11-442 / 11-642 / 11-742:
Search Engines

Introduction

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

1

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

2

Course Philosophy

This course is about search engine <u>theory</u> and <u>practice</u>
- Beautiful theory, and ugly-but-effective heuristics
  - This field is developing rapidly, so theory is often insufficient
- The emphasis is on <u>what works</u>

This course covers <u>enterprise,</u> <u>web,</u> and <u>vertical</u> search
- They have many similarities
- They also have important differences
- All three are important commercially

3 © 2021, Jamie Callan

3

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 © 2021, Jamie Callan

4

Overview of Course Topics

- Text representation
- Search engine indexes
- Index construction
- Query structure
- Document structure
- Unsupervised ranking
- Feature-based ranking
- Neural ranking

- Page features
- Evaluation
- Search log analysis
- Diversity
- Personalization
- Federated & vertical search

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

5

Administrative Information: Teaching Assistants

Vinay Damodaran
vdamodar@andrew

Ziqi Deng
ziqideng@andrew

Harsh Jain
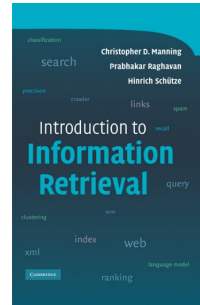harshjai@andrew

Ran Yao
rany2@andrew

6

Administrative Information:
Textbook

Textbook

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

An online version is available

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

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

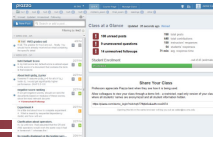**Access to some files is restricted to .cmu.edu**

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

Administrative Information: Course Discussion Forum

There is a Piazza discussion forum
- https://piazza.com/cmu/spring2021/1142

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

© 2021, Jamie Callan

9

Administrative Information: Canvas

We will use Canvas for three purposes
1. To send email to you
2. For short quizzes
3. To report homework and quiz grades to you

Everything else is done via the course website or Piazza

10

© 2021, Jamie Callan

10

Administrative Information:
Homework Descriptions

Homework assignments (6): Develop software, use it to run experiments
- Implement two exact-match Boolean retrieval algorithms
- Implement two best-match retrieval algorithms
- Implement pseudo relevance feedback
- Implement features and test learning to rank (LeToR)
- Implement neural reranking (?)
- Implement diversified ranking algorithms

11

Administrative Information:
Homework Policies

All homework must be submitted via the course website
- Due by 11:59 pm of the due date (Pittsburgh timezone)

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

12

Administrative Information: Programming Skills

This course requires <u>good</u> Java programming skills

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

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

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

© 2021, Jamie Callan

13

Quizzes

Quizzes (6): Short tests for material covered in class & readings

- 20 minute quizzes conducted through Canvas
- The time of the quiz depends on your timezone
    - EST, CST, and PRT: During class
    - PST, JST, and CTT: During class or at the alternate time
- Alternate time
    - TBD once I know who is in the course
    - Something like 9:00 pm Pittsburgh time

© 2021, Jamie Callan

14

Administrative Information:
Grading

Homework (65%)

- 5 assignments × 11% per assignment +
  1 assignment × 10% per assignment = 65%

Quizzes (35%)

- There are 6 quizzes
  - Drop the lowest score
  - 5 quizzes × 7% each = 35%

© 2021, Jamie Callan

15

Recommended Effort

Task	Hours / Week 11-642 11-742	Hours / Week 11-442
Attend class	3	3
Readings	1-2	1-2
Homework programming (× 2 weeks)	5	5
Homework experiments + report (× 2 weeks)	3	
Total	**12-13**	**9-10**

Your experience may vary, but this is how the course is designed

© 2021, Jamie Callan

16

Administrative Information: Class Participation

I strongly recommend against multitasking during class

- I spend a lot of time in Zoom meetings
- I understand the appeal of multitasking
- Data: Students that pay attention in class tend to get better grades

17

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
  - Raise your hand or post a question to chat

Do you have any questions right now?

18

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

19

Introduction to
Shallow Language Processing

A Great Choice.
Review by topjimmy5150
★ ★ ★ ★ ★ April, 21 2003

I have been looking and looking for a new camera to replace our bulky, but simple and reliable (but only fair picture taker) Sony Mavica FD73. My other choice (Besides the more expensive Nikon Coolpix 3100) was the (also more expensive) Sony Cybershot P72. I recommend any of these cameras, and I was set to buy the Sony, but at the last minute I cheaped out and bought the 2100. No regrets. I bought the camera (along with 128mb memory card (the stock 16mb card will be kept in the bag as a spare) and carrying case) at the new Best Buy in Harrisburg, PA. I also bought a set of 4 Nickle-Metal Hydride rechargable batteries and charger at Walmart for less than $20. I keep 2 in the camera and two in the charger/in the camera bag along with the original Lithium battery pack as spares.

Hands down, the best feature of this camera is it's compact design. It is very small. My family likes to go camping during the summer, and last year we found the Mavica too
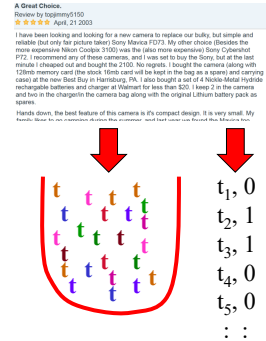
(topjimmy5150, Epinions.com)

20

Introduction to
Shallow Language Processing

Search engines use a shallow form of language understanding

- Discard word order
- Discard some words (e.g., "the")
- Transform words into <u>terms</u>
  - Map multiple words to the same term
    - » e.g., "cat", "cats"
  - …

The result is a <u>bag of words</u> (or feature vector)

Why does it work?

$t_1, 0$
$t_2, 1$
$t_3, 1$
$t_4, 0$
$t_5, 0$
: :

21

© 2021, Jamie Callan

21

Introduction to
Shallow Language Processing

Search engines consider two <u>types</u> 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

22

© 2021, Jamie Callan

22

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

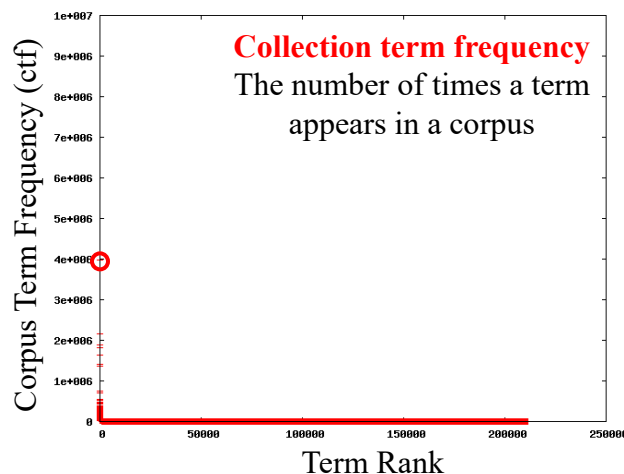**Tokens:**
Word occurrences

Types:
Unique words

23

© 2021, Jamie Callan

23

Term Frequency in the WSJ

Collection term frequency
The number of times a term appears in a corpus

Wall Street Journal (1987-1992)

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

Tokens:
Word occurrences

Types:
Unique words

24

© 2021, Jamie Callan

24

Term Frequency in the WSJ

Collection term frequency
The number of times a term appears in a corpus

Corpus Term Frequency (ctf)

Term Rank

Half of the vocabulary

Wall Street Journal (1987-1992)

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

Tokens:
Word occurrences
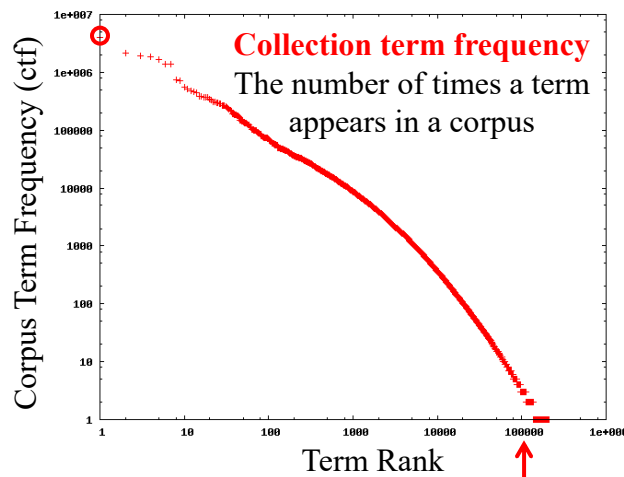
Types:
Unique words

25

© 2021, Jamie Callan

25

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

26

© 2021, Jamie Callan

26

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

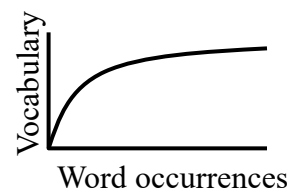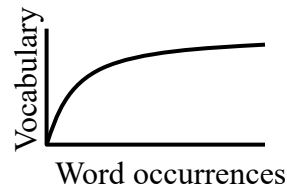## Statistical Properties of Text: Heaps' "Law"

