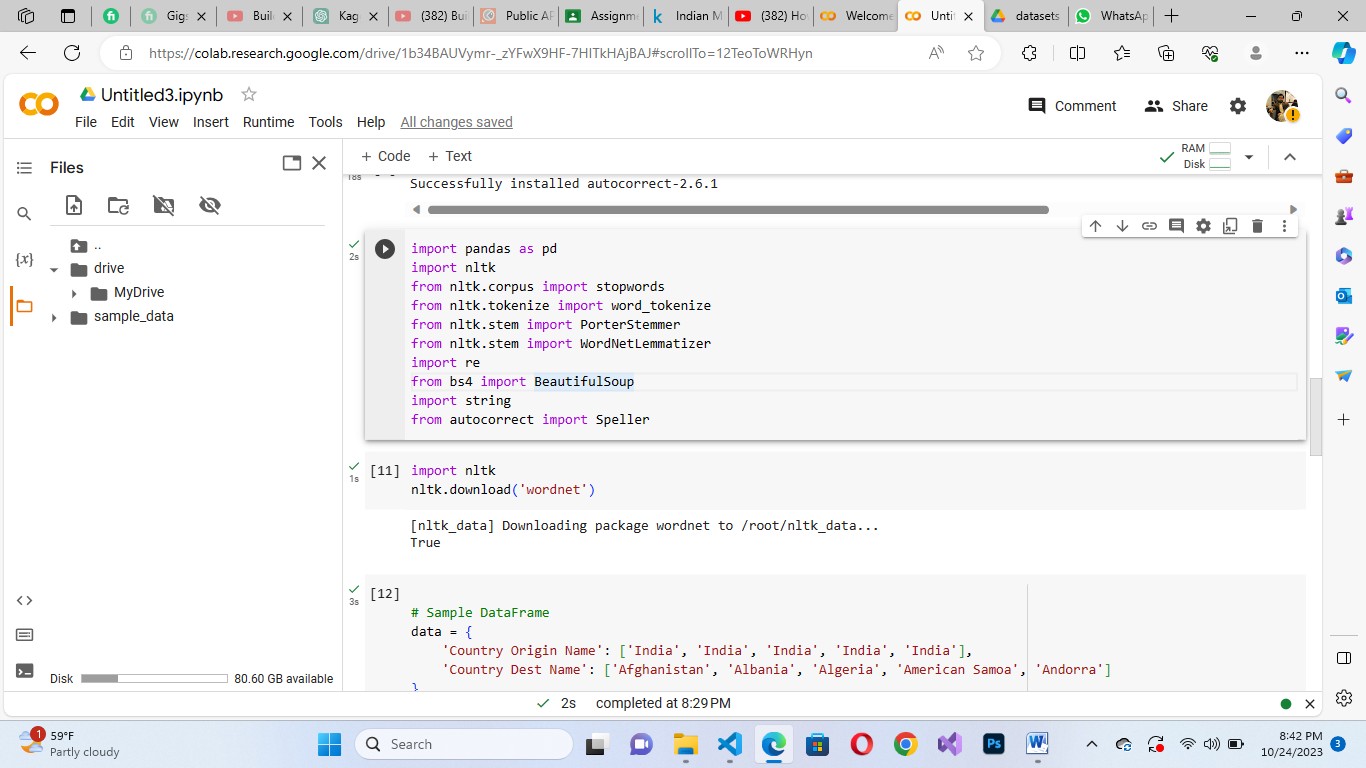
**DATASET LINK**: [Indian Migration History (kaggle.com)](https://www.kaggle.com/datasets/rajacsp/indian-migration-history)

