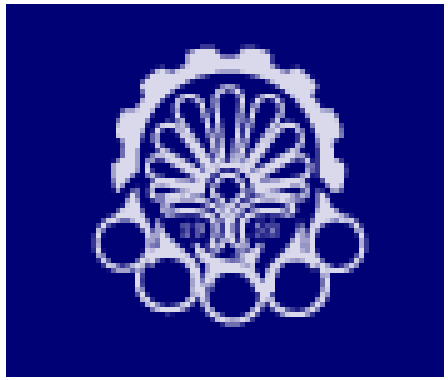


Statistical Natural Language Processing

Lecture 3: Matematical Fundations

Dr. Momtazi
Amirkabir University of Technology

Outline

1. Zipt's Law
2. Probability Theory

Outline

1. Zipt's Law
2. Probability Theory

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	2.1010
7	that	10594	1.0428
8	is	10102	0.9943
9	was	9815	0.9661
10	He	9542	0.9392
11	for	9489	0.9340
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	0.08520
4	to	26154	2.5744	0.10297
5	a	23363	2.2996	0.11498
6	in	21345	2.1010	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	0.9340	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

$Freq \cdot Rank \approx c$

Zipf's Law

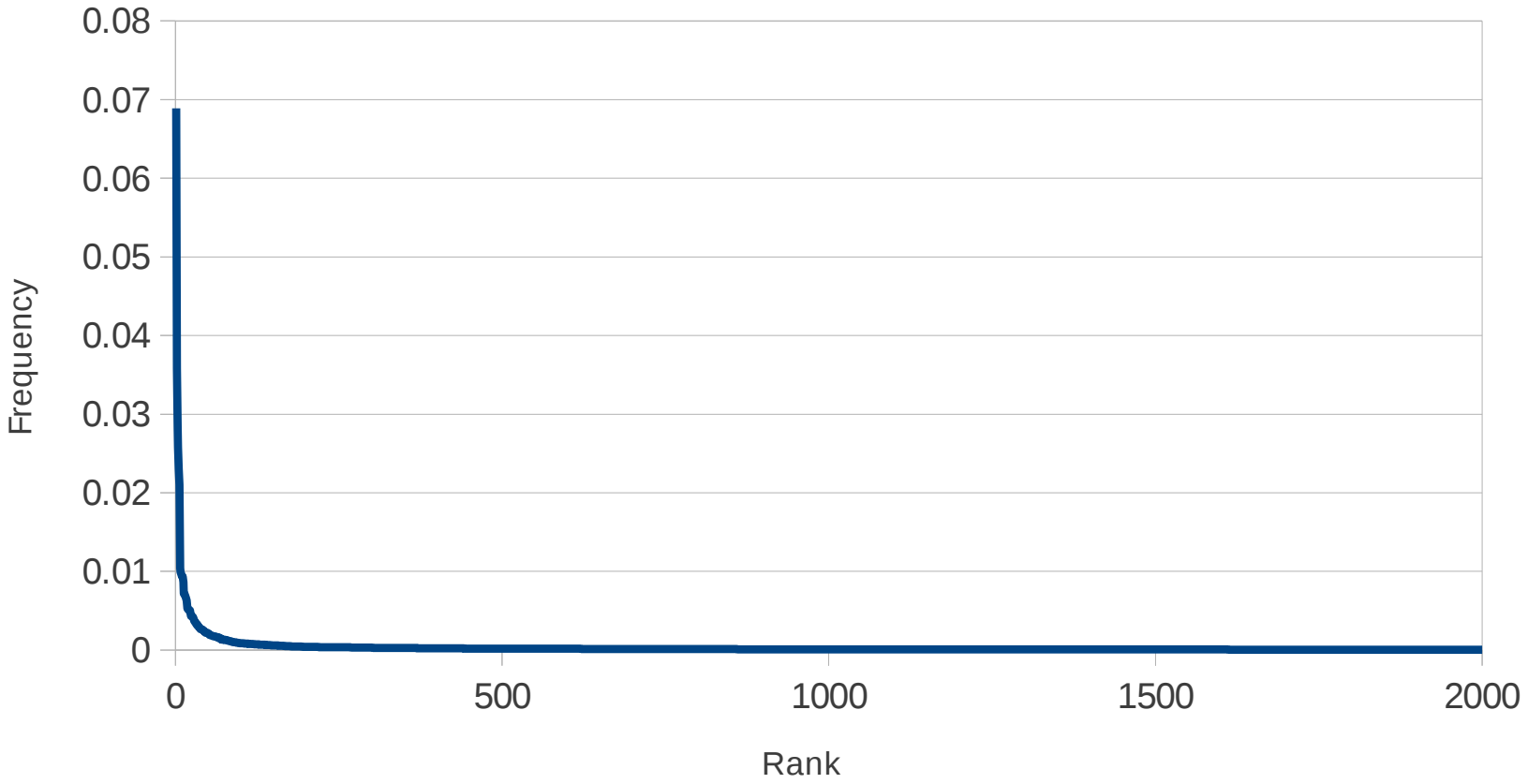
- The frequency of any word is inversely proportional to its rank in the frequency table
- Given a corpus of natural language utterances, the most frequent word will occur approximately
  - twice as often as the second most frequent word,
  - three times as often as the third most frequent word,
  -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

