Assignment 3
Natural Language Processing

Kanika Saini (2016047)

Your task in this assignment is to write a python program that will be able to generate and classify sentences based on some corpus .

Generative Model

- **Comp.graphics**
    - **Unigram**
        - Unigram sentence generation. (Choosing by max frequency)
            - the a I to of and it in you is that for on my Apr with From .
        - Unigram sentence generation. (Choosing randomly)
            - 460-8302 bars/clip-ons Wrangler phone BACKGROUND .
    - **Bigram**
        - How to the bike . I 'm not be a few
        - Consider that the bike . I 'm not be a few weeks I was a bike , and the
    - **Trigram**
        - I have a backup helmet (XL) , and I 'm not sure that if I could do ... Newsgroups :
        - You are wrong . As far as I 've never had a friend shopping for her first motorcycle . I 've never.
- **Rec.motorcycles**
    - **Unigram**
        - Unigram sentence generation. (Choosing by max frequency)
            - the to a of and I is for in it you that on be with or have are .
        - Unigram sentence generation. (Choosing randomly)
            - submit 13:15:56 show Multiplot assign .
    - **Bigram**
        - How do n't have a few posts a lot of the same as a good choice is a program . I have to the
        - Consider a few posts a lot of the same as a good choice
    - **Trigram**

- I have a copy of the above programs . Contact : Bill Johnston , (415) 924-8640 , (415)
- You are the same as the Usenet-standard JFIF format . The current version is 2.1 , available from Simtel20 and mirror sites.

Assumptions

- The corpus is not cleaned. Used as given.
- Word tokenizing is done via NLTK. Sentence tokenizing is avoided.
- In unigram, two implementations are there:
    - Max probability
    - Random selection
    - If else added to avoid excess punctuations
- In bigram and trigram, next max probability word is chosen and then deleted from the list to avoid repetitions.
- In bigram and trigram, starting word is given by the user.
- Discriminative model - add-one smoothing is used.
- <UNK> tag replaces words with frequency <= 2 in training data. Add-one smoothin is used.