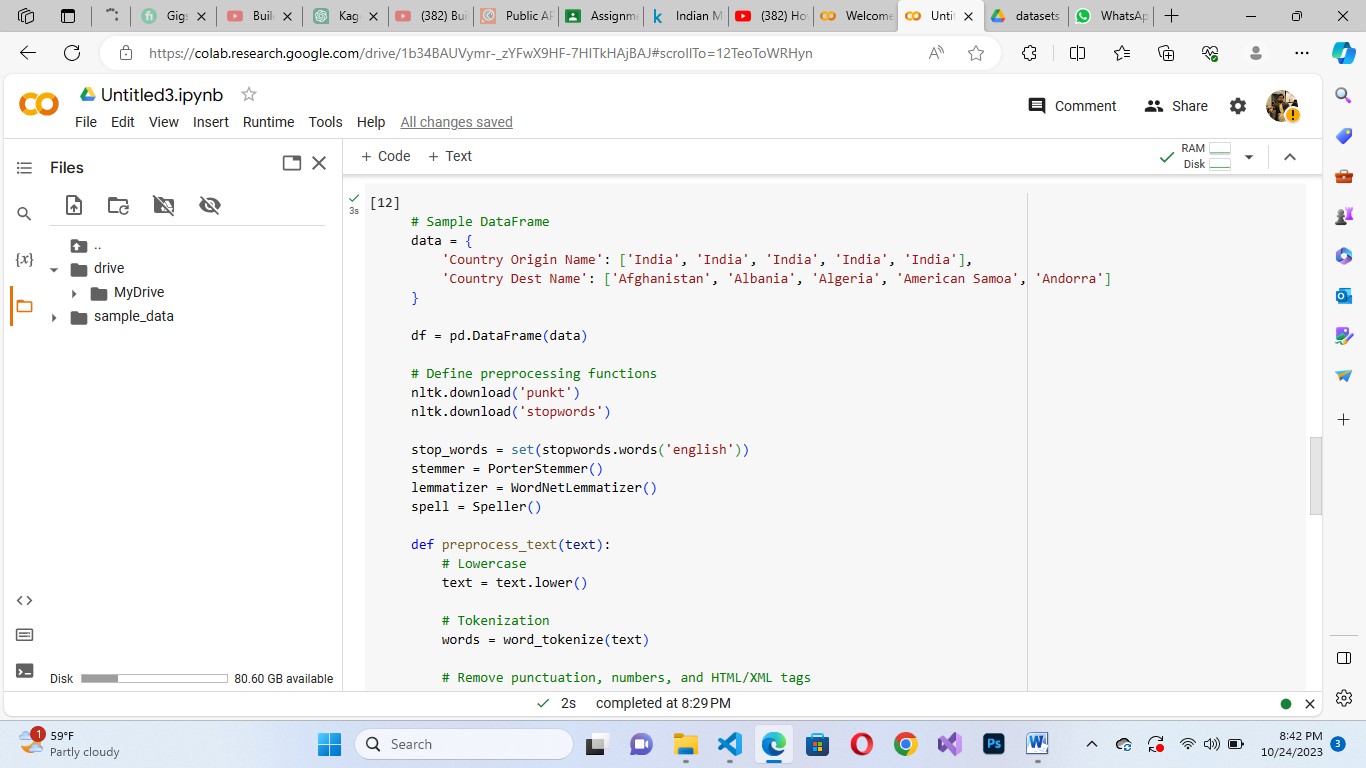
SCREENSHOTS OF ALL PREPROCESSING:

