
**11-442 / 11-642:
Search Engines**

Introduction

Jamie Callan
Carnegie Mellon University
callan@cs.cmu.edu

Outline

What this course is about

- Course philosophy
- Overview of course topics
- Administrative information
- Questions and answers

Introduction to shallow language processing

- The bag of words model
- Heaps' Law
- Zipf's Law

Course Philosophy

This course is about search engine theory and practice

- Beautiful theory, and ugly-but-effective heuristics
 - This field is developing rapidly, so theory is often insufficient
- The emphasis is on what works

This course covers enterprise, web, and vertical search

- They have many similarities
- They also have important differences
- All three are important commercially

3

© 2017, Jamie Callan

Course Philosophy

This is a computer science course

- Algorithms, data structures, computational complexity
- Open-source search software (Indri, Lucene)

You will leave this course with hands-on experience

- Query operators, document ranking, text mining, ...
- Developing software using search engine APIs
- Testing software using standard datasets and evaluation tools
- Lucene, web data

4

© 2017, Jamie Callan

Overview of Course Topics

- Web crawling
- Text representation
- Search engine indexes
- Index construction
- Query structure
- Unsupervised ranking
- Feature-based ranking
- Federated & vertical search
- Page features
- Evaluation
- Search log analysis
- Diversity
- Personalization
- Enterprise search

**This is a conceptual overview
It doesn't match the syllabus exactly**

5

© 2017, Jamie Callan

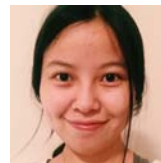
Administrative Information: Teaching Assistants



Rishabh Arora
rishabha@andrew



Haoming Chen
haomingc@andrew



Zhuyun Dai
zhuyund@cs



Akshat Gaur
agaur@andrew



Qinyu Hu
qinyuh@andrew



Qing Liu
qingl2@andrew



Arpita Pyreddy
mpyreddy@andrew



**Varshini
Ramaseshan**
vramases@andrew

6

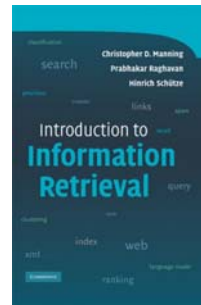
Administrative Information: Textbook

Textbook

- *Introduction to Information Retrieval*, Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze

An online version is available

- <http://nlp.stanford.edu/IR-book/>
- Section numbering may differ slightly
 - Be careful when doing readings



7

© 2017, Jamie Callan

Administrative Information: Course Web Page

Follow the link from Jamie's web page

- <http://boston.lti.cs.cmu.edu/classes/11-642/>
- Syllabus
- Lecture notes
- Reading assignments and copies of papers (when necessary)
- Homework assignments, data files, software, evaluation tools



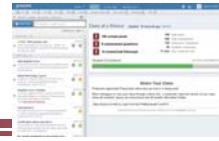
Access to some files is restricted to .cmu.edu

- When you are off-campus, use CMU's VPN service
- See the course Web page for instructions

8

© 2017, Jamie Callan

Administrative Information: Course Discussion Forum



There is a Piazza discussion forum

- <https://piazza.com/cmu/spring2017/1144211642/>

Main purpose: Students exchanging information with students

Secondary purpose: Instructor and TAs answering questions

- We will answer reasonable questions when we have time

Advice that shouldn't be necessary but seems to be...

- This is not a substitute for doing the assigned readings
- Search for similar questions before you post
- Don't wait until the last minute to post a question

9

© 2017, Jamie Callan

Canvas

We will use Canvas for three purposes

1. For you to submit reading summaries each week
2. To send email to you
3. To report homework grades to you

Everything else is done via the course website or Piazza

10

© 2017, Jamie Callan

Administrative Information: Reading Summaries

Every Tuesday before class you must submit a brief summary that describes the readings you did for that week

- Length: ½ - 1 page
- Grading scale:
 - 0: We doubt you read the material or understood it (rare)
 - ✓-: Summary lacks key information
 - ✓: Great! Just what we wanted
 - ✓+: Really great (but you spent too much time on it)

✓ is the best combination of grade and effort

11

© 2017, Jamie Callan

Administrative Information: Homework Descriptions (Tentative)

Assignments (5)

