CSCE 5290 – NATURAL LANGUAGE PROCESSING

Project 1

**Introduction:** N-gram model is used in NLP. It is used to predict the probability of the word sequence based on its previous words.

The Different N-gram models are Unigram model, Bigram model, Trigram model and N-gram model.

The Functions used to calculate the probability to predict the next word was based on Markov’s assumption i.e., P(word/count) = Count (text, word)/count(text)

In the example: Birds are excellent indicators of biodiversity change.

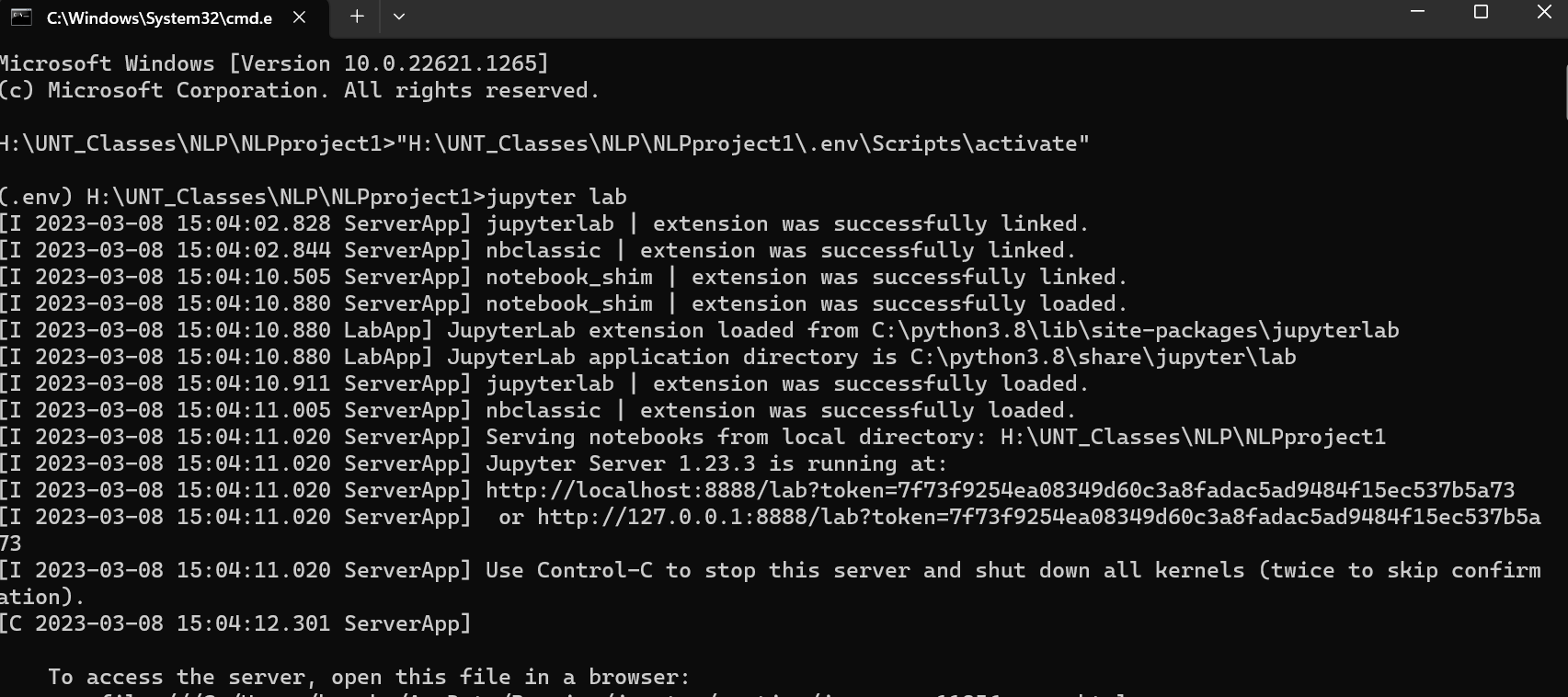
The unigram output would be [birds][are][excellent][indicators][of][biodiversity][change]

