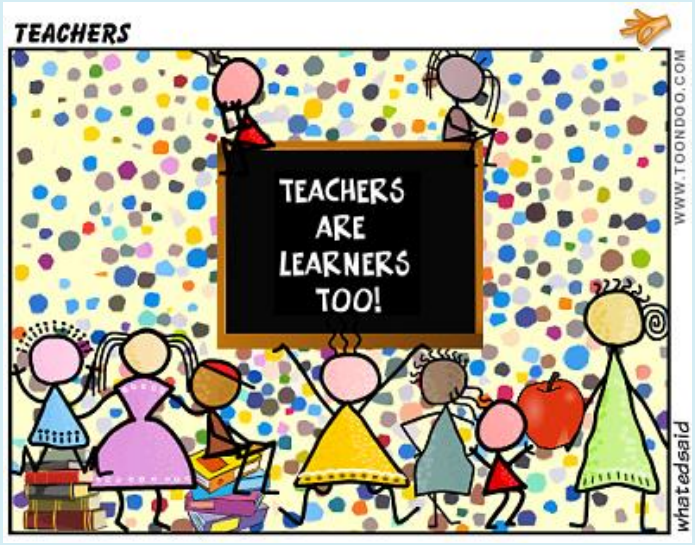
CMPS 143
Introduction to Natural Language Processing

Natural Language and Dialogue Systems
http://nlds.soe.ucsc.edu

So, who is "Staff"?

CMPS 130 - 01	Computational Model	LEC	TuTh	02:00PM-03:45PM	Tantalo,P.	▲	175	175
CMPS 143 - 01	Natural Lang Process	LEC	TuTh	10:00AM-11:45AM	Staff	●	30	27
CMPS 160 - 01	Intro Comp Graphics	LEC	MWF	09:30AM-10:40AM	Pang,A.	●	129	88

- Instructor: Elahe Rahimtoroghi (elahe@soe.ucsc.edu)
  - PhD Candidate in Computer Science (4th year)
  - Office hours: Thursday 2-3:30 – E2 255
  - Advisor: Prof. Marilyn Walker
  - Natural Language and Dialogue Systems (NLDS) Lab



Tutor and Lab Sections

- Tutor: Jiaqi Wu (jwu64@ucsc.edu)
  - PhD students in Computer Science
  - Great with Python and NLP toolkit (NLTK)!

- Sections:
  - Get help on the assignments and projects
  - **Tuesday – Thursday**
  - **4:00 – 6:00 pm**
  - **Social Sciences I Mac Lab (Room 135)**

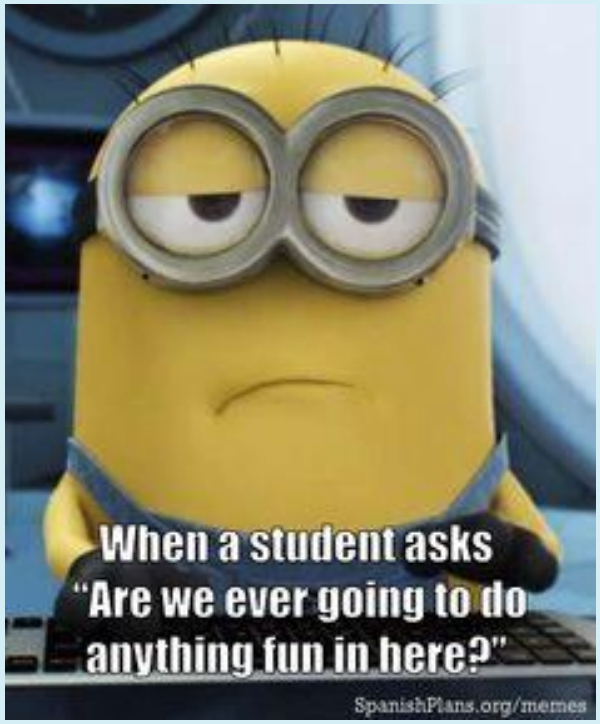
What is Natural Language Processing

- The field of natural language processing (NLP) develops theories and computational models of natural language data (the language that humans speak and write).
  - **Natural Language Understanding (NLU)**
  - **Natural Language Generation (NLG)**

Why is NLP growing now?