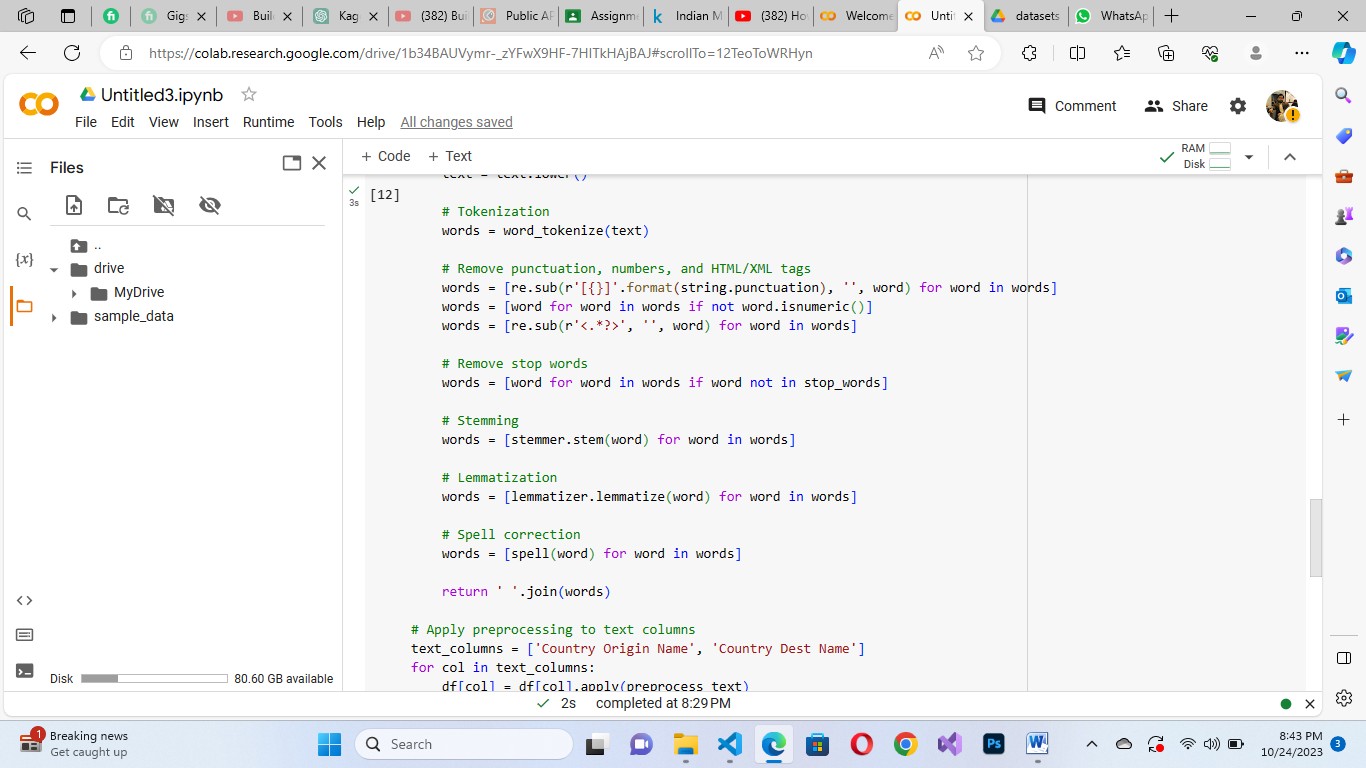


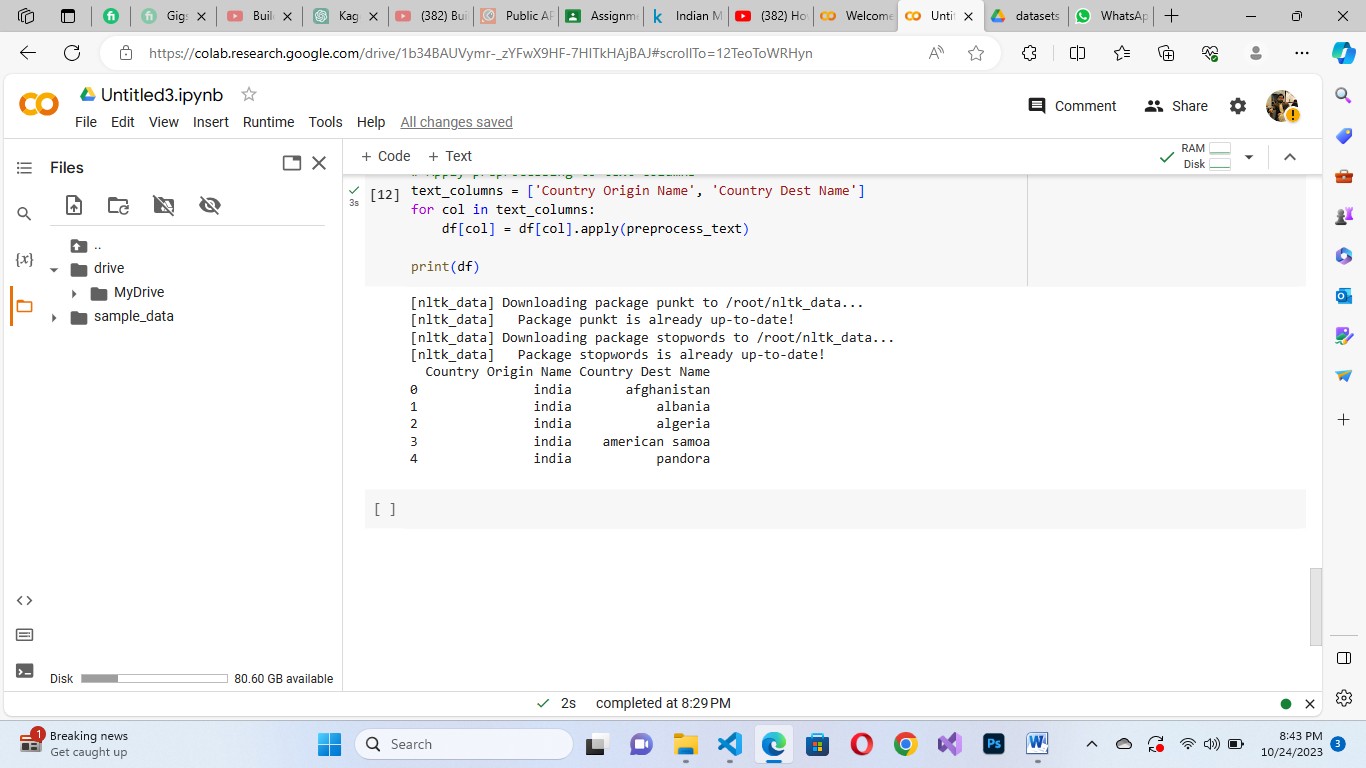












# QUESTION NO 2

