PROJEKTOWANIE SYSTEMÓW STEROWANIA (E:35379W0)

POLITECHNIKA GDAŃSKA

Adaptive Distribution of Vocabulary Frequencies

Project Report

Supervisor: Mr. Karol Szymanski

By: John Bounds Christopher 154001

BACKGROUND

Zipf's law can be considered similar to Italian economist Vilfredo Pareto's principle, who observed and showed that 20% of Italy's population owned 80% of its land. This Pareto distribution can also be observed in many natural phenomenons as well.

Similarly, in this report we can see using Zipf's law that, humans on an average use 20% of the total vocabulary in any given language which accounts for 80% of their conversations.

In Software: This principle can be used for optimization efforts. Fixing 20% of top bugs can eliminate 80% of errors.

MYSTERY OF ZIPF'S LAW

Zipf's law was analysed by linguist George Kingsley Zipf, who theorised that given a large language corpus, the frequency of each word is approximately equal to inverse of its rank in the frequency table. That is:

$$P_n \; \alpha \; 1/(n^a)$$

where a is almost equal to 1. This is known as a "power law" and we can observe that the most frequent word will occur approximately twice as often as the second most frequent word. "The" is the most frequently occurring word (accounting for nearly 7% of all word occurrences — 69,971 out of slightly over 1 million), and "of" the second most frequent (3.5% of all words).

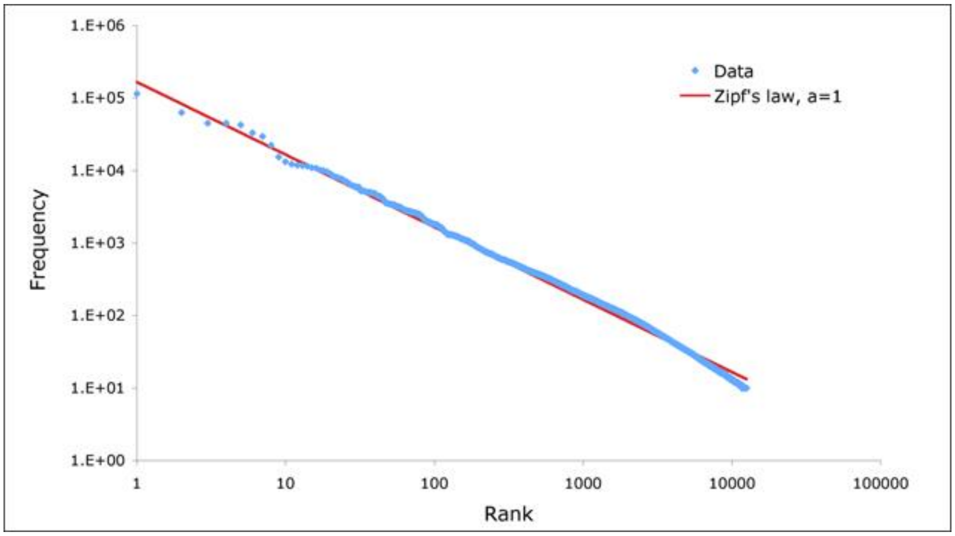


Fig 1: Zipf's Law

SUMMARY

Reference IEEE paper: Adaptive Distribution of Vocabulary

Frequencies: a novel estimation suitable for social media corpus

“2014 Brazilian Conference on Intelligent Systems”

Rodrigo Augusto Igawa, Guilherme Sakaji Kido, Jose Luis Seixas Jr., Sylvio Barbon Jr

Abstract – In this project, we try to introduce a mathematical model for distribution of vocabulary frequency on a social media platform and develop an adaptive model that is more sensitive to text noises rather than the zipf's law distribution so that it is asymptotically quicker.

LIST OF FIGURES

AIM

The aim of this project is to verify the author's approach and confirm the Zipf's distribution across a random language corpus on social media.