- Implement two unranked Boolean retrieval algorithms
- Implement two popular document ranking algorithms
- Implement pseudo relevance feedback
- Implement features and test learning to rank (LeToR)
- Implement diversified ranking algorithms

12

© 2017, Jamie Callan

Administrative Information: Homework Policies

All homework must be submitted via the course website

- Due by 11:59 pm of the due date

Late homework

- Deduct 10% for first day late, 5% for each additional day late

Don't fall behind

- HW2 builds on HW1
- HW3 builds on HW2
- ...
- It is difficult to catch up later

13

© 2017, Jamie Callan

Administrative Information: Programming Skills

This course requires good Java programming skills

- A good knowledge of Java classes and structure
- Good object-oriented programming skills
 - E.g., to be comfortable creating new subclasses
- Very comfortable with recursion and inheritance
- Good debugging skills

If you have less than 1-2 years of programming experience you may have difficulty in this course

- The TAs will help you with search engine knowledge
- The TAs won't help you with basic programming skills

14

© 2017, Jamie Callan

Administrative Information: Grading

Reading summaries (10%)

- $15 \text{ weeks} \times 0.67\% \text{ per summary} = 10\%$

Homework (50%)

- $5 \text{ assignments} \times 10\% \text{ per assignment} = 50\%$

Midterm exam (20%)

- Covers the first half of the course

Final exam (20%)

- Covers the second half of the course

15

© 2017, Jamie Callan

Administrative Information: Course Policy on Academic Integrity

The policy for this course is simple

- You must be the author of everything that you submit
- Revising or modifying someone else's work does not make you the author

It is okay to discuss homework with other students

- You may share ideas, experiences, insights, and lessons learned
- You may not share detailed pseudo code
- You may not share source code or parts of a report

16

© 2017, Jamie Callan

Administrative Information: Course Policy on Academic Integrity

We will check software and reports for signs of cheating

Penalties

- Failure of the course
 - Five students failed last semester due to cheating
- Possibly other penalties from your graduate program or CMU

**If you are having problems meeting your deadlines
... submit the assignment late**

- A late penalty lowers your grade a little
- Cheating causes you to fail the course

17

© 2017, Jamie Callan

Administrative Information: Class Participation

Laptops and mobile phones must be kept closed during class

- Students who pay attention in class typically get better grades



18

© 2017, Jamie Callan

Administrative Information: Class Participation

Ask questions!

- Questions make the class more interesting
- Questions guide me toward the information that you need

Ask questions during class

- If you are confused, probably others are confused, too
- Help everyone to become less confused

Do you have any questions right now?



19

© 2017, Jamie Callan

Outline

What this course is about

- Course philosophy
- Overview of course topics
- Administrative information
- Questions and answers

Introduction to shallow language processing

- The bag of words model
- Heaps' Law
- Zipf's Law

20

© 2017, Jamie Callan

Introduction to Shallow Language Processing

Search engines consider two types of language properties

1. Language-dependent properties
 - E.g., lexical characteristics, morphology, syntax, ...
 - Covered in a later lecture
2. Language-independent properties
 - Today's focus

23

© 2017, Jamie Callan

Statistical Properties of Text



Rank	Term	Count	Rank	Term	Count
1	the	4,352,160	101	9	80,490
2	of	2,134,125	102	most	80,409
3	to	2,023,402	103	such	80,037
4	a	1,811,373	104	time	80,014
5	in	1,546,782	105	no	78,459
6	and	1,507,140	106	into	78,208
7	s	855,190	107	only	78,150
8	that	787,792	108	trading	78,133
9	for	780,138	109	many	77,578
10	is	605,988	110	so	77,099
11	said	528,481	111	now	76,281
12	it	510,102	112	based	75,798
:	:	:	:	:	:

Wall Street Journal (1987-1992)

