PROJEKTOWANIE SYSTEMÓW STEROWANIA (E:35379W0)

Adaptive Distribution of Vocabulary Frequencies

Project Presentation

Supervisor: Mr. Karol Szymanski

By: John Bounds Christopher 154001

BACKGROUND

- Italian economist Vilfredo Pareto’s principle observed that 20% of Italy's population owned 80% of its land
- Observed in many natural phenomenons as well
- 20% of the total vocabulary in any given language accounts for 80% of their conversations

- **In Software:** used for optimization efforts. Fixing 20% of top bugs can eliminate 80% of errors

MYSTERY OF ZIPF'S LAW

- Linguist George Kingsley observed that the frequency of each word is approximately equal to inverse of its rank in the frequency table. That is:

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

- "The" is the most frequently occurring word (accounting for nearly 7% of all word occurrences — 69,971 out of slightly over 1 million

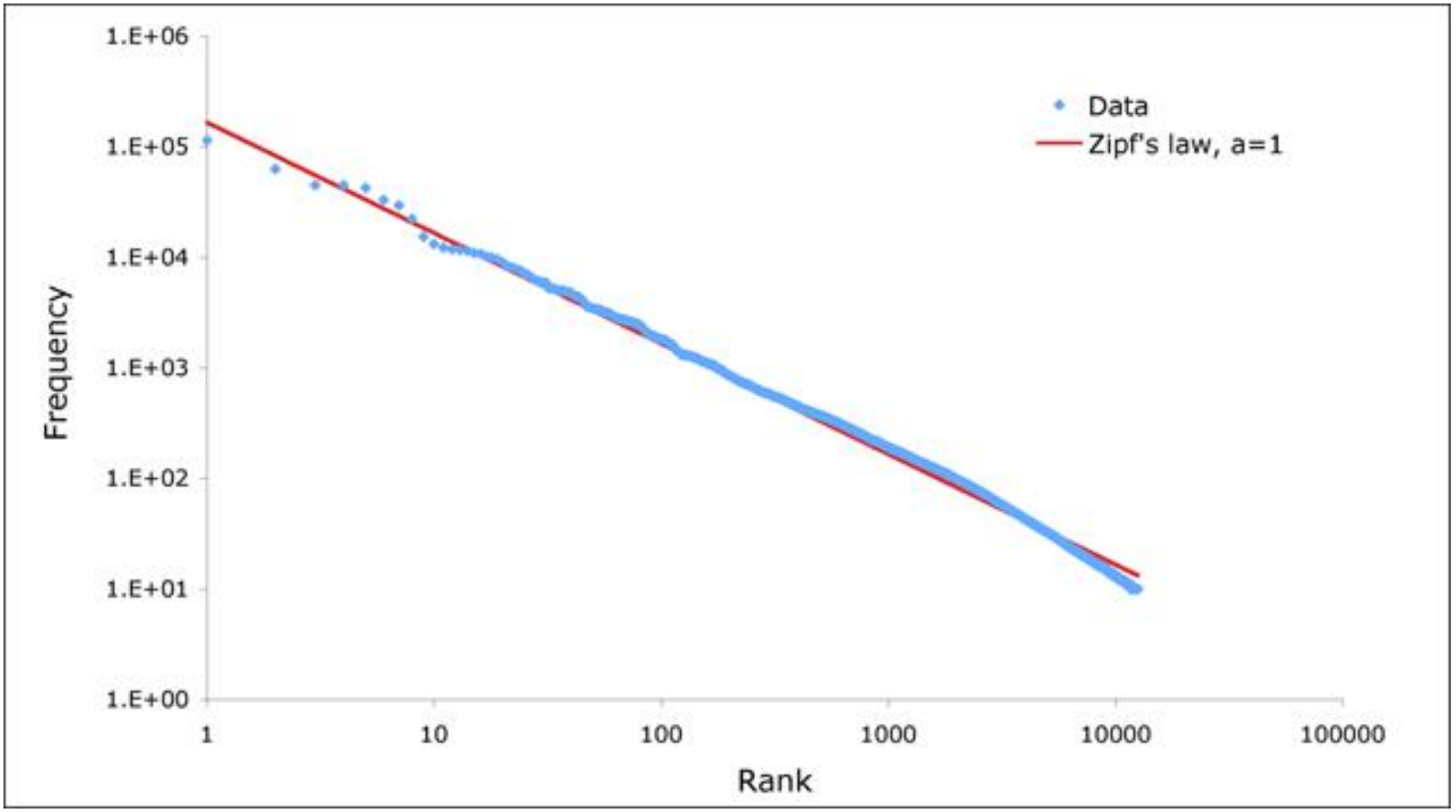

Fig 1: Zipf's Law

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

- Build a mathematical model for distribution of vocabulary frequency on a social media platform

