Amirkabir University of Technology
(Tehran Polytechnic)

Natural Language Processing

Lecture 3: Zipf's Law

Amirkabir University of Technology
Dr Momtazi

Outline

- **Zipf's Law**

Zipf's Analysis

- Count the frequency of all the words in a corpus
- Sort the words by frequency
- Rank: position of a word in the sorted list

Word Frequency

Rank	Word	Count	Freq(%)
1	The	69970	6.8872
2	of	36410	3.5839
3	and	28854	2.8401
4	to	26154	2.5744
5	a	23363	2.2996
6	in	21345	2.101
7	that	10594	1.0428
8	is	10102	0.9943
9	was	9815	0.9661
10	He	9542	0.9392
11	for	9489	0.934
12	it	8760	0.8623
13	with	7290	0.7176
14	as	7251	0.7137
15	his	6996	0.6886
16	on	6742	0.6636
17	be	6376	0.6276
18	at	5377	0.5293
19	by	5307	0.5224
20	I	5180	0.5099

Word Frequency

Rank	Word	Count	Freq(%)	Freq x Rank
1	The	69970	6.8872	0.06887
2	of	36410	3.5839	0.07167
3	and	28854	2.8401	0.0852
4	to	26154	2.5744	0.10297
5	a	23363	2.2996	0.11498
6	in	21345	2.101	0.12606
7	that	10594	1.0428	0.07299
8	is	10102	0.9943	0.07954
9	was	9815	0.9661	0.08694
10	He	9542	0.9392	0.09392
11	For	9489	0.934	0.10274
12	It	8760	0.8623	0.10347
13	With	7290	0.7176	0.09328
14	As	7251	0.7137	0.09991
15	His	6996	0.6886	0.10329
16	On	6742	0.6636	0.10617
17	Be	6376	0.6276	0.10669
18	At	5377	0.5293	0.09527
19	By	5307	0.5224	0.09925
20	I	5180	0.5099	0.10198

5

Zipf's Law

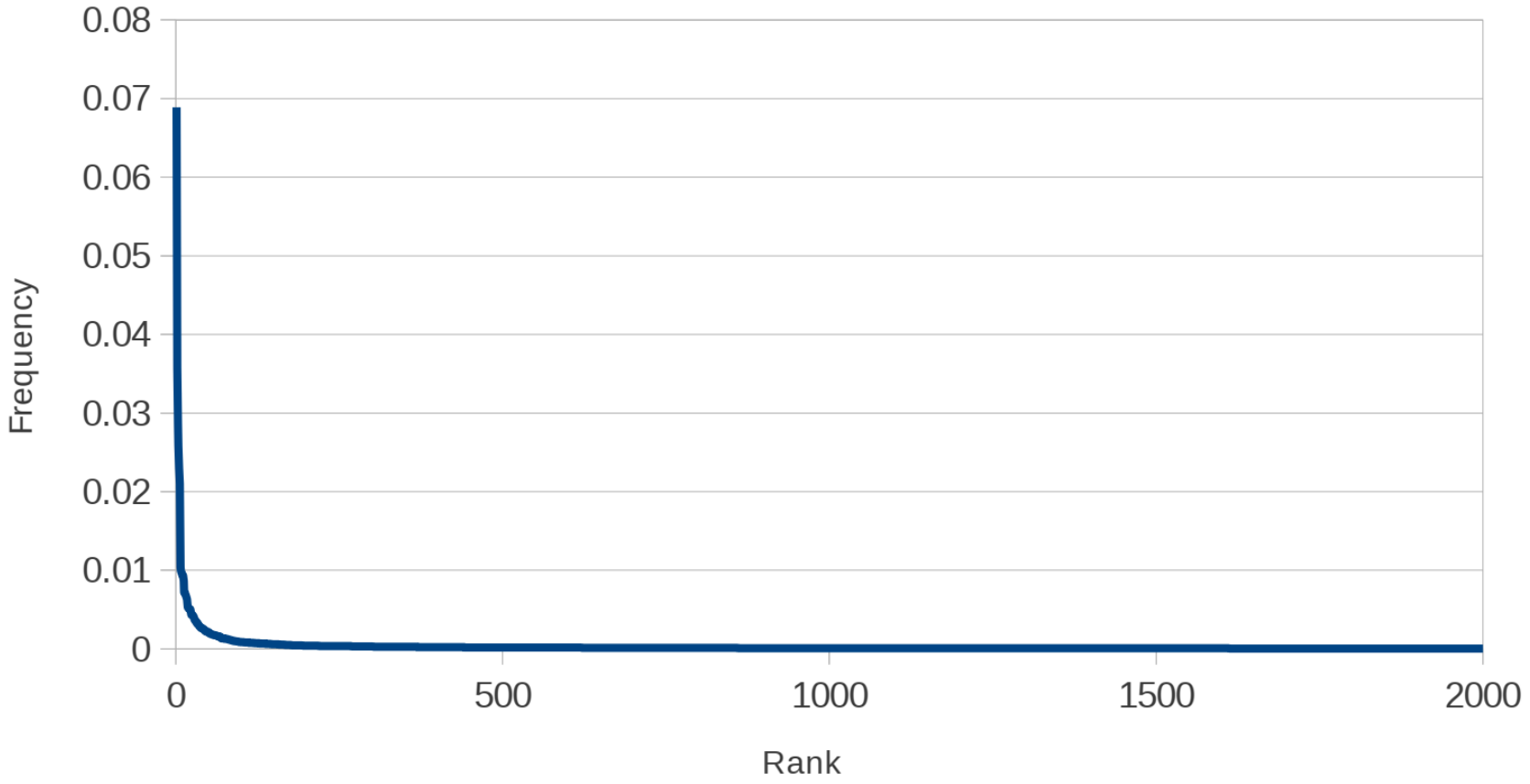
- The frequency of any word is inversely proportional to its rank in the frequency table
- Given a corpus of natural language utterances, the most frequent word will occur approximately
  - twice as often as the second most frequent word,
  - Three times as often as the third most frequent word,
  - ...