- **So much NLP data**: online forum discussions, posts on Facebook, LinkedIn and other social sites, books, journal and newspaper articles, phone calls, radio broadcasts, YouTube user generated videos or public lectures, web pages, weblogs, tweets, user reviews, email, and comment streams on newspaper or magazine articles.
- **So many NLP applications**: Education, Economics, Digital Humanities, Biological, Environmental, Medical and Health Informatics, Conversational Interfaces, Search, Indexing, Business intelligence, Translation, Computational advertising.
- **So many interested parties:** Research, companies and government funding agencies, as well as an area of inherent theoretical interest.

Why are you taking this class?



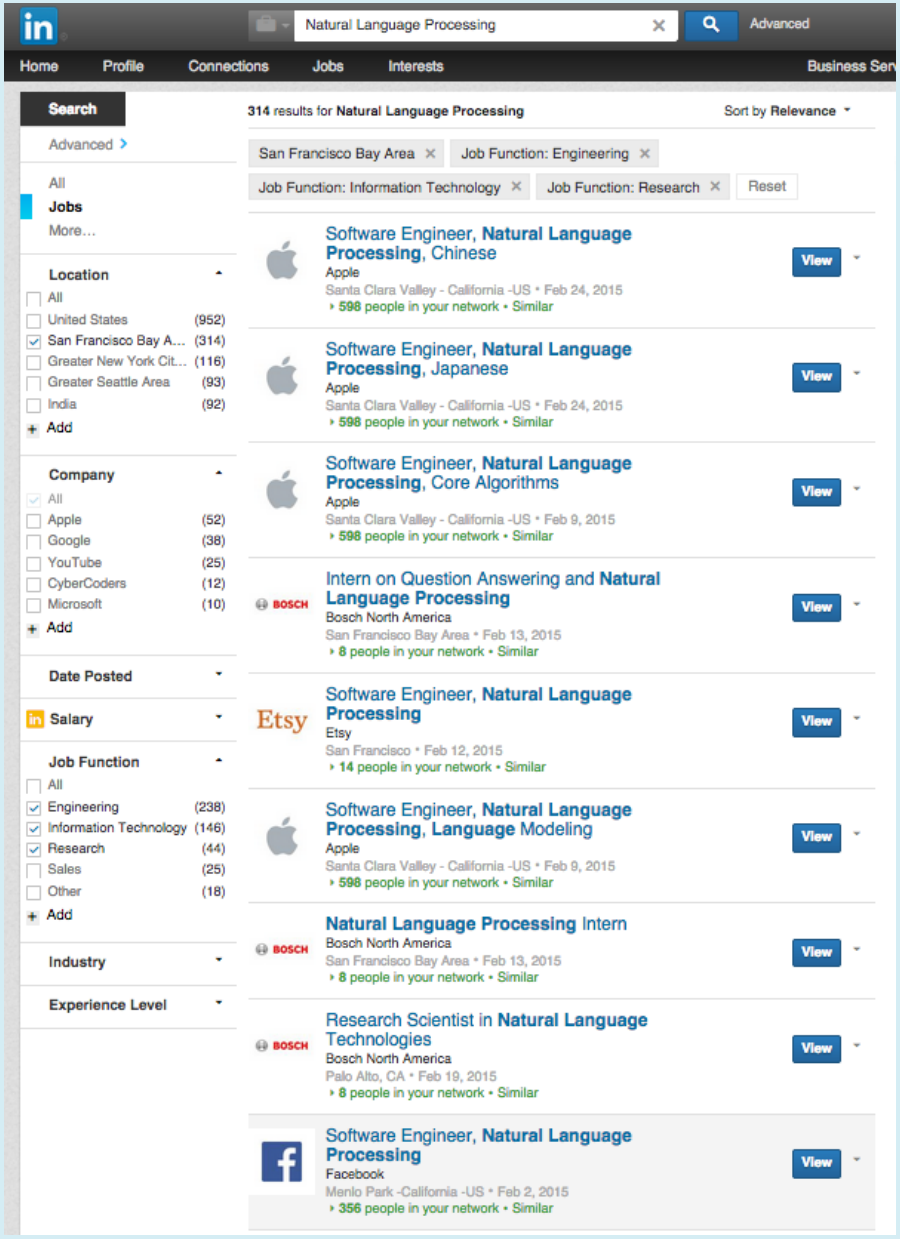
Before I tell you
why you want to

Why should NLP be interesting to you?

