Introduction

(SNLP tutorial)

Vilém Zouhar, Awantee Deshpande, Julius Steuer

April 21, 2021

Overview

- Hello
- Topics (15 minutes)
- Requirements
- Materials
- Assignments
- Homework
- Zipf's Law (30 minutes)
- QA

Hello

Who am I?

Hello

Who am I?

Who are you?

🤔

Topics

Task: Pick one not yet taken + why do you find it interesting.

- Language properties, Zipf's Law, basic statistical formalism
- Entropy, basic information theory (Shannon's game, entropy-based quantities, code lengths)
- Language modelling, back-off models (interpolation, discounting)
- Text classification, basic algorithms (kNN, decision trees, SVM, . . .)
- Word sense disambiguation, basic algorithms (dictionary-, translation-, collocation-based)
- Information retrieval, latent semantic analysis, singular value decomposition
- Machine translation, word alignment, beamsearch
- POS tagging, named entity recognition
-
    - sequence labeling (hidden markov chains / models, conditional random fields)

Requirements

Tutorial Requirements (exam admission)

- 70% of mandatory points (∼10 assignments, 10 points each)
- Tutorial points only for exam admission (no final grade influence)

Tutorial Bonus Points

- ∼2pts for extra excercises in the assignments
- 1pt for participating and *talking* in an tutorial
- Presenting a solution to an excercise (∼5 points)
- ▸ Presentable excercises are marked in the assignment sheet
- ▸ Let individual tutors known if you wish to present (first come - first serve)
- ▸ Every group can present *at most* once, about 10 to 15 minutes

Final Project

- 25% of the final grade
- Details TBD

Transfer from last year

- Possible
- Do project and exam

What’s available

- Lectures by prof. Klakow (recorded)
- Tutorials
- Corrected homework
- Consultations
- ▸ Only in specific cases
- ▸ By default **no** email and **no** chat
- ▸ Better ask during the lecture / tutorials
- Public forum (please use Piazza)
- ▸ Ask questions
- ▸ Other students will also benefit from the answers
- ▸ You can answer someone else’s issue

Assignments

- Mandatory groups of 2
- Usually 3 excerises per one assignment
- ▸ Can't be changed later (very special exceptions)
- Jupyter notebook templates
- ▸ Assignment + solution in the same notebook
- ▸ Can use Google Collab or local runtime
- Only one submission per group
- ▸ Submit through Teams

Dates / Times

- Lecture: Fridays 8:30-10:00
- Tutorials:
- ► Awantee: TODO
- ► Julius: TODO
- ► Vilém: TODO
- Assignments
- ► Release (usually) Friday 23:59
- ► Deadline (next) Friday 23:59 (also in Teams)
- Exam: (TBD) 30. Jul.

Tutorial Content

- Review of the topic (per demand)
- Presentation of the past assignment
- Troubleshooting current assignment

Current Homework

- Notebook instructions
- Stick breaking
- Zipf's law on words
- Bonus: Zipf's law on characters

Languages

Language

$L \subseteq \Sigma^*$ (all possible substrings by elements of alphabet Σ)

Languages

Language

$L \subseteq \Sigma^*$ (all possible substrings by elements of alphabet Σ)

- $\Sigma_1 = \{$a,b,...,z,ü,ä,ö$\}$
- $\Sigma_2 = \{$A,G,C,T$\}$
- $\Sigma_3 = \{$def,True,:,print,...$\}$
- $\Sigma_4 = \{$SELECT,INSERT,DROP,...$\}$
- $\Sigma_5 = \{$hallo,ja,nein,...$\}$
- $\Sigma_6 = \{$+,-,=,1,2,3...$\}$
- $\Sigma_7 = \{$+,-,=,1,2,3...$\}$

Languages

Language

$L \subseteq \Sigma^*$ (all possible substrings by elements of alphabet Σ)

- $\Sigma_1 = \{$a,b,...,z,ü,ä,ö$\}$
- $\Sigma_2 = \{$A,G,C,T$\}$
- $\Sigma_3 = \{$def,True,:,print,...$\}$
- $\Sigma_4 = \{$SELECT,INSERT,DROP,...$\}$
- $\Sigma_5 = \{$hallo,ja,nein,...$\}$
- $\Sigma_6 = \{$+,-,=,1,2,3...$\}$
- $\Sigma_7 = \{$+,-,=,1,2,3...$\}$

- ‘Oberfläche’ $\in L_1$ (German words)
- ‘..GATTCCAATCAG’ $\in L_2$ (DNA)
- ‘while True: f()’ $\in L_3$ (Python)
- ‘SELECT * FROM tbl;’ $\in L_4$ (SQL)
- ‘Wie geht's dir?’ $\in L_5$ (German)
- ‘4=5’ $\in L_6$ (arithmetics)
- ‘1=2+=3333=’ $\in L_7$ (???)

