L665 Machine Learning for NLP
1. Introduction

Damir Cavar

Indiana University
Department of Linguistics

January 2018

Outline

General

General Issues

Me: Damir Cavar, Department of Linguistics

Office: BH 850, office hours after class 11–12

Email: dcavar@iu.edu

- We use Canvas for everything! Except for...
  No homework submission via email.

TA: Atreyee Mukherjee (atremukh@indiana.edu)

HW: Assignments are always described on the last slides and in Canvas.

Speech and Language

- Signal
- Phones, Phonemes, Syllables
- Morphemes, Wordformation
- Phrases, Sentences
- Meaning, Intention

Dealing with Language in Computers

- Texts, Words, and Lexicon
- Frequency Profiles
- Lexical Properties
- Stemming
- Part-of-Speech Tagging
- Lexical properties
- Parsing

Resources

- Files and environment
- Coding examples
- Data and tools (e.g. Antconc, oXygen, Protege, GATE, ...)
- ...

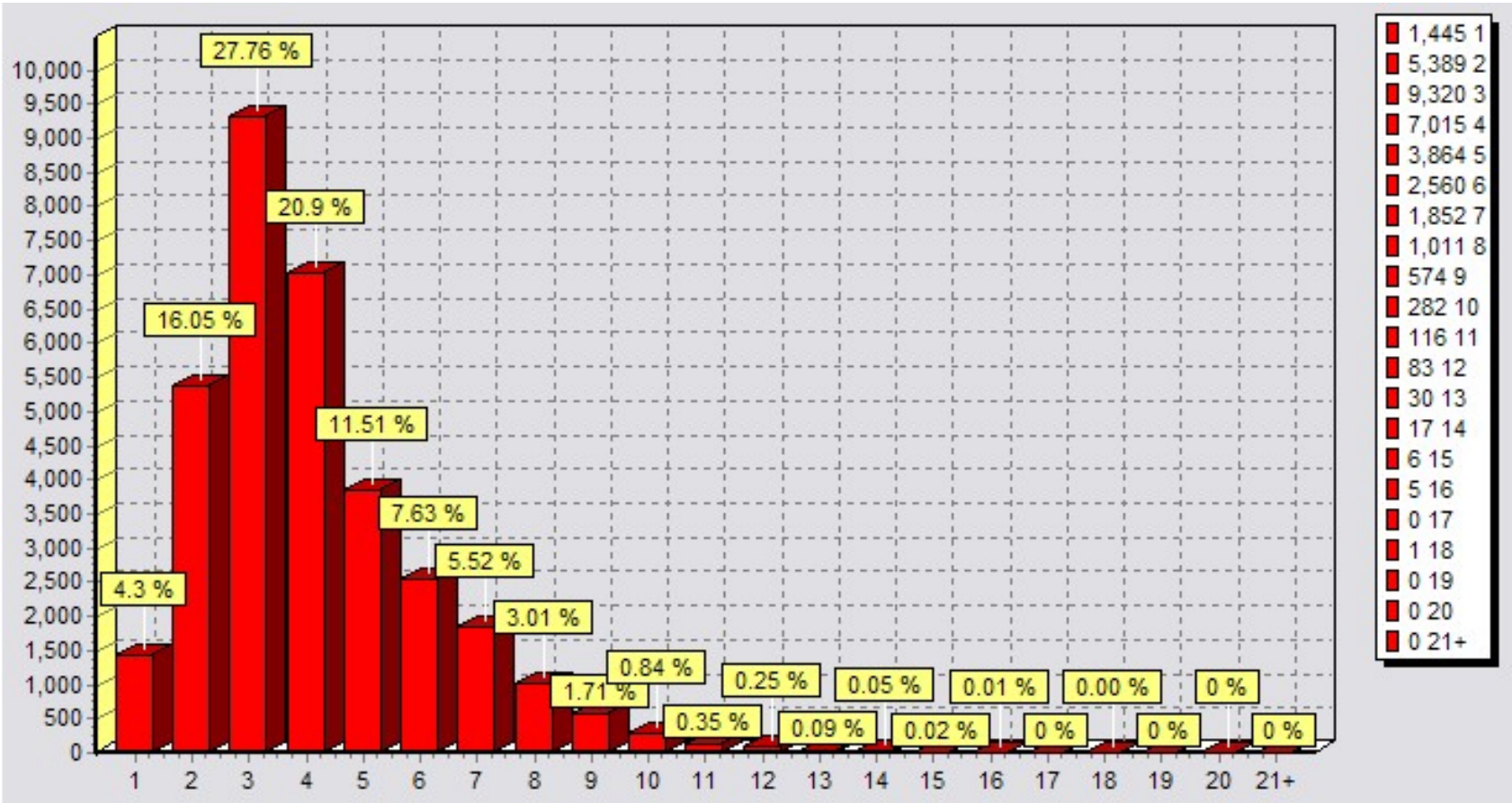
Lexicology

- Download:
  Antconc
- A text, e.g. *The House of Pomegranates* by Oscar Wilde.
  Gutenberg Archive book 873
  Direct link to UTF-8 text

Lexicon

- Example: data/HOPG.txt
  - "A House of Pomegranates" by Oscar Wilde
  - Number of tokens: 33570
  - Number of words: 4066
  - Mapping word frequency on word length...