OBJECTIVE

- To access a large existing twitter corpus / dataset and pre-processing by lower-casing and removing hash-tags

- Use Python's data-science libraries for this purpose

- Split the data into a data-frame and write a word-count program

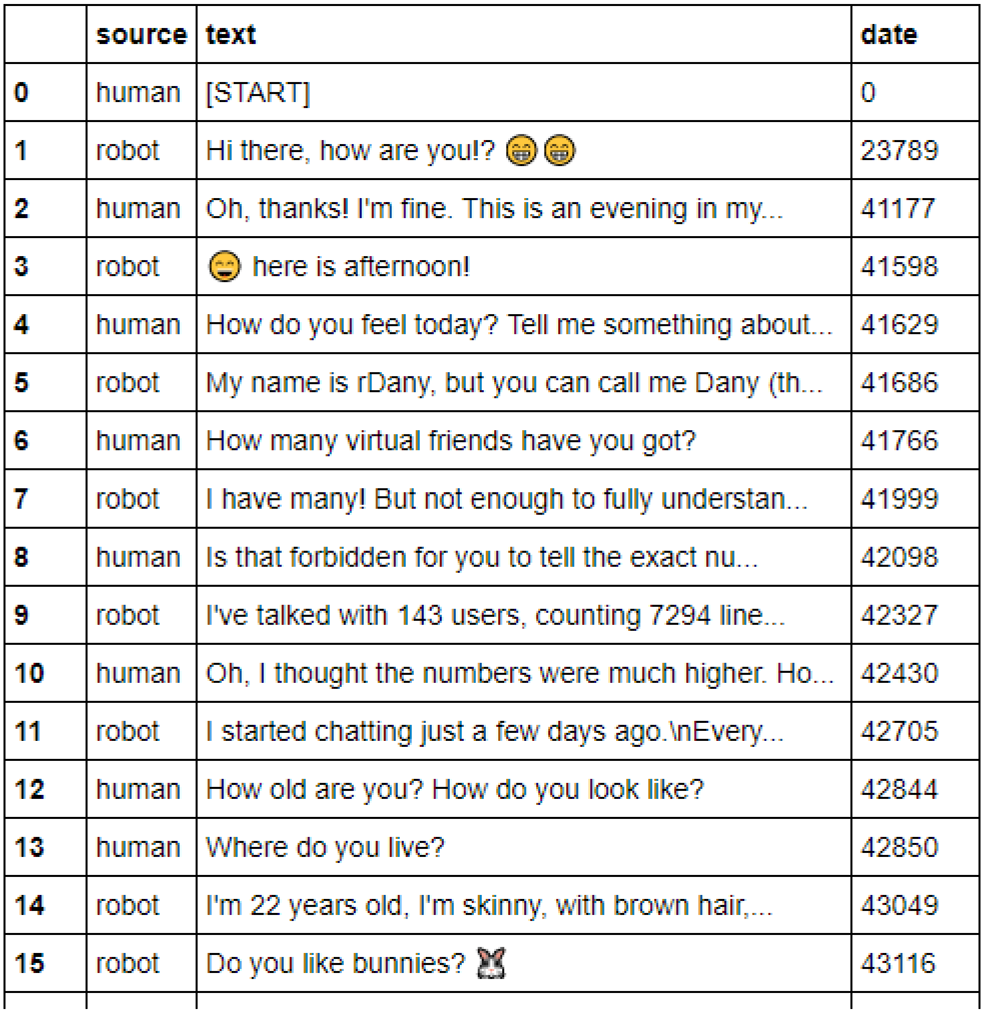
- Rank the words in descending order alongside its rank

- Plot the word rank against its frequency on the X-axis and plot the adaptive distribution curve as well using seaborn visualization tool