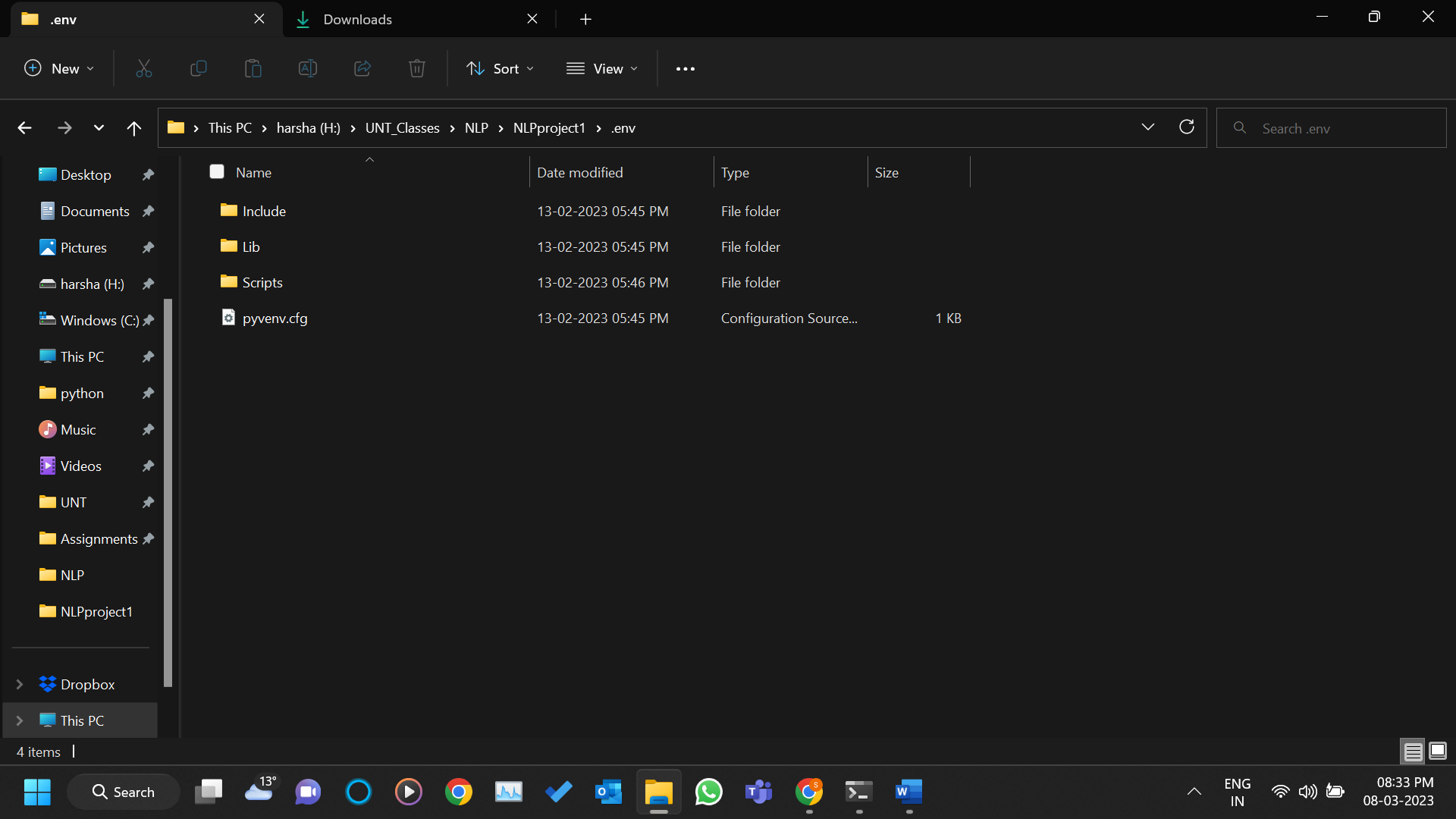
While the bigram output is [birds are] [are excellent] [excellent indicators] [indicators of ][of biodiversity][biodiversity changes]

For Tri gram output would be [birds are excellent] [are excellent indicators] [excellent indicators of] [indicators of biodiversity ][of biodiversity changes]

