Lab2

EMPIRICAL LAWS

Date: 02-03-2021

Objective

The objective this lab is to implement and analyse various empirical laws discussed in the theory class. We will implement the following statistics from various text corpus

- Type to token ratio
- Zipf's law
- Heap's law
- Automatic readability index

Tools

- NLTK
- Matplotlib

Instructions

- Use Jupyter/Colab notebook for implementation
- Write your observations in the notebook as markdown text
- Convert the notebook to PDF
- Submit the PDF before due time.

Steps

Type to token ratio

1. Take 5 corpus from NLTK: shakespeare-caesar.txt', 'shakespeare-hamlet.txt', 'shakespeare-macbeth.txt'
2. Write a function find TTR
    a. For the window of 1000 words

        i. Find the list of tokens (use tokenize function)
        ii. Find the unique tokens
        iii. Compute TTR
    b. Return the moving average
3. Find TTR for each document

Zipf's Law

1. Select a corpus (say 'bible-kjv.txt') from NLTK
2. Find the unique tokens
3. Find the frequency of each token
4. Sort the tokens in the decreasing order of frequency
5. Plot rank vs frequency (use log(rank) and log(freq))
6. Give your observations

Heap's Law

1. Select a corpus from NLTK
2. Find number of types for each window 100 tokens
3. Plot the statistics (number of tokens vs types)
4. Give your observations

ARI

1. Take a random text
2. Use NLTK functions for word and sentence segmentation
3. Compute ARI =

$$4.71\left(\frac{\text{characters}}{\text{words}}\right)+0.5\left(\frac{\text{words}}{\text{sentences}}\right)-21.43$$

4. Check the appropriate age group for the text

Score	Age	Grade Level
1	5-6	Kindergarten
2	6-7	First/Second Grade
3	7-9	Third Grade
4	9-10	Fourth Grade
5	10-11	Fifth Grade
6	11-12	Sixth Grade
7	12-13	Seventh Grade
8	13-14	Eighth Grade
9	14-15	Ninth Grade
10	15-16	Tenth Grade
11	16-17	Eleventh Grade
12	17-18	Twelfth grade
13	18-24	College student
14	24+	Professor

References

- Lectures 5&6
- NLP | How tokenizing text, sentence, words works