$\Longrightarrow$Rank of a word times its frequency is approximately a constant

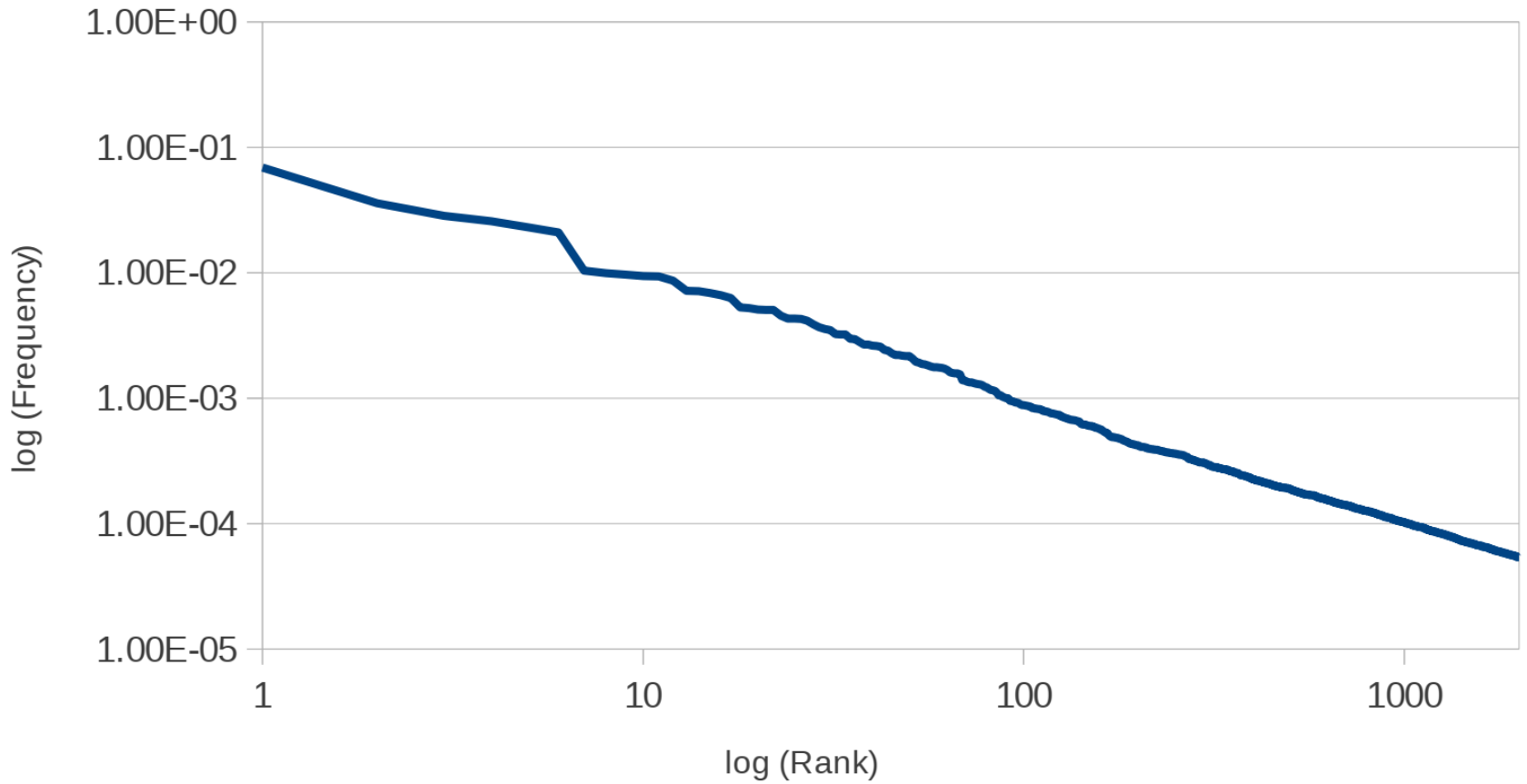
$$\text{Rank} \cdot \text{Freq} \approx c$$

$$c \approx 0.1 \text{ for English}$$

Zipf's Law


0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
Frequency
0
500
1000
1500
2000
Rank

Zipf's Law


1.00E+00
1.00E-01
1.00E-02
1.00E-03
1.00E-04
1.00E-05
log (Frequency)
1
10
100
1000
log (Rank)

Word Frequency

- Zipf's Law is not very accurate for very frequent and very infrequent words

Rank	Word	Count	Freq(%)	Freq x Rank
1	The	69970	6.8872	0.06887
2	of	36410	3.5839	0.07167
3	and	28854	2.8401	0.0852
4	to	26154	2.5744	0.10297
5	a	23363	2.2996	0.11498

Word Frequency

- Zipf's Law is not very accurate for very frequent and very infrequent words

Rank	Word	Count	Freq(%)	Freq x Rank
1000	current	104	0.0102	0.102
1001	spent	104	0.0102	0.1021
1002	eight	104	0.0102	0.1022
1003	covered	104	0.0102	0.1023
1004	Negro	104	0.0102	0.1024
1005	role	104	0.0102	0.10251
1006	played	104	0.0102	0.10261
1007	I'd	104	0.0102	0.10271
1008	date	103	0.0101	0.1018
1009	council	103	0.0101	0.1019
1010	race	103	0.0101	0.102

Further Reading

- Foundations of Statistical Natural Language Processing
  - Chapter 1