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

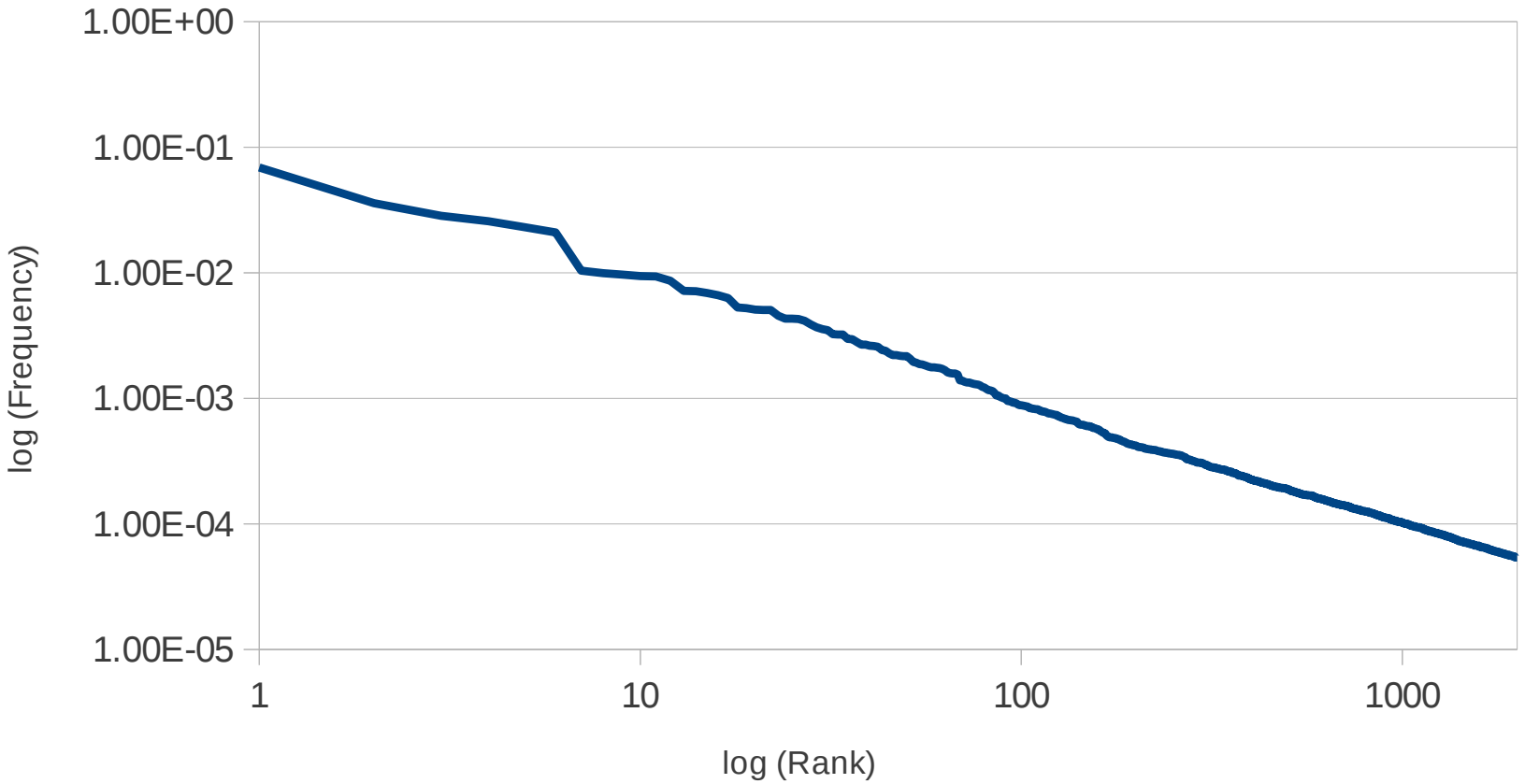
$$Rank \cdot Freq \approx c$$

$c \approx 0.1$ 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	0.08520
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	0.10200
1001	spent	104	0.0102	0.10210
1002	eight	104	0.0102	0.10220
1003	covered	104	0.0102	0.10230
1004	Negro	104	0.0102	0.10240
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	0.10180
1009	council	103	0.0101	0.10190
1010	race	103	0.0101	0.10201