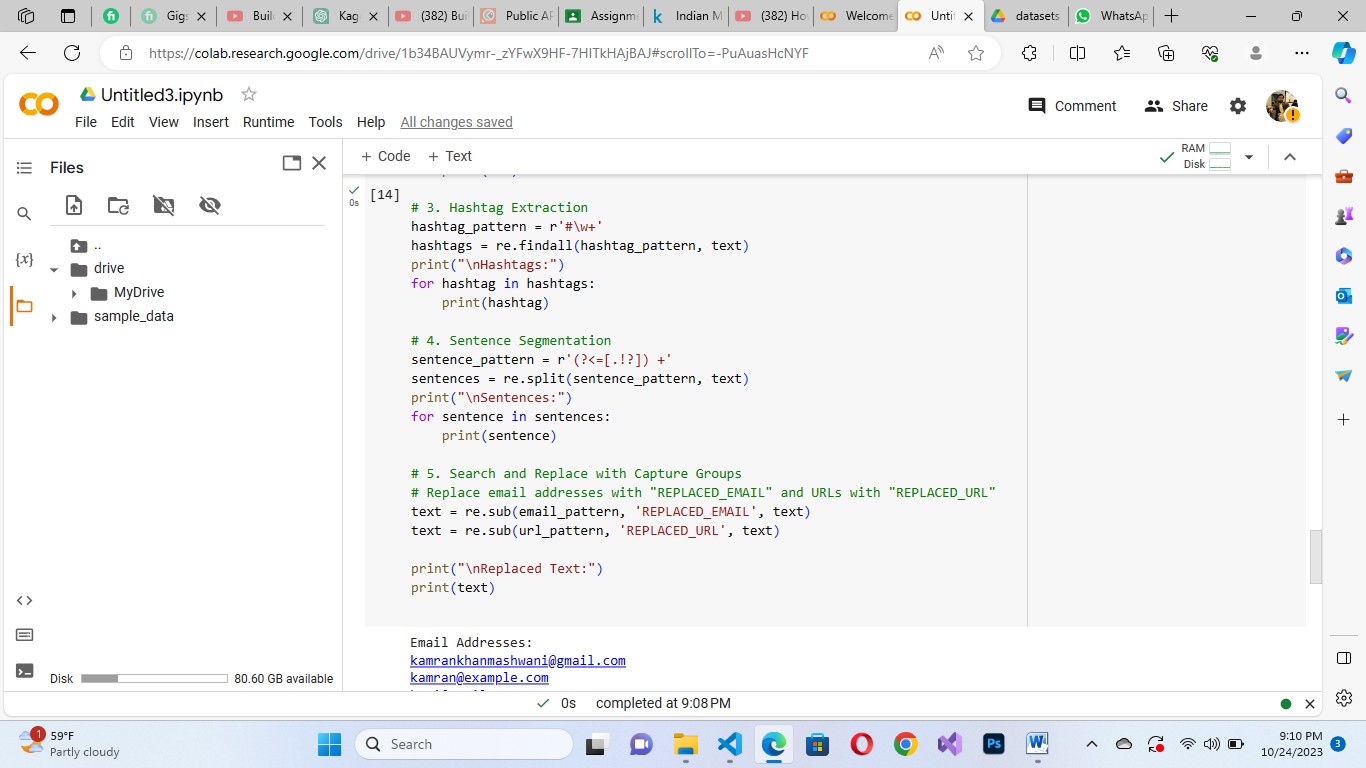
Write a program in python which utilizes regular expression (R.E) to perform the following tasks.