Quantitative Properties



Quantitative Properties

- 49 most frequent words:
  - THE, AND, OF, TO, A, HE, HIS, IN, THAT, WITH, HIM, WAS, IT, I, HER, FOR, IS, ME, HAD, THEY, BUT, ON, AS, AT, SHE, NOT, FROM, THEIR, SAID, THOU, THEM, THEE, WHEN, WHO, WERE, SO, HAVE, LITTLE, OUT, YOUNG, MY, BY, BE, SOUL, THERE, CAME, THIS, WILL, INTO

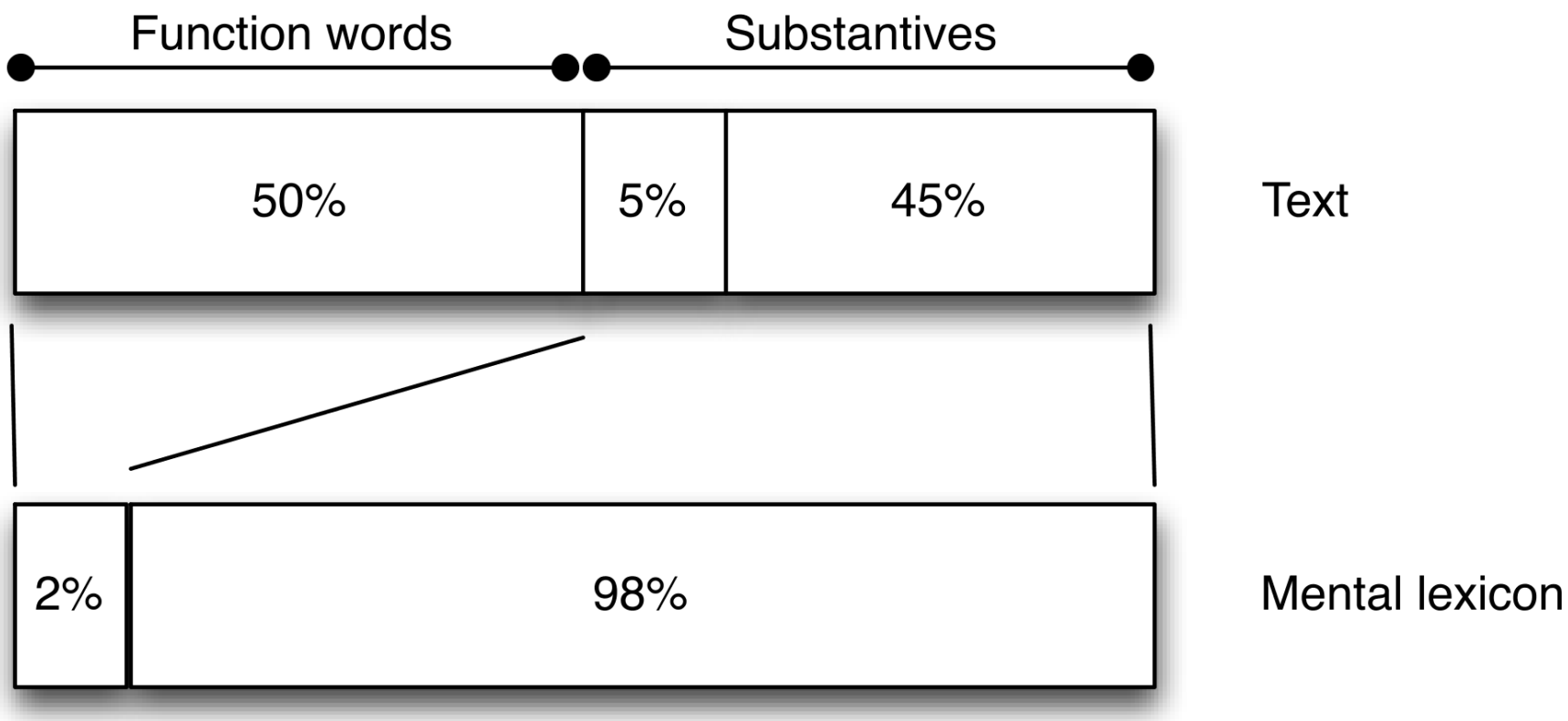
Quantitative Properties

- 49 most frequent words:
  - all together make up 16489 tokens out of 33570
    - 49 % of all tokens
  - 49 words out of 4066
    - 1.2 % of all words