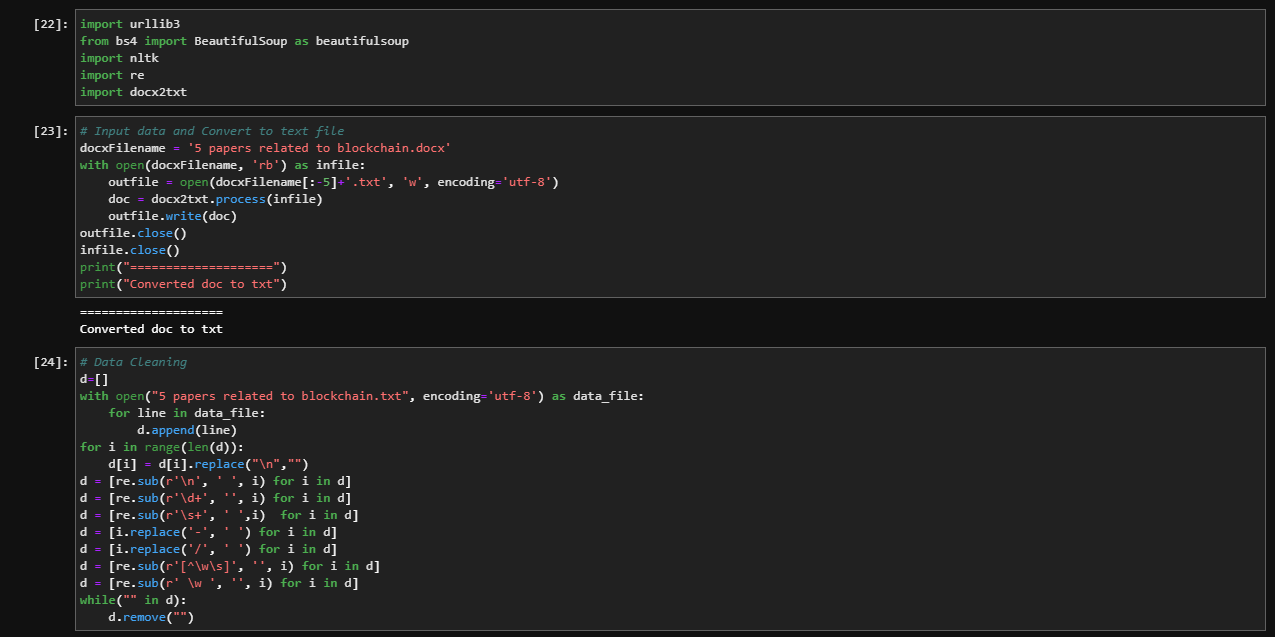
And for 4-gram output would be [birds are excellent indicators] [are excellent indicators of] [excellent indicators of biodiversity] [indicators of biodiversity changes]