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

Vocabulary vs. Word occurrences

The vocabulary never stops growing

- Misspellings, names, new words, …

Statistical Properties of Text: Heaps' "Law"

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

$V=KN^{\beta}$

K : usually $10 \leq \boxed{K} \leq 100$

β : usually $0.4 \leq \boxed{\beta} \leq 0.6$ for English

N : total number of word occurrences

Is it a good fit?
- WSJ 87-93, Predicted: $V = \boxed{25} \times (69,000,000)^{\boxed{0.5}} = 208K$ terms
- WSJ 87-93, Actual: $V = 211K$ terms

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

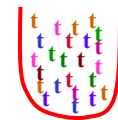## Statistical Properties of Text: Zipf's "Law"

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**
    – Rank terms 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$_3$
    :   :   :

Page 15

Statistical Properties of Text: Zipf's "Law"

Empirical observation: $\qquad P(t_R) = \dfrac{ctf_{t_R}}{N} \approx \dfrac{A}{R} \qquad A \approx 0.1$ 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, …

31

Statistical Properties of Text: Zipf's "Law"

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

Simple manipulation provides a different expression

$$\frac{ctf_{t_R}}{N} = \frac{A}{R}$$

$$R \times ctf_{t_R} = A \times N$$

$$\text{Rank} \times \text{Frequency} = \text{Constant} \qquad (\text{Constant} = 0.1 \times N) \qquad \textbf{\textcolor{red}{Zipf's Law}}$$

Both forms are useful

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

32

Statistical Properties of Text:
Zipf's "Law"

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

- **Method 1:** $P(t_1) = \dfrac{0.1}{1} = 10\%$ **so, 10,000 occurrences**

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

33

Statistical Properties of Text:
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	1118725	855,190	107	only	73,188	78,150
8	that	978,885	787,792	108	trading	72510	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
⋮	⋮ ⋮	⋮ ⋮ ⋮ ⋮	⋮ ⋮ ⋮ ⋮	⋮	⋮ ⋮	⋮ ⋮ ⋮	⋮ ⋮ ⋮

34

Statistical Properties of Text:
Zipf's "Law"

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

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

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

35

Statistical Properties of Text:
Zipf's "Law"

In a 100,000 word corpus, <u>how many terms</u> occur 50 times?

- **Rank of last term that occurs 51 times** $\qquad R = \dfrac{0.1 \times 100,000}{51} = 196.1$

- **Rank of last term that occurs 50 times** $\qquad R = \dfrac{0.1 \times 100,000}{50} = 200$

- **Take the difference**
  - $3.9 \approx$ About 4 terms occur 50 times
  - The answer is essentially the same (4.1) if we use 50 and 49

36

Statistical Properties of Text:
Zipf's "Law"

In a 100,000 word corpus, <u>what is the rank of</u> the last term?

- **Method 2:** $R = \dfrac{0.1 \times 100,000}{1} = 10,000$

  – The last term occurs just once, and 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

© 2021, Jamie Callan

37

Statistical Properties of Text:
Zipf's "Law"

In a 100,000 word corpus, <u>what proportion</u> of terms occur once?

- **Rank of last term that occurs 2 times** $R = \dfrac{0.1 \times 100,000}{2} = 5,000$

- **Rank of last term that occurs 1 time** $R = \dfrac{0.1 \times 100,000}{1} = 10,000$

- **Do the math…**
  – About 10,000 – 5,000 = 5,000 terms occur just once
  – The vocabulary contains about 10,000 terms
  – Thus about 50% of the vocabulary occurs just once
    » This is an over-estimate (but convenient)

© 2021, Jamie Callan

38

Statistical Properties of Text: Zipf's "Law"

What <u>proportion</u> 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…**

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

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

Statistical Properties of Text: What Does Zipf's Law Tell Us?

A <u>few</u> terms are very common…
- Most frequent term is <u>10%</u> of all tokens
- Most frequent 5 terms are <u>23%</u> of all tokens
- Most frequent 100 terms are <u>52%</u> of all tokens

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

<u>Most</u> terms are very rare…
- <u>50%</u> of the terms occur once
- <u>83%</u> of the terms occur fewer than <u>5 times</u>
- <u>91%</u> of the terms occur fewer than <u>10 times</u>

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

Statistical Properties of Text:
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

41

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

42

**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., term frequency in a document collection (ctf_t)
- 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 <u>predictably</u>

43

**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 <u>in a specific corpus</u>
- 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

44

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

45