Outline

Motivation

- Statistical NLP aims to do statistical inference for the field of NL
- Statistical inference consists of taking some data (generated in accordance with some unknown probability distribution) and then making some inference about this distribution.
- In order to do this, we need a model of the language.
- Probability theory helps us finding such model

Probability Space

- How likely it is that something will happen
- Sample space Ω is listing of all possible outcome of an experiment
- Event A is a subset of Ω
- Probability function (or distribution)

$$P : \Omega \rightarrow [0, 1]$$

Prior Probability

- Prior probability: the probability before we consider any additional knowledge

$$P(A)$$

Example

- A fair coin is tossed 3 times. What is the chance of 2 heads?
- Solution:
  - The sample space is:
    $\Omega = HHH, HHT, HTH, HTT, THH, THT, TTH, TTT$
  - The event of interest is: $A = HHT, HTH, THH$
  - Each of the basic outcomes in Ω is equally likely, and thus has probability 1/8

$$P(A) = \frac{|A|}{|\Omega|} = \frac{3}{8}$$

Probability of a Word

- Positive

$$P(w = w_i) \geq 0 \quad \forall w_i \in W$$

- Normalized

$$\sum_{w_i \in W} P(w_i) = 1$$

- Additive

$$P(w = w_i \vee w = w_j) = P(w = w_i) + P(w = w_j) \quad \forall w_i \neq w_j$$

Probability of a Sequence of Words

$$P(\text{“to be or not to be”}) = P(w_1 = \text{“to”}, w_2 = \text{“be”}, ...w_6 = \text{“be”})$$

- Shorthand notation:

If we have a specific sequence $w_1, w_2, w_3, ...w_N$
We denote the probability of this specific sequence by
$P(w_1, w_2, w_3, ...w_N)$

Conditional Probability

- Sometimes we have partial knowledge about the outcome of an experiment
- Conditional (or Posterior) Probability
- Suppose we know that event B is true
- The probability that A is true given the knowledge about B is expressed by

$$P(A|B)$$

Conditional Probability

- Definition:

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

- Interpretation: $P(A|B)$ is the probability that A is observed given that the predecessor item is B

Statistical Independence

- Definition:

$$P(A|B) = P(A)$$

- Consequence:

$$P(A, B) = P(A).P(B)$$

Bayes theorem

- Definition:

$$P(B|A).P(A) = P(A|B).P(B)$$

- Proof:

$$P(B|A).P(A) = \frac{P(A,B)}{P(A)}.P(A) = P(A,B)$$

$$P(A|B).P(B) = \frac{P(A,B)}{P(B)}.P(B) = P(A,B)$$

Example

S:stiff neck, M: meningitis

$$P(S|M) = 0.5 \quad P(M) = 1/50,000 \quad P(S) = 1/20$$

I have stiff neck, should I worry?

$$P(M|S) = \frac{P(S|M).P(M)}{P(S)}$$

$$P(M|S) = \frac{0.5 \times 1/50,000}{1/20} = 0.0002$$

Bayes Decomposition

- Write joint probability as product of conditional probabilities

$$P(A, B) = P(A) \cdot P(B|A)$$

$$P(A, B, C, D) = P(A) \cdot P(B|A) \cdot P(C|A, B) \cdot P(D|A, B, C)$$

$$P(w_1, w_2, w_3, w_4) = P(w_1) \cdot P(w_2|w_1) \cdot P(w_3|w_1, w_2) \cdot P(w_4|w_1, w_2, w_3)$$

$$P(w_1, w_2, ...w_n) = P(w_1) \cdot P(w_2|w_1) \cdot P(w_3|w_1, w_2) \cdots P(w_n|w_1, w_2, w_3, ..., w_{n-1})$$

Bayes Decomposition

- Write joint probability as product of conditional probabilities

$$P(S) = \prod_{i=1}^{n} P(w_i | w_1, w_2, ..., w_{i-1})$$

Random Variables

- A random variable (RV) X is a variable whose possible values are numerical outcomes of a random phenomenon.
- Types of random variables:
  - Discrete
  - Continuous

Expectation value

- The Expectation is the mean or average of a RV
- Suppose random variable X can take value x_1 with probability p_1, value x_2 with probability p_2, and so on, up to value x_k with probability p_k. Then the expectation of this random variable X is defined as

$$E[X] = x_1 p_1 + x_2 p_2 + ... + x_k p_k$$

- Example:
  Let X represent the outcome of a roll of a fair six-sided die.

$$E[X] = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5$$

Variance

- The variance of a RV is a measure of whether the values of the RV tend to be consistent over trials or to vary a lot

$$Var(X) = E((X - E(X))^2) = E(X^2) - E^2(X)$$

Further Reading

- FSNLP
  - Chapter 2