Languages

Language

$L \subseteq \Sigma^*$ (all possible substrings by elements of alphabet Σ)

- $\Sigma_1 = \{\texttt{a,b,...,z,ü,ä,ö}\}$
- $\Sigma_2 = \{\texttt{A,G,C,T}\}$
- $\Sigma_3 = \{\texttt{def,True,:,print,...}\}$
- $\Sigma_4 = \{\texttt{SELECT,INSERT,DROP,...}\}$
- $\Sigma_5 = \{\texttt{hallo,ja,nein,...}\}$
- $\Sigma_6 = \{\texttt{+,-,=,1,2,3...}\}$
- $\Sigma_7 = \{\texttt{+,-,=,1,2,3...}\}$

- ‘`Oberfläche`’ $\in L_1$ (German words)
- ‘`..GATTCCAATCAG`’ $\in L_2$ (DNA)
- ‘`while True: f()`’ $\in L_3$ (Python)
- ‘`SELECT * FROM tbl;`’ $\in L_4$ (SQL)
- ‘`Wie geht's dir?`’ $\in L_5$ (German)
- ‘4=5’ $\in L_6$ (arithmetics)
- ‘1=2+=3333=’ $\in L_7$ (???)

Usually defined by the alphabet and production rules (Automata and Grammar).

Zipf's Law

1. Sort words/entries by frequency f(x)
2. r(x) = position in the sorted list
   - Then $f(x) \propto \frac{1}{r(x)^\gamma}$ (γ parameter)
3. Most common entry m.
   - Then $f(x) = \frac{f(m)}{r(x)^\gamma}$

Zipf's Law

1. Sort words/entries by frequency f(x)
2. r(x) = position in the sorted list
   - Then $f(x) \propto \frac{1}{r(x)^{\gamma}}$ (γ parameter)
3. Most common entry m.
   - Then $f(x) = \frac{f(m)}{r(x)^{\gamma}}$

Rank	Frequency	Predicted ($\gamma = 0.7$)
4	606	$1507/4^{0.7} = 571$
5	490	$1507/5^{0.7} = 488$
...		
11	261	$1507/11^{0.7} = 281$
12	252	$1507/12^{0.7} = 264$

Rank	Word [3]	Frequency
1	the	1507
2	and	714
3	to	703
4	a	606
5	of	490
6	she	484
7	said	416
8	it	346
9	in	345
10	was	328
11	I	261
12	you	252
13	as	237
14	Alice	221

Zipf's Law

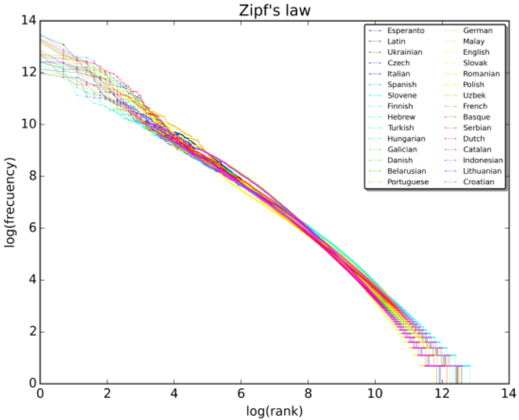

Figure 1: log-log plot of Zipf's Law applied to natural languages [4]

We are plotting:

$$\log\left(\frac{C}{\exp(x)^{\gamma}}\right)$$

$$\log(C) - \gamma \log \exp(x)$$

$$\log(C) - \gamma x$$

Zipf's Law Notes

- Works beyond natural languages

Rank	City	Population
1	New York, N.Y.	8,491,079
2	Los Angeles, Calif.	3,928,864
3	Chicago, Ill.	2,722,389
4	Houston, Tex.	2,239,558
5	Philadelphia, Pa.	1,560,297
6	Phoenix, Ariz.	1,537,058
7	San Antonio, Tex.	1,436,697
8	San Diego, Calif.	1,381,069
9	Dallas, Tex.	1,381,069

Zipf's Law Notes

- Works beyond natural languages
- DNA subsequences of fixed lengths

Rank	City	Population
1	New York, N.Y.	8,491,079
2	Los Angeles, Calif.	3,928,864
3	Chicago, Ill.	2,722,389
4	Houston, Tex.	2,239,558
5	Philadelphia, Pa.	1,560,297
6	Phoenix, Ariz.	1,537,058
7	San Antonio, Tex.	1,436,697
8	San Diego, Calif.	1,381,069
9	Dallas, Tex.	1,381,069

Zipf's Law Notes

- Works beyond natural languages
- DNA subsequences of fixed lengths
- Code (programming languages)

Rank	City	Population
1	New York, N.Y.	8,491,079
2	Los Angeles, Calif.	3,928,864
3	Chicago, Ill.	2,722,389
4	Houston, Tex.	2,239,558
5	Philadelphia, Pa.	1,560,297
6	Phoenix, Ariz.	1,537,058
7	San Antonio, Tex.	1,436,697
8	San Diego, Calif.	1,381,069
9	Dallas, Tex.	1,381,069

Zipf's Law Notes

- Works beyond natural languages
- DNA subsequences of fixed lengths
- Code (programming languages)
- Population of cities (frequency is city population) [5]

Rank	City	Population
1	New York, N.Y.	8,491,079
2	Los Angeles, Calif.	3,928,864
3	Chicago, Ill.	2,722,389
4	Houston, Tex.	2,239,558
5	Philadelphia, Pa.	1,560,297
6	Phoenix, Ariz.	1,537,058
7	San Antonio, Tex.	1,436,697
8	San Diego, Calif.	1,381,069
9	Dallas, Tex.	1,381,069

Resources

1. UdS SNLP Class: https://teaching.lsv.uni-saarland.de/snlp/
2. Tutorial repository for these slides: https://github.com/zouharvi/uds-snlp-tutorial
3. Piazza: https://piazza.com/uni-saarland.de/spring2021/snlp