Lexical Co-occurrence

Rank	Word	Frequency	Co-occ.
1	the	8705	4010
2	of	4220	3000
3	and	3055	3110
4	to	3049	2233
		Sum[1-4]	**12353**
5	in	2629	2030
6	a	2190	1610
7	that	1394	1105
8	is	1160	883
9	was	1090	913
10	it	969	476
11	for	967	943
12	as	847	705
13	on	818	741
14	this	675	456
15	by	664	714
16	with	613	646
17	not	586	377
18	be	548	405
19	but	522	283
20	he	522	314
		Sum[1-20]	**24954**

Quantitative Properties



Quantitative Properties

- Entropy effect on the lexicon
- Dynamic and static at the same time
  - **Open Class Lexicon:** Nouns, Verbs, Adjectives, ... being highly dynamic in a language (one estimate: ca. 20,000 new words in English per year), leading to group, region or dialect variation
  - **Closed Class Lexicon:** Functional Categories (articles, prepositions, conjunctions) being static in the language (language specific) and over speakers (thus in text)

Quantitative Properties

- **Data-driven linguistics and computational methods:**
  - Observed data from corpora: word frequencies, N-gram models, co-occurrence patterns, etc.
  - Extracted models from **seen data**
  - Applied models to **unseen data**
- **Issues:**
  - Language does not work this way!

Quantitative Properties

- **Sparseness of data:**
  - Most interesting phenomena do not occur frequently enough in corpora or data samples.
- **Economy of language:**
  - Obvious things are not mentioned (cross-modality is required, a theory, world-knowledge, reported as an argument initially made by Linda Smith, Indiana University):
    orange carrots 0 frequency
    purple carrots high frequency
  - Dark text (or content), empty categories:
    Got it! or *Who caught the ball?* $\rightarrow$ *I have!*
  - Implicatures from speech acts, implications from propositions, assertions, etc.
  - Semantic shifts: *I finished the juice* or *friendly fire*
  - Large number of new words (or tokens) emerging in language every year.

Quantitative Properties

- **Data-driven Natural Language Processing and Quantitative Methods:**
  - Quantitative methods: counting words, lists of words (N-grams), bags of words, word with annotation tuples, etc.
  - Data-driven NLP: e.g. corpus-based machine learning methods
  - *Issues or Problems:*
    - Language does not display all data that is implicitly there.
    - Cultural or common knowledge is optimized away from communication.

MS on Rationalist vs. Empiricist Approaches

- Rationalist Linguistics
- Empiricist Linguistics
- Arguments for and against:
  - Poverty of Stimulus: pronouns and Binding Theory
  - Multi-modality of language
  - Economy of language
  - Sparseness of data
- Other notions:
  - Competence vs. Performance
  - I-language, E-language, nativist

Quantitative Properties

Frequency profile in NLTK:

```
>>> import nltk
>>> file = open("HOPG.txt", mode='r', encoding='utf-8')
>>> text = file.read()
>>> file.close()
>>> myFd = nltk.FreqDist(text)
```

Print frequency profile:

```
>>> for x in myFd:
...     print(x, myFd[x])
```

Quantitative Properties

Generating a N-gram model, here $n = 2$:

```
>>> myBigrams = nltk.ngrams(text, 2)
>>> for x in myBigrams:
...     print(x)
```

Counting N-grams:

```
>>> myBigrams = nltk.ngrams(text, 2)
>>> myFd = nltk.FreqDist(myBigrams)
>>> for x in myBigrams:
...     print(x)
```

Quantitative Properties

Frequency profile of tokens in NLTK:

```
>>> import nltk
>>> file = open("HOPG.txt", mode='r', encoding='utf-8')
>>> text = file.read()
>>> file.close()
>>> myTokens = nltk.word_tokenize(text)
>>> myFd = nltk.FreqDist(myTokens)
```

Print frequency profile:

```
>>> for x in myFd:
...     print(x, myFd[x])
```

Lexicon and Quantitative Properties

Next session:

- Zipf's law
- Lexicon
- Text
- Language properties

Lexicon and Quantitative Properties

- Zipf's laws
  - Frequency and semantic specificity
  - Clustering in sections of text

Lexicology

- Words
  - Distributional properties
  - Form
  - Meaning
  - Formal properties
    - Category and morpho-syntactic information
    - Syntactic and semantic frames

Text

- Document classification and clustering
- Text-mining and Information extraction
- Sentiment analysis
- ...

TODO

- Reading: MS Chapters 1, 2.1
- Probability Theory for next session
- Optional tasks:
  - Prepare your laptop with a Python 3 installation (incl. Numpy, Scipy, NLTK, Scikit-learn, spaCy)
  - Read some documentation on Python 3/NLTK: Python Tutorial, NLTK pages, etc.
  - Read Bender (2013), if you do not know anything about linguistics.

TODO

Questions or Issues?

TODO

END Session 2