**Virtual Environment:**

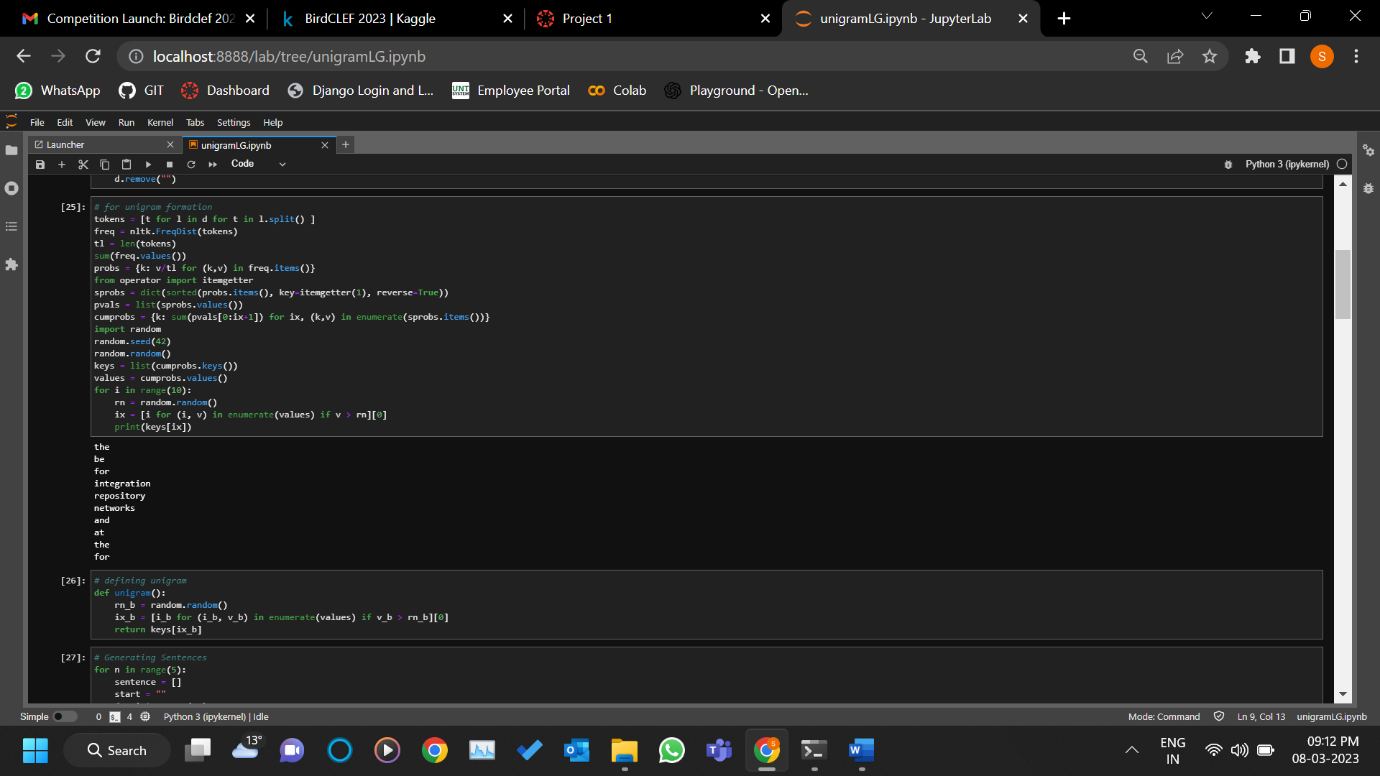
****

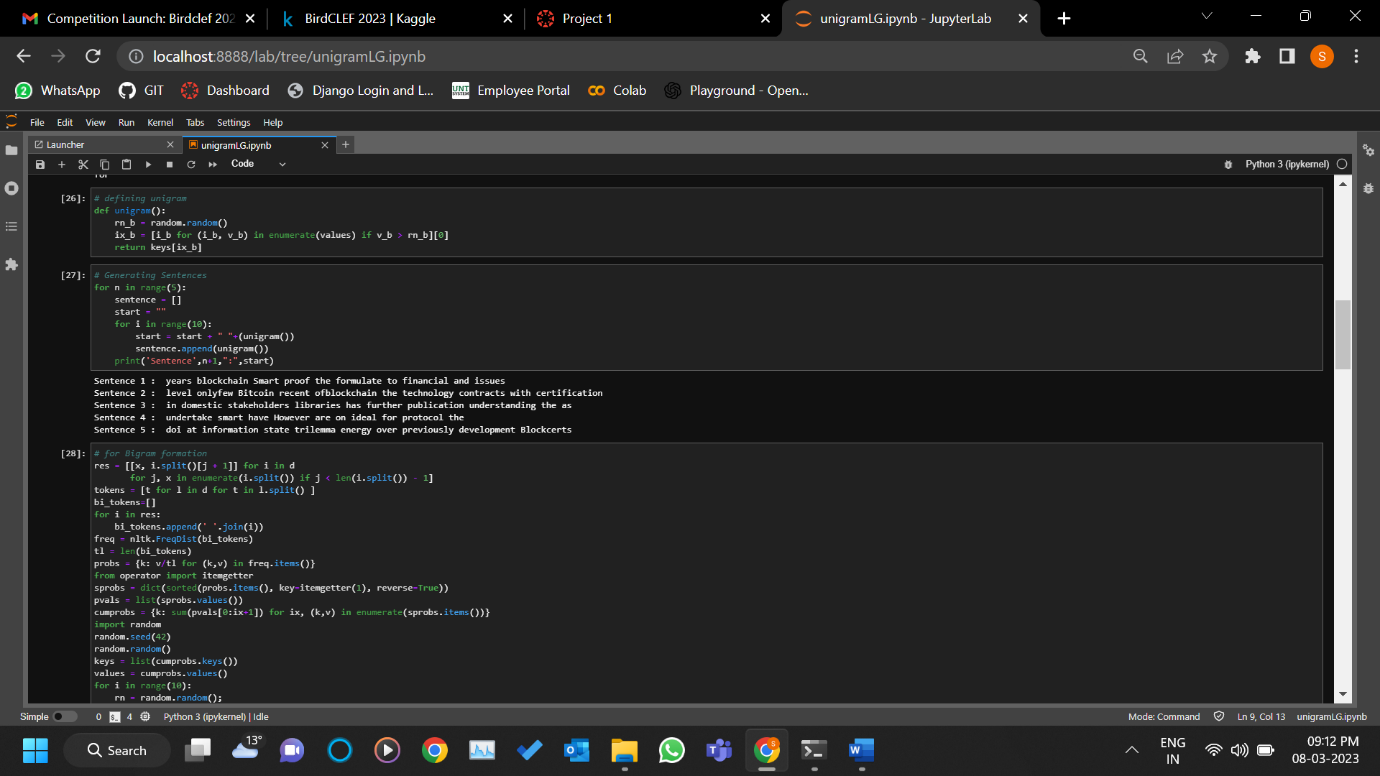


**Here we are importing libraries that are required for the project and then importing doc file and then converting into the text file and cleaning the data**