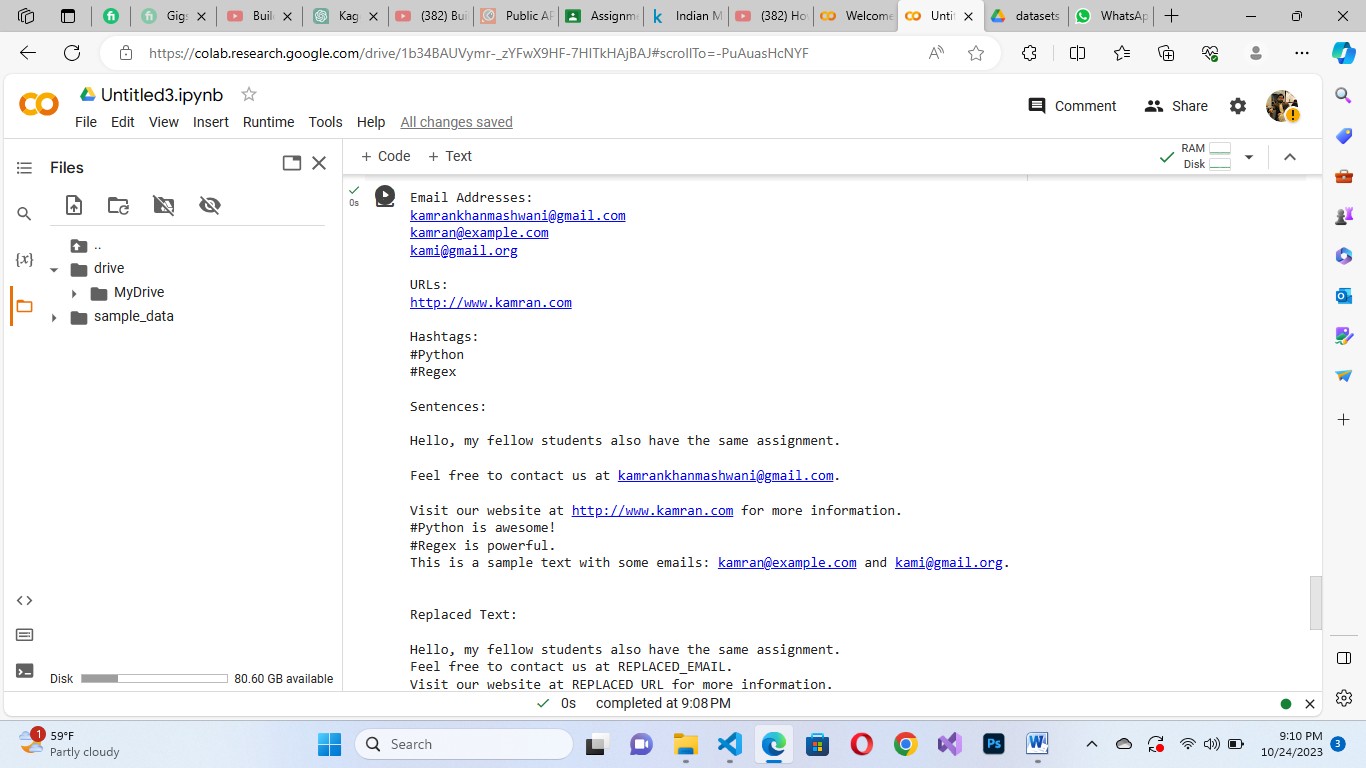
Email Address Extraction URL Extraction

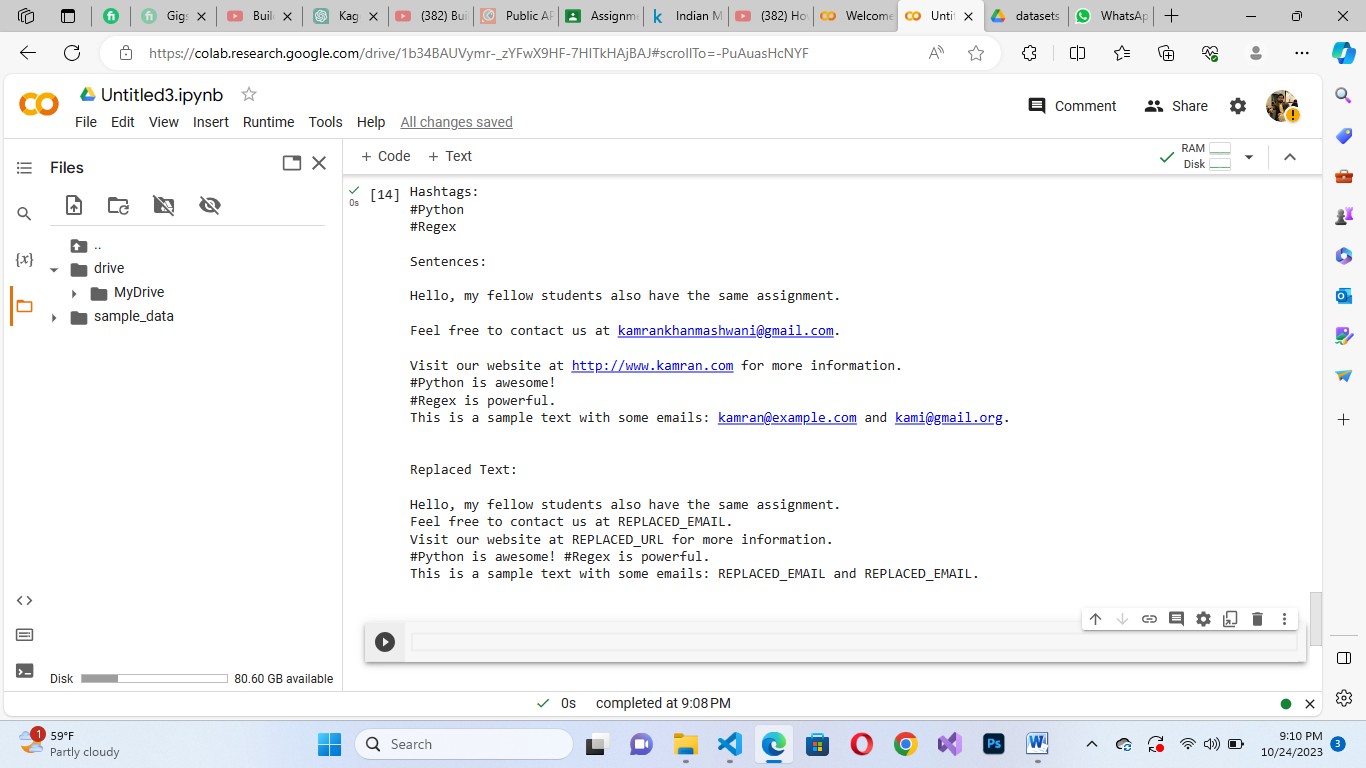
Hashtag Extraction Sentence Segmentation

Search and replace with capture groups. (search and replace certain text.)

# SOLUTION:







**QUESTION 3**

**Text Prediction using N-Gram models. Given a sequence of words, predict the next word using bigrams or trigrams.**

SOLUTION:

