Assignment - 1 (100 marks)
n-gram Language Model

Fine Print :)

- Try your assignment with a smaller corpus initially.
- Once the output of your functions is consistent with your expectations, you may proceed to extract all the content and perform the tasks to complete every task.
- If you are familiar with functional-style programming, use it.
- Write at least 1-3 lines of comments for every function.
- You may use any JSON library to extract the text from the files.
- You may use NLTK or SpaCy for lemmatization and stemming operations, if required.
- You may use regex libraries to remove unwanted content from the corpus.
- Keep the processed corpus safe. You will need it for all your assignments.
- Share the working code only.
- Python notebooks should be available on the Colab platform (Google).
- Please make sure that all the results are available when you share them. Incomplete python notebooks will not be evaluated.
- Use the email id ramaseshan.ta@gmail.com to share your notebook and provide viewing rights.
- I will not run/change your Python notebook.
- If I do not see any results associated with each assignment, I will NOT evaluate it.
- Use the following naming convention for your file.
  - The first part of the filename should be your Firstname.
  - The second part of the filename should be your roll number.
  - The third part of your assignment should be Assignment_0X, where X is the assignment number.
  - Example - Ramaseshan_XYZ202201_Assignment_01.py
- **Assignments will not be graded if they are sent to my personal email id or to ramaseshan.ta@gmail.com as attachments. Only if you share (using the share option of Colab) the assignments (python notebooks) with the email id (ramaseshan.ta@gmail.com) mentioned above, I will consider them as SUBMITTED**

1 Corpus Creation - 10 marks

You will first create the corpus suitable for building a language model using bigrams and trigrams. The JSON encoded corpus is found. here. The compressed file contains around 56500+ JSON files.

- Create corpus using 50K files.

. Use the following code to read contents from ***body_text*** key.

```
# sample code
import json

def extract_body_text(filename:str) -> str:
    '''
    Note: This function will extract body_text from a
    single file
    '''
    file = open(filename)
    paper_content = json.load(file)
    body_text = ""

    if 'body_text' in paper_content:
        for bt in paper_content['body_text']:
            body_text = body_text + bt['text']
    return ( body_text + '\n').lower()
```

2 Preprocessing - 10 marks

Analyze the corpus carefully. Recommend a set of Preprocessing steps and implement them. Keep the preprocessed file safe. You will need it for other assignments.

3 Find the Vocabulary Count - 10 marks

Find the vocabulary count

4 Bigram and Trigram Language Models - 40 marks

Build bigram and trigram models. Save the models in a folder. Use Laplacian or add-1 smoothing for computing the probability score.

DO NOT use the following code to build the model:

```
def generate_bigrams(filename):
    ....
    for w1, w2 in ngrams(word_tokenize(sentence),2,
            pad_left=True, pad_right=True,
            left_pad_symbol='<s>',
            right_pad_symbol='</s>'):
        model[w1][w2] += 1
        ....
```

Instead use Counter from the collections module to build the models.

```
from collections import Counter
def generate_bigrams(filename):
    bigram_token_freq = Counter()
    with open(filename, encoding='utf8',mode='r') as fd:
        ....
        ....
        ....
            for sentence in sentences:
                bigram_token_freq.update(....)
        .....
        .....
```

Counter from collections module

5 Predicting the missing text - 15 marks

For the following sentences, find the missing word/words. For every missing word, list top ten most_common words with probability. Use all two models to predict the missing words.

all houses were _____ ventilated

it aims to develop an integrated _____ to reach mmps exposed to malaria with prevention diagnosis and treatment _____ by involving non-health _____ stakeholders from provincial to community level

this is because engineers do not work in _____ but rather as a team

6 Perplexity - 15 marks

Find the perplexity score for the following sentences

it appears that the overall code stroke volume has decreased since the covid- pandemic.

half a century ago hypertension was not treatable.

sarahs tv is broadcasting an advert for private healthcare.

Use bigram and trigram models to compute the perplexity score.