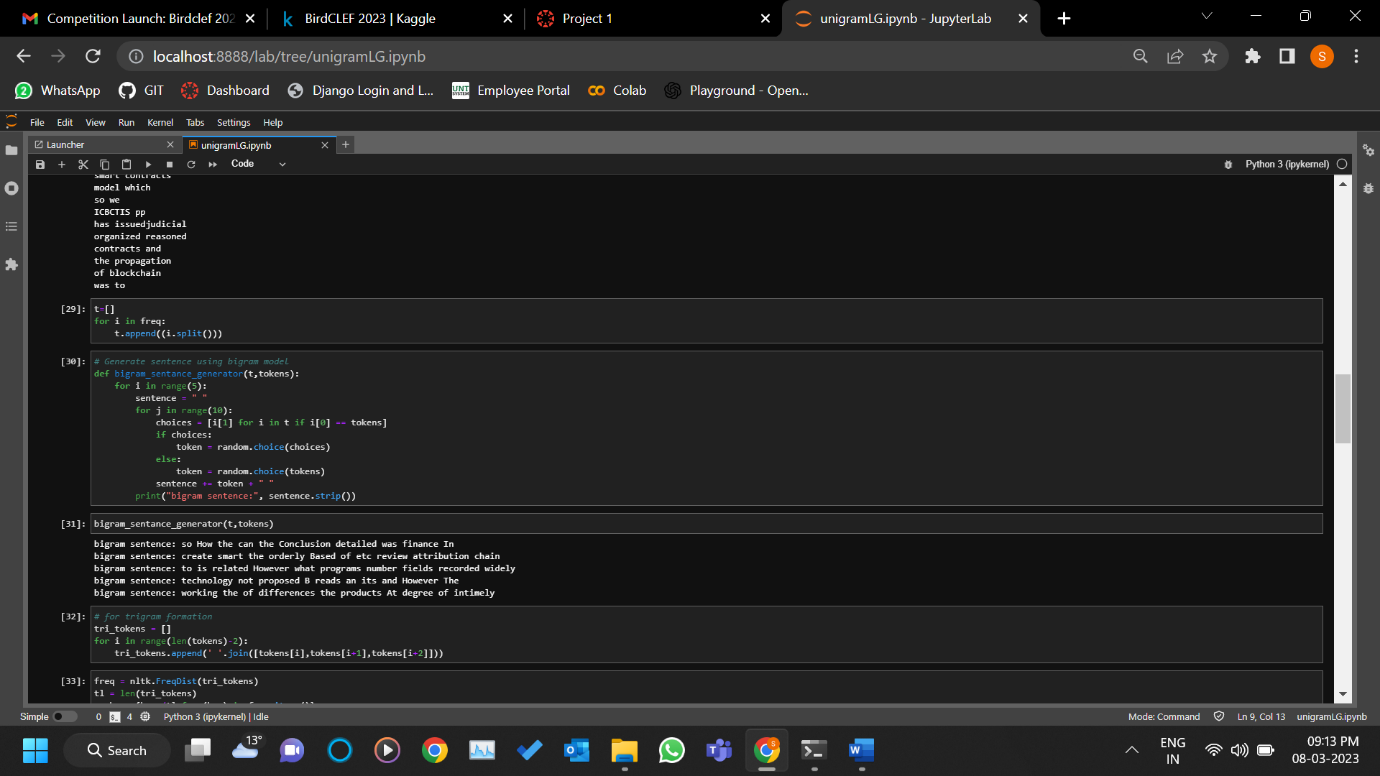
****

**Code for the unigrams and here we are creating unigrams and the generating the sentence for the unigrams**

****

****

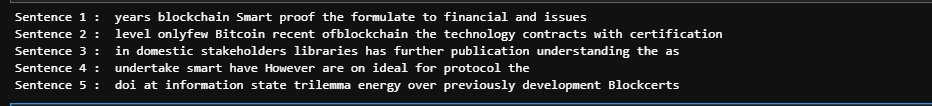
**Code for the bigrams and here we are creating bigrams and the generating the sentence for the bigrams**

****

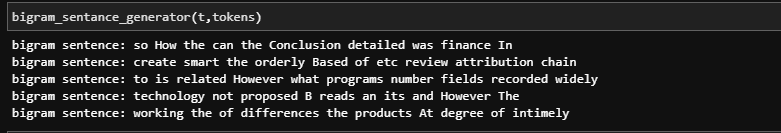
**Code for the trigrams and here we are creating trigrams and the generating the sentence for the trigrams**