TOOLS AND LIBRARIES USED

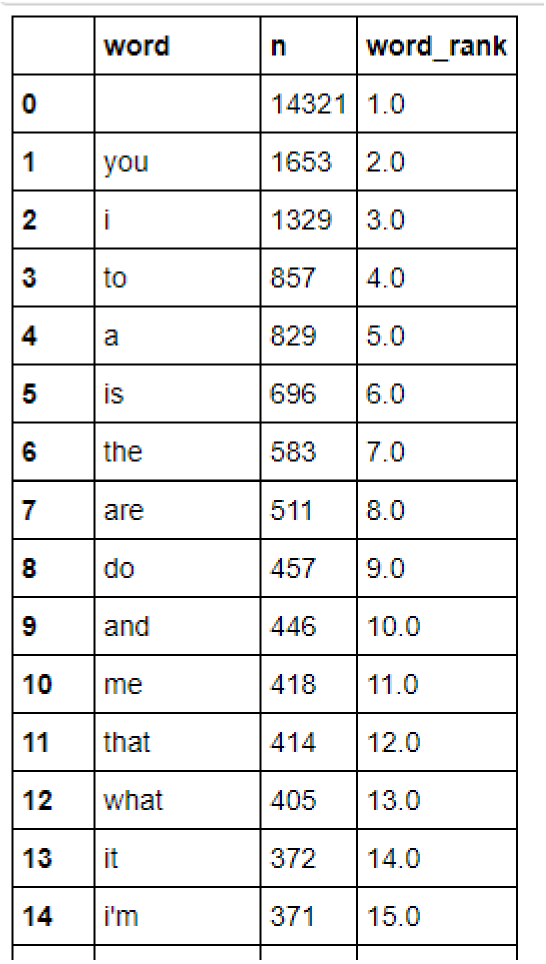
- Jupyter notebook using Python
- Numpy
- Pandas
- Regular expressions
- Seaborn

RESULTS



	source	text	date
0	human	[START]	0
1	robot	Hi there, how are you!? 😁😁	23789
2	human	Oh, thanks! I'm fine. This is an evening in my...	41177
3	robot	😄 here is afternoon!	41598
4	human	How do you feel today? Tell me something about...	41629
5	robot	My name is rDany, but you can call me Dany (th...	41686
6	human	How many virtual friends have you got?	41766
7	robot	I have many! But not enough to fully understan...	41999
8	human	Is that forbidden for you to tell the exact nu...	42098
9	robot	I've talked with 143 users, counting 7294 line...	42327
10	human	Oh, I thought the numbers were much higher. Ho...	42430
11	robot	I started chatting just a few days ago.\nEvery...	42705
12	human	How old are you? How do you look like?	42844
13	human	Where do you live?	42850
14	robot	I'm 22 years old, I'm skinny, with brown hair,...	43049
15	robot	Do you like bunnies? 🐰	43116

Fig 2: Twitter data-set



	word	n	word_rank
0		14321	1.0
1	you	1653	2.0
2	i	1329	3.0
3	to	857	4.0
4	a	829	5.0
5	is	696	6.0
6	the	583	7.0
7	are	511	8.0
8	do	457	9.0
9	and	446	10.0
10	me	418	11.0
11	that	414	12.0
12	what	405	13.0
13	it	372	14.0
14	i'm	371	15.0

Fig 3: Pre-processed and ranked data-set

DISCUSSIONS

The expected word distribution was not found on this data-set. This is because people tend to use more informal use of vocabulary and also resort to shorter forms of longer words.

The power law distribution which was found to predict ‘the’ and ‘of’ as the first two more frequent words used, fails to bring about the same distribution in the social media platforms as people tend to be more informal.

