AIM: N-gram Language Model

THEORY:

Language modeling is the way of determining the probability of any sequence of words. Language modeling is used in a wide variety of applications such as Speech Recognition, Spam filtering, etc. In fact, language modeling is the key aim behind the implementation of many state-of-the-art Natural Language Processing models.

Methods of Language Modelings:

Two types of Language Modelings:

- **Statistical Language Modelings:** Statistical Language Modeling, or Language Modeling, is the development of probabilistic models that are able to predict the next word in the sequence given the words that precede. Examples such as N-gram language modeling.
- **Neural Language Modelings:** Neural network methods are achieving better results than classical methods both on standalone language models and when models are incorporated into larger models on challenging tasks like speech recognition and machine translation. A way of performing a neural language model is through word embeddings.

N-gram

N-gram can be defined as the contiguous sequence of n items from a given sample of text or speech. The items can be letters, words, or base pairs according to the application. The N-grams typically are collected from a text or speech corpus (A long text dataset).

N-gram Language Model:

An N-gram language model predicts the probability of a given N-gram within any sequence of words in the language. A good N-gram model can predict the next word in the sentence i.e the value of p(w|h)

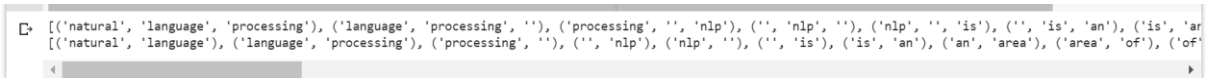
Example of N-gram such as unigram (“This”, “article”, “is”, “on”, “NLP”) or bi-gram (‘This article’, ‘article is’, ‘is on’,’on NLP’).

A) N GRAM FOR ENGLISH:

SOURCE CODE:

```
import re
from nltk.util import ngrams
s = "Natural Language processing (NLP) is an area of computer science and artfical intelligenc
e concerned with the interactions between computer and human (natural) languages."
s = s.lower()
s = re.sub(r'[^a-zA-Z0-9\s]',' ', s)
tokens = [token for token in s.split(" ") if token != " "]

output = list(ngrams(tokens, 3))
print(output)

output = list(ngrams(tokens, 2))
print(output)
```

OUTPUT:



3-GRAM:

[('natural', 'language', 'processing'), ('language', 'processing', ''), ('processing', '', 'nlp'), ('', 'nlp', ''), ('nlp', '', 'is'), ('', 'is', 'an'), ('is', 'an', 'area'), ('an', 'area', 'of'), ('area', 'of', 'computer'), ('of', 'computer', 'science'), ('computer', 'science', 'and'), ('science', 'and', 'artfical'), ('and', 'artfical', 'intelligence'), ('artfical', 'intelligence', 'concerned'), ('intelligence', 'concerned', 'with'), ('concerned', 'with', 'the'), ('with', 'the', 'interactions'), ('the', 'interactions', 'between'), ('interactions', 'between', 'computer'), ('between', 'computer', 'and'), ('computer', 'and', 'human'), ('and', 'human', ''), ('human', '', 'natural'), ('', 'natural', ''), ('natural', '', 'languages'), ('', 'languages', '')]

2-GRAM:

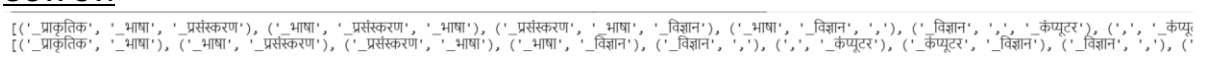
[('natural', 'language'), ('language', 'processing'), ('processing', ''), ('', 'nlp'), ('nlp', ''), ('', 'is'), ('is', 'an'), ('an', 'area'), ('area', 'of'), ('of', 'computer'), ('computer', 'science'), ('science', 'and'), ('and', 'artfical'), ('artfical', 'intelligence'), ('intelligence', 'concerned'), ('concerned', 'with'), ('with', 'the'), ('the', 'interactions'), ('interactions', 'between'), ('between', 'computer'), ('computer', 'and'), ('and', 'human'), ('human', ''), ('', 'natural'), ('natural', ''), ('', 'languages'), ('languages', '')]

B) N-GRAM FOR HINDI:

SOURCE CODE:

```
#!pip install inltk
from inltk.inltk import setup
from inltk.inltk import tokenize
from nltk.util import ngrams

#setup("hi")
hindi_text = "प्राकृतिक भाषा प्रसंस्करण भाषा विज्ञान, कंप्यूटर विज्ञान, और कृत्रिम बुद्धिमत्ता का ए क उपक्षेत्र है, जो कंप्यूटर और मानव भाषा के बीच पारस्परिक क्रियाओं से संबंधित है, विशेष रूप से कं प्यूटर को बड़ी मात्रा में प्राकृतिक भाषा डेटा को संसाधित और विश्लेषण करने के लिए कैसे प्रोग्राम किया जाता है।"

hindiTokens = tokenize(hindi_text, "hi")

hindiOutput = list(ngrams(hindiTokens, 3))
print(hindiOutput)

hindiOutput = list(ngrams(hindiTokens, 2))
print(hindiOutput)
```

OUTPUT:



3-GRAM:

[('_प्राकृतिक', '_भाषा', '_प्रसंस्करण'), ('_भाषा', '_प्रसंस्करण', '_भाषा'), ('_प्रसंस्करण', '_भाषा', '_विज्ञान'), ('_भाषा', '_विज्ञान', ','), ('_विज्ञान', ',', '_कंप्यूटर'), (',', '_कंप्यूटर', '_विज्ञान'), ('_कंप्यूटर', '_विज्ञान', ','), ('_विज्ञान', ',', '_और'), (',', '_और', '_कृत्रिम'), ('_और', '_कृत्रिम', '_बुद्धिमत्ता'), ('_कृत्रिम', '_बुद्धिमत्ता', '_का'), ('_बुद्धिमत्ता', '_का', '_एक'), ('_का', '_एक', '_उपक्षेत्र'), ('_एक', '_उपक्षेत्र', '_है'), ('_उपक्षेत्र', '_है', ','), ('_है', ',', '_जो'), (',', '_जो', '_कंप्यूटर'), ('_जो', '_कंप्यूटर', '_और'), ('_कंप्यूटर', '_और', '_मानव'), ('_और', '_मानव', '_भाषा'), ('_मानव', '_भाषा', '_के'), ('_भाषा', '_के', '_बीच'), ('_के', '_बीच', '_पारस्परिक'), ('_बीच', '_पारस्परिक', '_क्रियाओं'), ('_पारस्परिक', '_क्रियाओं', '_से'), ('_क्रियाओं', '_से', '_संबंधित'), ('_से', '_संबंधित', '_है'), ('_संबंधित', '_है', ','), ('_है', ',', '_विशेष'), (',', '_विशेष', '_रूप'), ('_विशेष', '_रूप', '_से'), ('_रूप', '_से', '_कंप्यूटर'), ('_से', '_कंप्यूटर', '_को'), ('_कंप्यूटर', '_को', '_बड़ी'), ('_को', '_बड़ी', '_मात्रा'), ('_बड़ी', '_मात्रा', '_में'), ('_मात्रा', '_में', '_प्राकृतिक'), ('_में', '_प्राकृतिक', '_भाषा'), ('_प्राकृतिक', '_भाषा', '_डेटा'), ('_भाषा', '_डेटा', '_को'), ('_डेटा', '_को', '_संसाधित'), ('_को', '_संसाधित', '_और'), ('_संसाधित', '_और', '_विश्लेषण'), ('_और', '_विश्लेषण', '_करने'), ('_विश्लेषण', '_करने', '_के'), ('_करने', '_के', '_लिए'), ('_के', '_लिए', '_कैसे'), ('_लिए', '_कैसे', '_प्रोग्राम'), ('_कैसे', '_प्रोग्राम', '_किया'), ('_प्रोग्राम', '_किया', '_जाता'), ('_किया', '_जाता', '_है'), ('_जाता', '_है', '।')]

2-GRAM:

[('_प्राकृतिक', '_भाषा'), ('_भाषा', '_प्रसंस्करण'), ('_प्रसंस्करण', '_भाषा'), ('_भाषा', '_विज्ञान'), ('_विज्ञान', ','), (',', '_कंप्यूटर'), ('_कंप्यूटर', '_विज्ञान'), ('_विज्ञान', ','), (',', '_और'), ('_और', '_कृत्रिम'), ('_कृत्रिम', '_बुद्धिमत्ता'), ('_बुद्धिमत्ता', '_का'), ('_का', '_एक'), ('_एक', '_उपक्षेत्र'), ('_उपक्षेत्र', '_है'), ('_है', ','), (',', '_जो'), ('_जो', '_कंप्यूटर'), ('_कंप्यूटर', '_और'), ('_और', '_मानव'), ('_मानव', '_भाषा'), ('_भाषा', '_के'), ('_के', '_बीच'), ('_बीच', '_पारस्परिक'), ('_पारस्परिक', '_क्रियाओं'), ('_क्रियाओं', '_से'), ('_से', '_संबंधित'), ('_संबंधित', '_है'), ('_है', ','), (',', '_विशेष'), ('_विशेष', '_रूप'), ('_रूप', '_से'), ('_से', '_कंप्यूटर'), ('_कंप्यूटर', '_को'), ('_को', '_बड़ी'), ('_बड़ी', '_मात्रा'), ('_मात्रा', '_में'), ('_में', '_प्राकृतिक'), ('_प्राकृतिक', '_भाषा'), ('_भाषा', '_डेटा'), ('_डेटा', '_को'), ('_को', '_संसाधित'), ('_संसाधित', '_और'), ('_और', '_विश्लेषण'), ('_विश्लेषण', '_करने'), ('_करने', '_के'), ('_के', '_लिए'), ('_लिए', '_कैसे'), ('_कैसे', '_प्रोग्राम'), ('_प्रोग्राम', '_किया'), ('_किया', '_जाता'), ('_जाता', '_है'), ('_है', '।')]

C) <u>N-GRAM FOR MARATHI:</u>

<u>SOURCE CODE:</u>

```
#!pip install inltk
from inltk.inltk import setup
from inltk.inltk import tokenize
from nltk.util import ngrams
```

```
#setup("mr")
marathi_text = "दापोलीतील बंद असलेल्या रिसॉर्टमधून सांडपाणी हे समुद्रात जात असल्याचं कारण देत ईडीने आज आपल्यावर कारवाई केली, यामध्ये मनी लॉड्रिंगचा संबंध नाही अशी माहिती राज्याचे मंत्री अनिल परब यांनी दिली आहे. दापोलीतील रिसॉर्ट हे सदानंद कदम यांच्या मालकीचे असून त्याच्याशी माझा काही संबंध नाही असंही अनिल परब म्हणाले."

marathiTokens = tokenize(marathi_text, "mr")

marathiOutput = list(ngrams(marathiTokens, 3))
print(marathiOutput)

marathiOutput = list(ngrams(marathiTokens, 2))
print(marathiOutput)
```

OUTPUT:

```
[('_दापोली', 'तील', '_बंद'), ('तील', '_बंद', '_असलेल्या'), ('_बंद', '_असलेल्या', '_रिसॉर्ट'), ('_असलेल्या', '_रिसॉर्ट', 'मधून'), ('_रिसॉर्ट', 'मधून', '_सांडपाणी'), ('मधून', '_सांडप
[('_दापोली', 'तील'), ('तील', '_बंद'), ('_बंद', '_असलेल्या'), ('_असलेल्या', '_रिसॉर्ट'), ('_रिसॉर्ट', 'मधून'), ('मधून', '_सांडपाणी'), ('_सांडपाणी', '_हे'), ('_हे', '_समुद्रात'), ('_
```

3-GRAM:

[('_दापोली', 'तील', '_बंद'), ('तील', '_बंद', '_असलेल्या'), ('_बंद', '_असलेल्या', '_रिसॉर्ट'), ('_असलेल्या', '_रिसॉर्ट', 'मधून'), ('_रिसॉर्ट', 'मधून', '_सांडपाणी'), ('मधून', '_सांडपाणी', '_हे'), ('_सांडपाणी', '_हे', '_समुद्रात'), ('_हे', '_समुद्रात', '_जात'), ('_समुद्रात', '_जात', '_असल्या'), ('_जात', '_असल्या', 'चं'), ('_असल्या', 'चं', '_कारण'), ('चं', '_कारण', '_देत'), ('_कारण', '_देत', '_ई'), ('_देत', '_ई', 'डी'), ('_ई', 'डी', 'ने'), ('डी', 'ने', '_आज'), ('ने', '_आज', '_आपल्या'), ('_आज', '_आपल्या', 'वर'), ('_आपल्या', 'वर', '_कारवाई'), ('वर', '_कारवाई', '_केली'), ('_कारवाई', '_केली', ','), ('_केली', ',', '_यामध्ये'), (',', '_यामध्ये', '_मनी'), ('_यामध्ये', '_मनी', '_लॉ'), ('_मनी', '_लॉ', 'ं'), ('_लॉ', 'ं', 'ड्रि'), ('ं', 'ड्रि', 'ंग'), ('ड्रि', 'ंग', 'चा'), ('ंग', 'चा', '_संबंध'), ('चा', '_संबंध', '_नाही'), ('_संबंध', '_नाही', '_अशी'), ('_नाही', '_अशी', '_माहिती'), ('_अशी', '_माहिती', '_राज्य'), ('_माहिती', '_राज्य', 'ाचे'), ('_राज्य', 'ाचे', '_मंत्री'), ('ाचे', '_मंत्री', '_अनिल'), ('_मंत्री', '_अनिल', '_पर'), ('_अनिल', '_पर', 'ब'), ('_पर', 'ब', '_यांनी'), ('ब', '_यांनी', '_दिली'), ('_यांनी', '_दिली', '_आहे'), ('_दिली', '_आहे', '.'), ('_आहे', '.', '_दापोली'), ('.', '_दापोली', 'तील'), ('_दापोली', 'तील', '_रिसॉर्ट'), ('तील', '_रिसॉर्ट', '_हे'), ('_रिसॉर्ट', '_हे', '_सदानंद'), ('_हे', '_सदानंद', '_कदम'), ('_सदानंद', '_कदम', '_यांच्या'), ('_कदम', '_यांच्या', '_मालकीचे'), ('_यांच्या', '_मालकीचे', '_असून'), ('_मालकीचे', '_असून', '_त्याच्याशी'), ('_असून', '_त्याच्याशी', '_माझा'), ('_त्याच्याशी', '_माझा', '_काही'), ('_माझा', '_काही', '_संबंध'), ('_काही', '_संबंध', '_नाही'), ('_संबंध', '_नाही', '_असं'), ('_नाही', '_असं', 'ही'), ('_असं', 'ही', '_अनिल'), ('ही', '_अनिल', '_पर'), ('_अनिल', '_पर', 'ब'), ('_पर', 'ब', '_म्हणाले'), ('ब', '_म्हणाले', '.')]

2-GRAM:

[('_दापोली', 'तील'), ('तील', '_बंद'), ('_बंद', '_असलेल्या'), ('_असलेल्या', '_रिसॉर्ट'), ('_रिसॉर्ट', 'मधून'), ('मधून', '_सांडपाणी'), ('_सांडपाणी', '_हे'), ('_हे', '_समुद्रात'), ('_समुद्रात', '_जात'),

('_जात', '_असल्या'), ('_असल्या', 'चं'), ('चं', '_कारण'), ('_कारण', '_देत'), ('_देत', '_ई'), ('_ई', 'डी'), ('डी', 'ने'), ('ने', '_आज'), ('_आज', '_आपल्या'), ('_आपल्या', 'वर'), ('वर', '_कारवाई'), ('_कारवाई', '_केली'), ('_केली', ','), (',', '_यामध्ये'), ('_यामध्ये', '_मनी'), ('_मनी', '_लॉ'), ('_लॉ', 'ं'), ('ं', 'ड्रि'), ('ड्रि', 'ंग'), ('ंग', 'चा'), ('चा', '_संबंध'), ('_संबंध', '_नाही'), ('_नाही', '_अशी'), ('_अशी', '_माहिती'), ('_माहिती', '_राज्य'), ('_राज्य', 'ाचे'), ('ाचे', '_मंत्री'), ('_मंत्री', '_अनिल'), ('_अनिल', '_पर'), ('_पर', 'ब'), ('ब', '_यांनी'), ('_यांनी', '_दिली'), ('_दिली', '_आहे'), ('_आहे', '.'), ('.', '_दापोली'), ('_दापोली', 'तील'), ('तील', '_रिसॉर्ट'), ('_रिसॉर्ट', '_हे'), ('_हे', '_सदानंद'), ('_सदानंद', '_कदम'), ('_कदम', '_यांच्या'), ('_यांच्या', '_मालकीचे'), ('_मालकीचे', '_असून'), ('_असून', '_त्याच्याशी'), ('_त्याच्याशी', '_माझा'), ('_माझा', '_काही'), ('_काही', '_संबंध'), ('_संबंध', '_नाही'), ('_नाही', '_असं'), ('_असं', 'ही'), ('ही', '_अनिल'), ('_अनिल', '_पर'), ('_पर', 'ब'), ('ब', '_म्हणाले'), ('_म्हणाले', '.')]

CONCLUSION:

From this practical, I have learned how to implement the n-gram language model for English and other Indian regional languages.