Documents: 174K
Tokens: 69M
Terms (types): 211K
Megabytes: 533

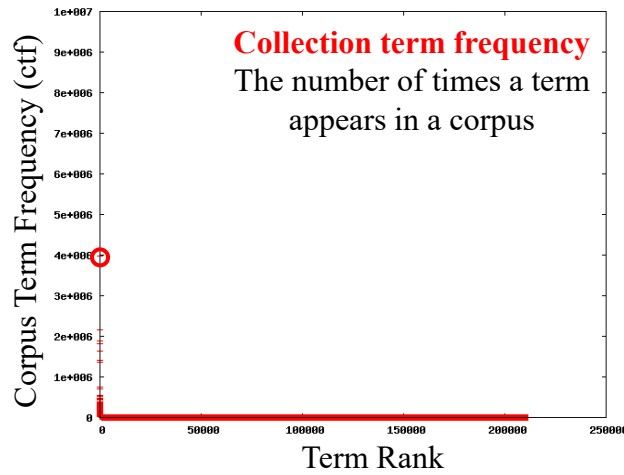
Tokens:
Word occurrences

Types:
Unique words

24

© 2017, Jamie Callan

Term Frequency in the WSJ



Wall Street Journal (1987-1992)

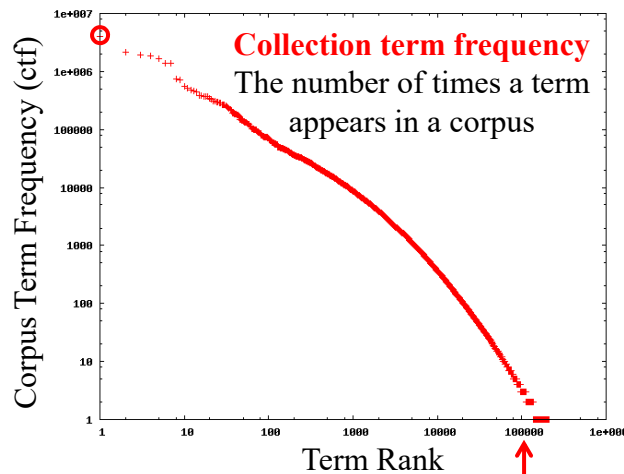
Documents: 174K
Tokens: 69M
Terms (types): 211K
Megabytes: 533

Tokens:
Word occurrences
Types:
Unique words

25

© 2017, Jamie Callan

Term Frequency in the WSJ



Wall Street Journal (1987-1992)

Documents: 174K
Tokens: 69M
Terms (types): 211K
Megabytes: 533

Tokens:
Word occurrences
Types:
Unique words

26

Half of the vocabulary

© 2017, Jamie Callan

Term Frequency

The term frequency distribution is very skewed

- A few really frequent terms, many very rare terms

This pattern is a property of how humans use language

- It holds across different languages and types of documents

What changes?

- The slope may change
- The position of an individual term will change
 - E.g., “linux” may be frequent in one corpus, rare in another

27

© 2017, Jamie Callan

Statistical Properties of Text

Two “laws” describe how words are used in human languages

- **Heaps’ Law:** The size of the vocabulary
- **Zipf’s Law:** The term frequency distribution

28

© 2017, Jamie Callan

Statistical Properties of Text: Heaps' "Law"

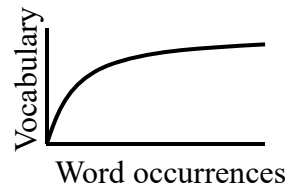
**Heaps' Law predicts the number of distinct terms
(vocabulary size, V)**

$$V = KN^\beta$$

K : usually $10 \leq K \leq 100$

β : usually $0.4 \leq \beta \leq 0.6$ for English

N : total number of word occurrences



The vocabulary never stops growing

- Misspellings, names, new words, ...

29

© 2017, Jamie Callan

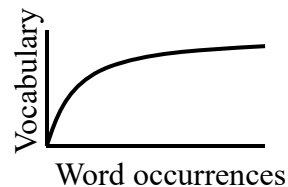
Statistical Properties of Text: Heaps' "Law"

**Heaps' Law predicts the number of distinct terms
(vocabulary size, V)**

K : usually $10 \leq K \leq 100$

β : usually $0.4 \leq \beta \leq 0.6$ for English

N : total number of word occurrences



Is it a good fit?