**Analysis:**

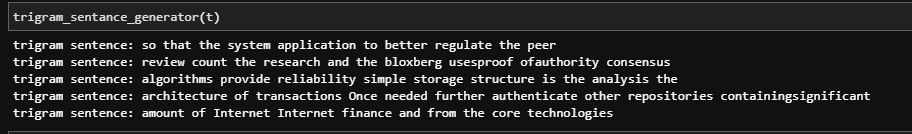
**unigram**

****

**Bigram**

****

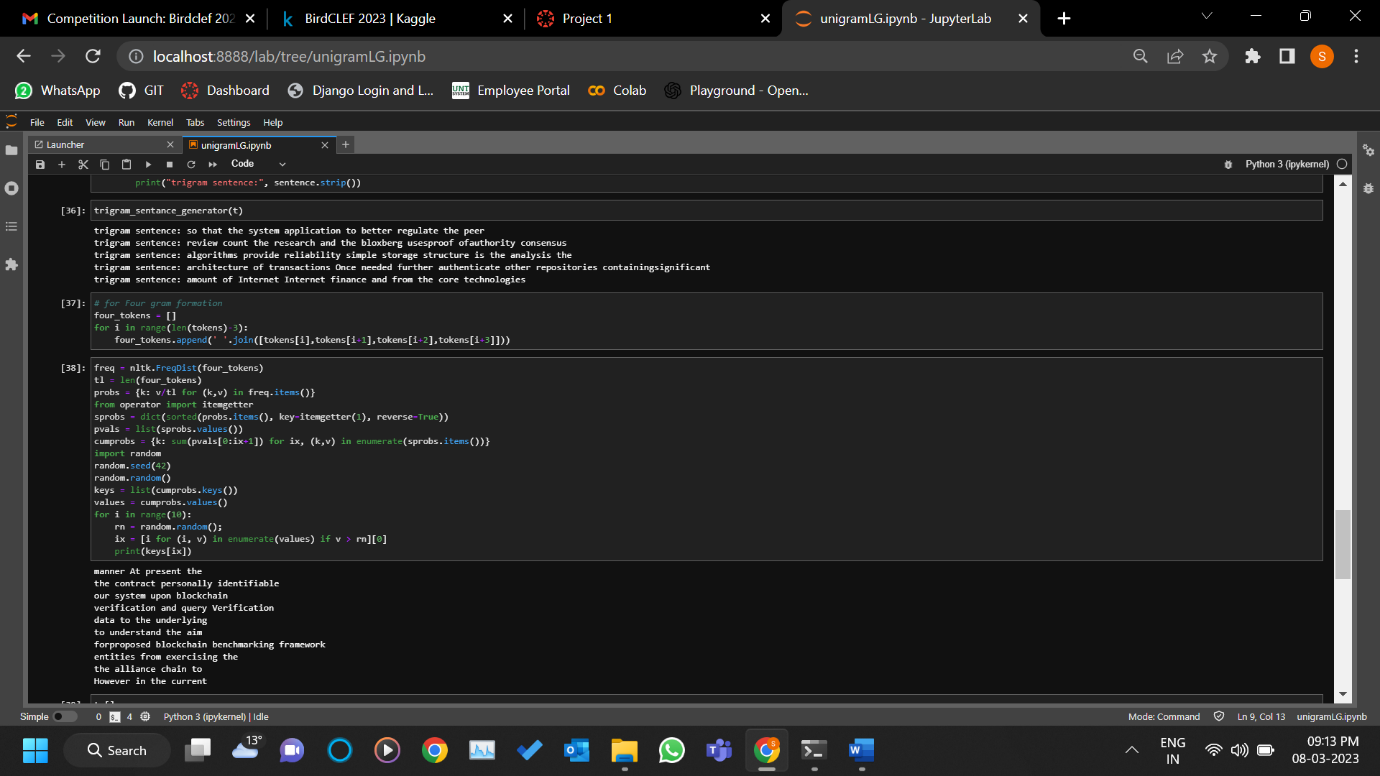
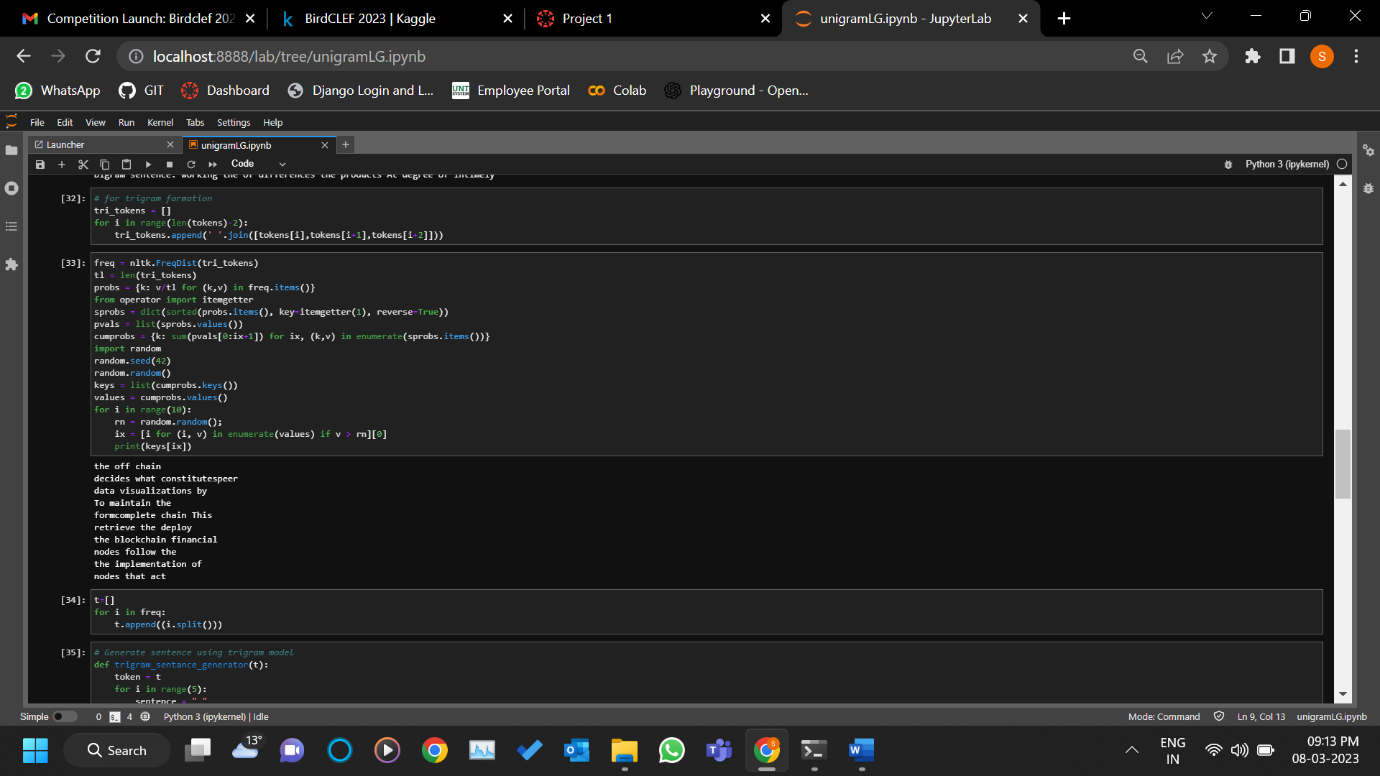
**Trigram**