- Develop an adaptive model that is more sensitive to text noises and asymptotically quicker

LIST OF FIGURES

OBJECTIVE

- To access a large existing twitter corpus / dataset
- Split the data into a data-frame and write a word-count program
- Rank the words in descending order
- Plot the graph using seaborn visualization tool

TOOLS AND LIBRARIES USED

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

RESULTS

Fig 2: Twitter data-set

	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

	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
- People tend to use more informal use of vocabulary
- People also use shorter forms of longer words to say time

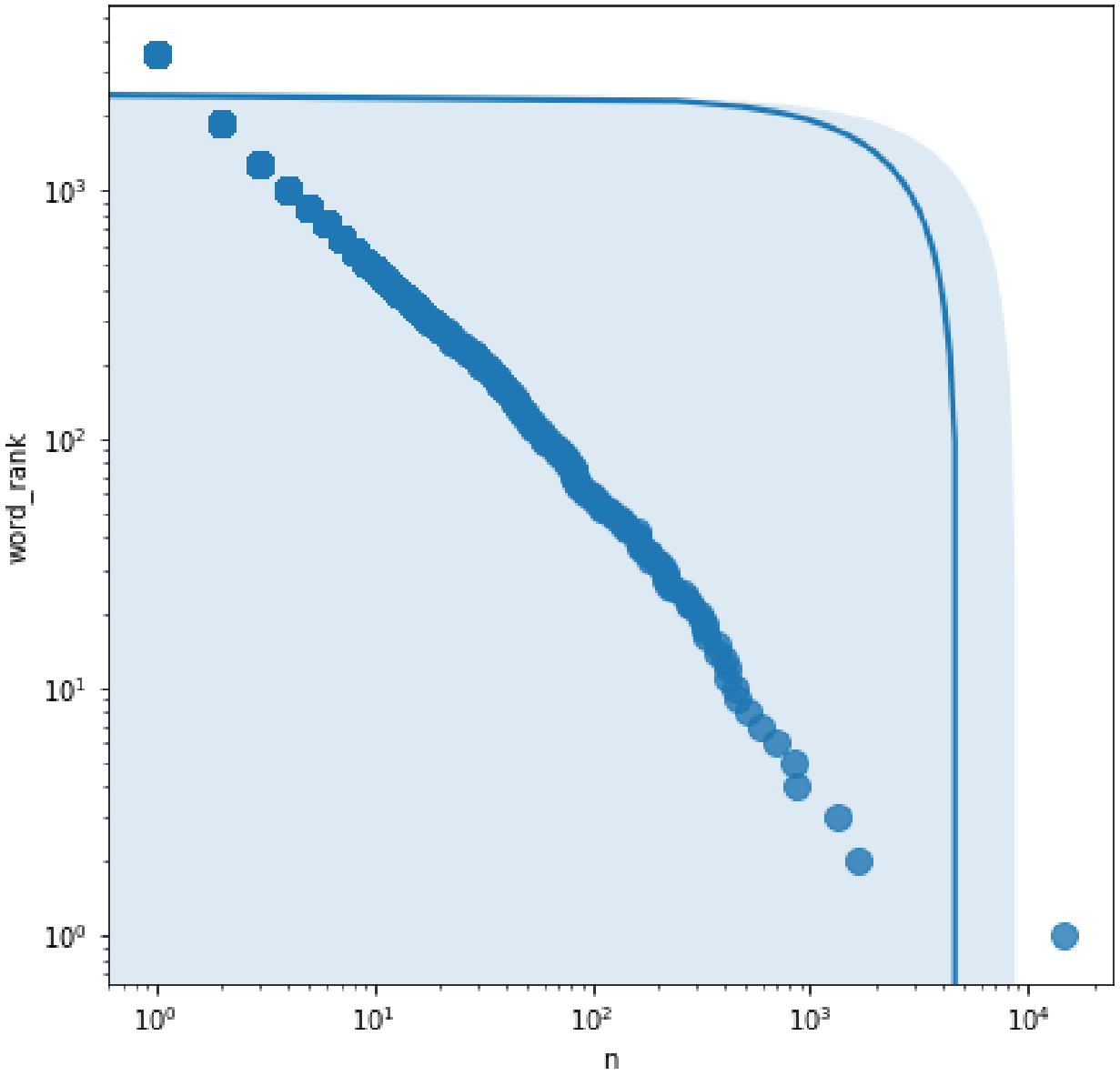

Fig 4: Visualization of pre-processed data-set

CONCLUSION

- Man tries to use more common word, which may not describe his intention so accurately than an uncommon word to express his real emotion / intent

- Man uses 20% of more commonly used and versatile words in 80% of his conversation as this is more productive for him

QUESTIONS ?

THANK YOU !!