- WSJ 87-93, Predicted: $V = 25 \times (69,000,000)^{0.5} = 208\text{K terms}$
- WSJ 87-93, Actual: $V = 211\text{K terms}$

Parameters converge quickly (e.g., in a few million words)

30

© 2017, Jamie Callan

Statistical Properties of Text: Zipf's "Law"

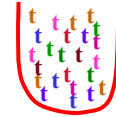
Zipf's Law relates a term's frequency to its rank

- **MLE probability of observing term t in corpus C**

$$P(t) = \frac{ctf_t}{N}$$

ctf_t : Collection term frequency – how often t occurs in C

N : Total word occurrences in corpus C



- **MLE probability of observing the R 'th ranked term**

– Terms are rank in descending order of frequency

$$P(t_R) = \frac{ctf_{t_R}}{N}$$

So far, this is obvious math

1.	t_1	ctf_1
2.	t_2	ctf_2
3.	t_3	ctf_3
:	:	:

31

© 2017, Jamie Callan

Statistical Properties of Text: Zipf's "Law"

Empirical observation: $P(t_R) = \frac{ctf_{t_R}}{N} \approx \frac{A}{R}$ $A \approx 0.1$ for English

- **So, what is the probability of the 4 most frequent terms?**

– $P(t_1) = 0.1 / 1 = 0.100$ (10% of the collection)

– $P(t_2) = 0.1 / 2 = 0.050$ (5% of the collection)

– $P(t_3) = 0.1 / 3 = 0.033$ (3.3% of the collection)

– $P(t_4) = 0.1 / 4 = 0.025$ (2.5% of the collection)

- **The 4 most frequent terms are 20.8% of word occurrences**

- **The 50 most frequent terms are 45% of word occurrences**

Note: We don't need to know the collection size, contents, ...

32

© 2017, Jamie Callan

Statistical Properties of Text: Zipf's "Law"

Empirical observation: $P(t_R) = \frac{ctf_{t_R}}{N} \approx \frac{A}{R}$ $A \approx 0.1$ for English

Simple manipulation provides a different viewpoint

$$\begin{aligned} \frac{ctf_{t_R}}{N} &= \frac{A}{R} \\ R \times ctf_{t_R} &= A \times N \end{aligned}$$

Rank \times Frequency = Constant (Constant = $0.1 \times N$) **Zipf's Law**

Both forms are useful

- **Method 1:** $P(t_R) \approx \frac{0.1}{R}$
- **Method 2:** Rank \times Frequency = $0.1 \times N$

33

© 2017, Jamie Callan

Statistical Properties of Text: Zipf's "Law"

In a 100,000 word corpus, how often is the most frequent (rank = 1) term expected to occur?

- **Method 1:** $P(t_1) = \frac{0.1}{1} = 10\%$ **so, 10,000 occurrences**
- **Method 2:** $1 \times ctf_{t_1} = 0.1 \times 100,000$ **so, 10,000 occurrences**

34

© 2017, Jamie Callan

Zipf's Law on WSJ '87-92

Rank	Term	Predicted	Actual	Rank	Term	Predicted	Actual
1	the	7,831,076	4,352,160	101	9	77,535	80,490
2	of	3,915,538	2,134,125	102	most	76,775	80,409
3	to	2,610,358	2,023,402	103	such	76,030	80,037
4	a	1,957,769	1,811,373	104	time	75,299	80,014
5	in	1,566,215	1,546,782	105	No	74,582	78,459
6	and	1,305,179	1,507,140	106	into	73,878	78,208
7	s	1,118,725	855,190	107	only	73,188	78,150
8	that	978,885	787,792	108	trading	72,510	78,133
9	for	870,120	780,138	109	many	71,845	77,578
10	is	783,107	605,988	110	so	71,192	77,099
11	said	711,915	528,481	111	now	70,550	76,281
12	it	652,590	510,102	112	based	69,920	75,798
:	:	:	:	:	:	:	:

35

© 2017, Jamie Callan

Statistical Properties of Text: Zipf's "Law"

In a 100,000 word corpus, what is the expected rank of the last term that occurs 50 times?

- Method 2:** $R = \frac{0.1 \times 100,000}{50} = 200$

– i.e., it is expected to be the 200th most frequent term

36

© 2017, Jamie Callan

Statistical Properties of Text: Zipf's "Law"

In a 100,000 word corpus, how many terms occur 50 times?

- Rank of last term that occurs 51 times $R = \frac{0.1 \times 100,000}{51} = 196.1$
- Rank of last term that occurs 50 times $R = \frac{0.1 \times 100,000}{50} = 200$
- Take the difference
 - About 4 terms occur 50 times
 - The answer is essentially the same if we use 50 and 49

37

© 2017, Jamie Callan

Statistical Properties of Text: Zipf's "Law"

In a 100,000 word corpus, what is the rank of the last term?

