Ministry of Education and Science of the Russian Federation
Peter the Great St. Petersburg State Polytechnic University
Institute of Computer Sciences and Technologies
Graduate School of Cyber-Physical Systems and Control

Practice Task – Ch 10

Bigram/Trigram Model (Markov Model) – Text Generator
Discipline: Intellectual Computing
4 April 2017

Student Group: 13541/8

Christopher W. Blake

Professor

Kuchmin A.Y

St. Petersburg
2017

Contents

Introduction

Chapter 10 of "AI Application Programming" by M. Tim Jones is about Markov models, specifically the bigram/trigram models, and their use in a text generator. A Markov model is created to represent the probabilistic relationship between words. The model provides the percentage possibility of a next word in a sentence. Using this likelihood, words from existing text are classified as "start", "middle", and "end" words. The number of occurrences of word pairs/triplets is also recorded. After having formed this model, a random start word is selected, and next words are predicted until an end word is found. Using this process a new sentence, which is statistically similar to the existing text, is created.

A sample C# program has been created to demonstrate this text generation system. The user simply needs to select a text file with sample sentences, and the bigram and trigram models are generated. The bigram or trigram model may also be selected. At this point, the user may press the "Generate" button to create new text.

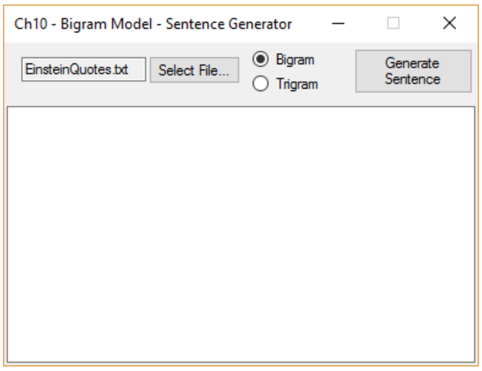


Figure 1: C# Sample Program

Background

Word Relationship

All words in the text file are compared in a pair-wise (bigram) or triplet-wise (trigram) fashion. Every combination of two/three words is added to a list of connections. The number of occurrences of this word pair is additionally counted and later used to determine probabilities. An example of this can been seen in figure 2, with the word "dial" and "delete". "Dial" has a 70% chance of having the word "number" after it. "Delete" has an 80% chance of the word "number" after it.

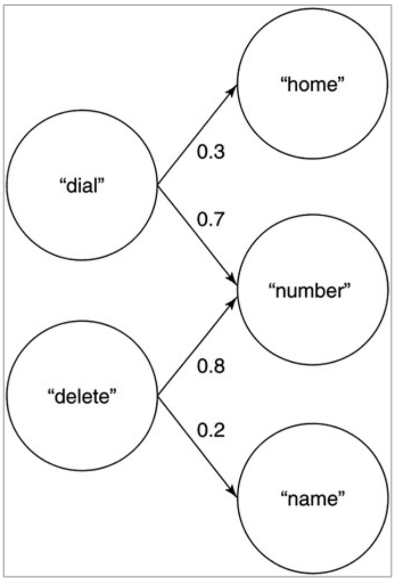


Figure 2: Word Pair Probability

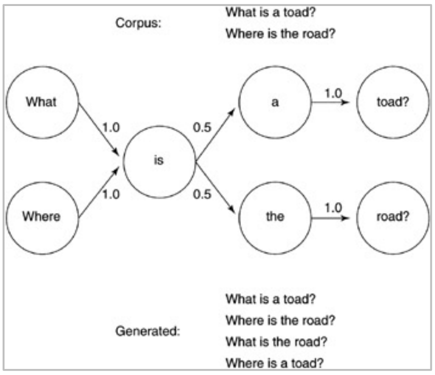
Word Categories

Each new word is added to a list of available words. It is also recorded where, in the sentence, this word was found. This is simply categorized into "Start", "Middle", and "End". Using the number of occurrences in each position, a word has a percentage degree to each category.

Algorithm

The below steps describes the general worfklow of building a sentence, from selecting a start word, to adding content, to ending the sentence. (See figure 3.)

1.) Select a random start word.
2.) Chose the next word
    a. Create a list of words with occurences greater than zero.



What is a toad?
Where is the road?

What is a toad?
Where is the road?
What is the road?
Where is a toad?

Figure 3: Word Selection Process

    b. Randomly pick a word.
    c. Generate a random value (0 to 1).
    d. Compare word probability to random number.
        i. If it passes, chose this word.
        ii. If it failes, return to step b.

3.) Add word to sentence.

4.) Check if word is of category “End”.

    a. Generate a random value (0 to 1).
    b. Compare percentage of “End” truthness to random number.
    c. If it failes, return to step 2. Do not end the sentence.
    d. If it passes, continue to step 5. End the sentence.

5.) Convert list of words to a string and present to the user.

Results

The following sentence were generated using a sample text file with 108 Einstein quotations, which resulted in 1830 bigram connections and 1910 trigram connections. The bigram model produced more interesting results which were sometimes difficult to read. The trigram model often produced readable sentences but it also had the tendency to reproduce existing sentences. Below are example generated sentences, which were thought provoking.

Bigram

```
1.  Insanity: doing it if it is more violent.
2.  Knowledge relates to overcoming the modernist's snobbishness.
3.  Bear in the shifting sand.
4.  Heroism on from the facts.
5.  And as it has been given a soul, everything that created it has been preserved,
    given superstitions.
6.  Look deep into your difficulties in the kiss the door directly to go away for truth
    leave elegance to overcome man's insecurity before me like an incorrigible
    nonconformist warmly acclaimed.
7.  Great pleasure indeed.
8.  Anger dwells only reads too much and artistic powers to stop questioning.
```

Trigram

```
1.  Schools, the collection of prejudices acquired by age eighteen.
2.  Common is a somewhat new kind of religion.
3.  Look into your hands as your inheritance in order to be lost.
```

Conclusion

A Markov network of nodes was used to successfully categorize sample sentences into bigram and trigram networks. Using these word pairs and triplets, new sentences similar to the original content were created. The bigram model produced difficult to read (but unique) results. The trigram model produced easier-to-read content but had the tendency to reproduce existing sentences.