****

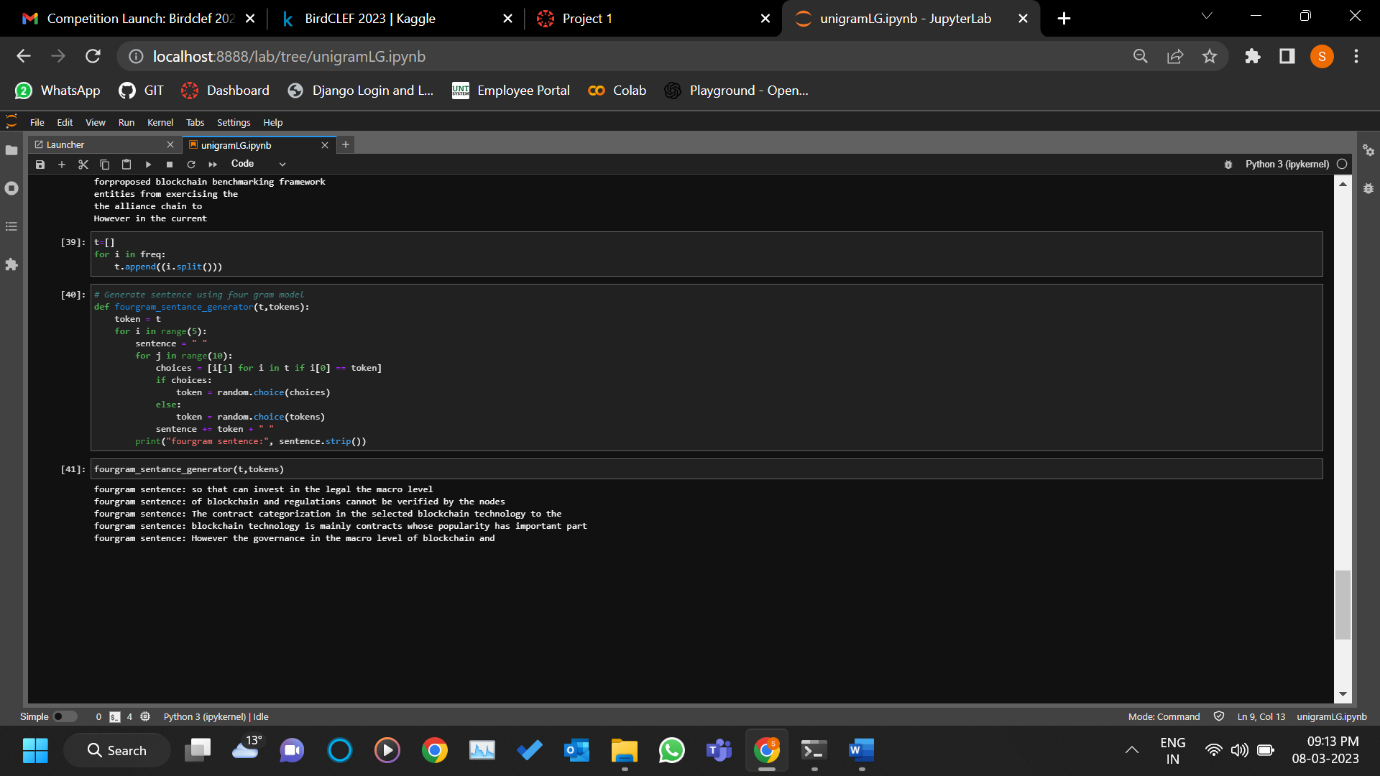
**Four gram:**

****

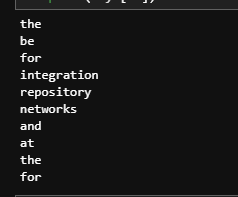
Here we can see that for unigram the sentences are not reasonable, whereas in bigram we can see that there is little relevance in the sentences. In Trigram the relevance is comparatively more whereas in Four gram you can see the sentences is very clear comparative unigram , bigram , trigram and fourgram.

****

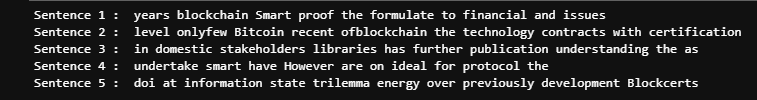
Code for the fourgrams and here we are creating fourgrams and the generating the sentence for the fourgrams



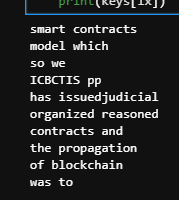
Here we are generating unigrams by calculating the cumulative probability



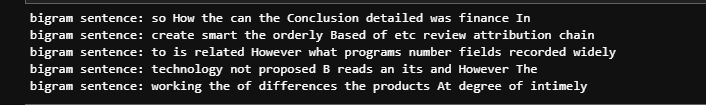
Generating the sentences using unigrams



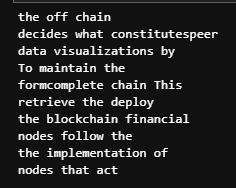
Here we are generating bigrams by calculating the cumulative probability



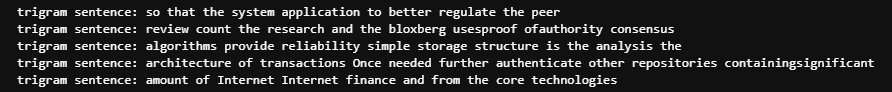
Generating the sentences using bigrams



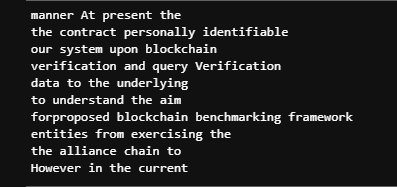
Here we are generating trigrams by calculating the cumulative probability



Generating the sentences using trigrams



Here we are generating bigrams by calculating the cumulative probability



Generating the sentences using four grams