- Method 2: $R = \frac{0.1 \times 100,000}{1} = 10,000$
 - i.e., it occurs at rank 10,000

You could also have used Heap's Law for this

- Heap's Law is more accurate
- In some situations, Zipf's Law is more convenient

38

© 2017, Jamie Callan

Statistical Properties of Text: Zipf's "Law"

In a 100,000 word corpus, what proportion of terms occur once?

- Rank of last term that occurs 2 times $R = \frac{0.1 \times 100,000}{2} = 5,000$
- Rank of last term that occurs 1 time $R = \frac{0.1 \times 100,000}{1} = 10,000$
- Do the math...
 - The vocabulary contains about 10,000 terms
 - About $10,000 - 5,000 = 5,000$ terms occur just once
 - Thus about 50% of the vocabulary occurs just once

39

© 2017, Jamie Callan

Statistical Properties of Text: Zipf's "Law"

What proportion of terms occur n times?

- Vocabulary size: $0.1 \times N$
- Number of terms that occur n times: $\frac{0.1 \times N}{n} - \frac{0.1 \times N}{n+1}$
- Do the math...

$$\left(\frac{\frac{0.1 \times N}{n} - \frac{0.1 \times N}{n+1}}{\frac{0.1 \times N}{1}} \right) = \frac{1}{n} - \frac{1}{n+1} = \frac{1}{n(n+1)}$$

Note: Independent of the corpus or its size

ctf	Proportion
1	0.500
2	0.167
3	0.083
4	0.050
5	0.033
≤ 5	0.833

40

© 2017, Jamie Callan

What Does Zipf's Law Tell Us?

A few terms are very common...

- Most frequent term is 10% of all tokens
- Most frequent 5 terms are 23% of all tokens →
- Most frequent 100 terms are 52% of all tokens

Rank	Proportion
1	0.100
2	0.050
3	0.033
4	0.025
5	0.020

Most terms are very rare...

- 50% of the terms occur once
- 91% of the terms occur fewer than 10 times →

Freq	Proportion
1	0.500
2	0.167
3	0.083
4	0.050
5	0.033

41

© 2017, Jamie Callan

Practical Uses of Heaps' and Zipf's Laws

These “laws” allow system designers to estimate the amount of storage needed for important data structures

- Size of term dictionary, distribution of inverted list sizes
- Estimates are independent of content type, language, ...

Inverted list sizes vary significantly

- The longest usually covers about 10% of all term occurrences
- About 50% of the vocabulary occurs just once
- About 83% of the vocabulary occurs fewer than 5 times
- Software must be good at handling objects of different sizes

42

© 2017, Jamie Callan

Introduction to Shallow Language Processing

There are stable, language-independent patterns in how people use human languages

These properties hold in a wide range of languages

- English, Spanish, German, Mandarin, Japanese, Arabic, ...

The most frequent words in one corpus may be rare in another corpus

- Typically “the” is the most common word in an English corpus

43

© 2017, Jamie Callan

Introduction to Shallow Language Processing

Search engines typically focus on term frequency counts

- E.g., term frequency in a document ($tf_{t,d}$)
- E.g., the number of documents that contain a term (df_t)
- This allows them to ignore many aspects of language usage
 - E.g., syntax and semantics

This approach works because term frequency is skewed predictably

44

© 2017, Jamie Callan

Introduction to Shallow Language Processing

Variations in term frequency make it difficult to construct good queries for exact-match retrieval models

- A rare word in an AND causes few or no documents to match
- A common word in an OR causes many documents to match

Most people do not have good intuitions about which words are frequent or rare in a specific corpus

- Thus queries may not behave as expected

This is one reason why search engines evolved from exact-match retrieval models to best-match retrieval models

45

© 2017, Jamie Callan

Outline

What this course is about

- Course philosophy
- Overview of course topics
- Administrative information
- Questions and answers

Introduction to shallow language processing

- The bag of words model
- Heaps' Law
- Zipf's Law

46

© 2017, Jamie Callan

One Last Thing ... The Waitlist

There is room for 32 more students

- But 61 people are on the waitlist

The course web page explains the waitlist priority policy

- If you are on the waitlist, read it

Priority is given to students that attend class

- Usually every student that attends the first two classes gets in
- Sign the attendance sheet at the end of each lecture
 - i.e., today and every day until you are admitted to the class