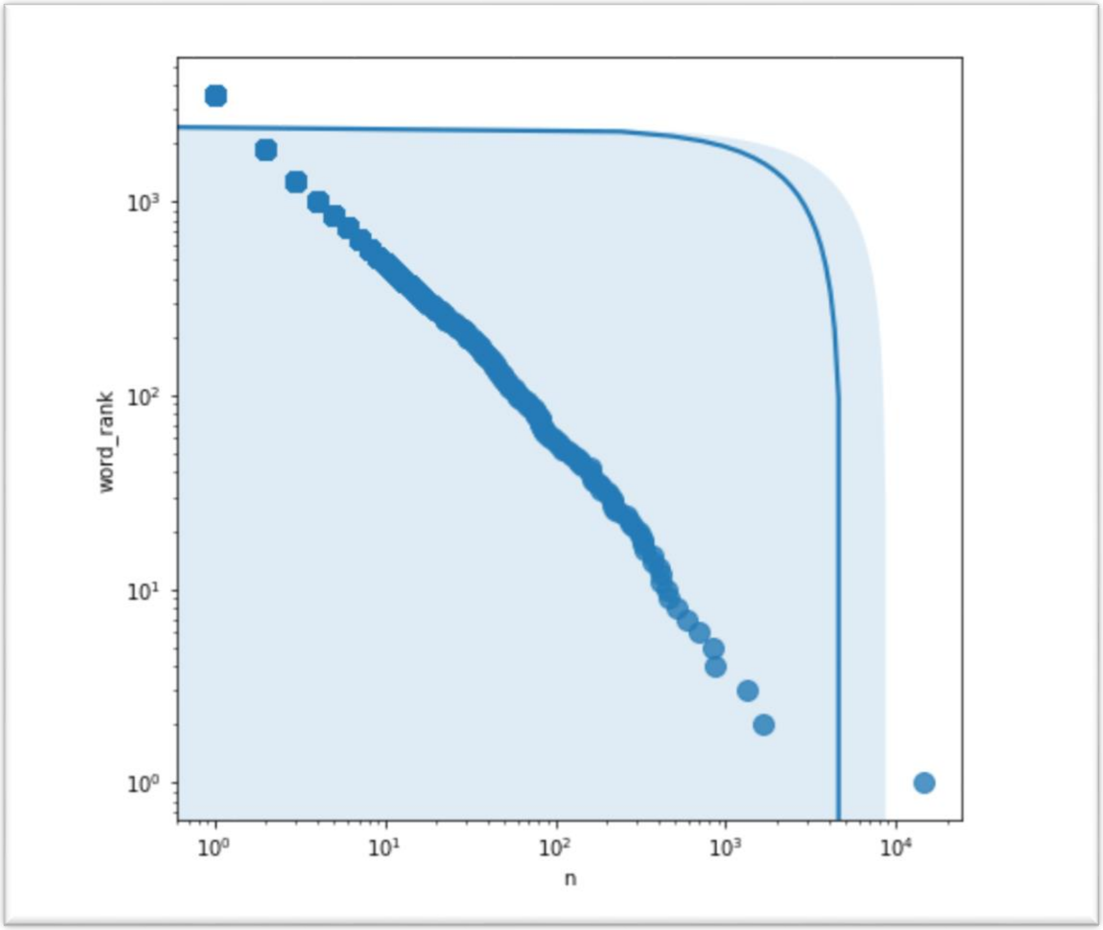


Fig 4: Visualization of pre-processed data-set

CONCLUSION

It can be argued that man is lethargic in trying to use a more uncommon word to express his real emotion / intent and would prefer a more common word, which may not describe his intention so accurately.

This goes on to show that man would use 20% of more commonly used and versatile words in 80% of his conversation as this is more productive for him.

APPENDICES

Adding libraries

```
import matplotlib.pyplot as plt

import numpy as np

import pandas as pd

import re

import seaborn as sns
```

Reading the dataframe

```
df = pd.read_csv('C:\\Users\\HP\\rdany_conversations_2016-03-01.csv')
```

Calling required columns of the dataframe

```
df = df[['source', 'text', 'date']]
```

Removing emoticons

```
df['emoji_count'] = df.text.apply(lambda x: len(re.findall(r'[\U0001f600-\U0001f650]', x)))

df['clean_text'] = df.text.apply(lambda x: re.sub('[^A-Za-z\']', ' ', x.lower()))
```

Writing the word count program

```
word_list = ' '.join(df.clean_text.values).split(' ')

words = pd.DataFrame(word_list, columns=['word'])

word_counts = words.word.value_counts().reset_index()

word_counts.columns = ['word', 'n']

word_counts['word_rank'] = word_counts.n.rank(ascending=False)
```

Initializing the visualization libraries

```
%matplotlib inline
```

```
ff,,  axax  ==  pltplt..subplotssubplots(figsize=(7, 7))

ax.set(xscale="log", yscale="log")

sns.regplot("n", "word_rank", word_counts, ax=ax, scatter_kws={"s": 100})
```