- **Most of what we think of as communication is done using language**
  - Texting, Talking, Facebook, Twitter, Email, Blogging
- **It is not a solved problem**. There is lots of room for innovation.
- **There are many many applications**: multiple different types of analysis of any of the data above: sentiment, structure, meaning, words
- **Skills needed to work on NLP data are useful for many other different kinds of tasks**: (scripting, data processing and cleaning, setting up experiments and testing methods, python, learning from data, statistical modeling)

Why is NLP Interesting?

- **There are lots of jobs!**
- Over 300 jobs within 100 miles of UCSC 95064 on Feb 15th, 2015



What kinds of NLP jobs are there?

- So many different applications for NLP

Siri Language Engineer (Cloud Services Localization) - British English
Apple - Santa Clara Valley - California -US
Posted 9 hours ago

Analytical Linguist, Knowledge Engine
Google - San Francisco, CA, USA
Posted 8 days ago

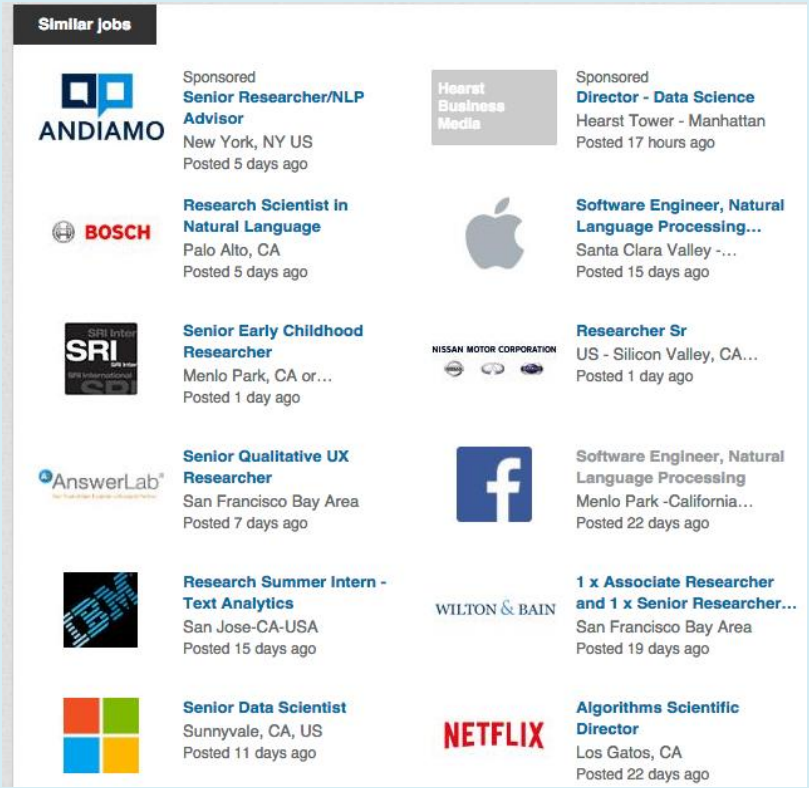
Software Engineer, Natural Language Processing
Facebook - Menlo Park -California -US
Posted 21 days ago

173 Jobs results for analytical linguist

Analytical Linguist, Knowledge Engine
Google

Bilingual Visual Data Evaluator
Leapforce

Korean Translator
Z-Axis Tech Solutions Inc



Let's look at a sample job

About this job

Job description

Facebook was built to help people connect and share, and over the last decade our tools have played a critical part in changing how people around the world communicate with one another. With over a billion people using the service and more than fifty offices around the globe, a career at Facebook offers countless ways to make an impact in a fast growing organization.

Facebook is seeking an Natural Language Processing Engineer to join our engineering team in Menlo Park. The ideal candidate will have industry experience solving language-related problems using statistical methods on vast quantities of data. Individuals in this role should be experts in machine learning and have experience working on machine translation, word-sense disambiguation, topic modeling, etc. The candidate will help Facebook build products that support idiomatic user input and expression in more than 70 languages, for products such as Open Graph, News Feed, and Search.

Responsibilities

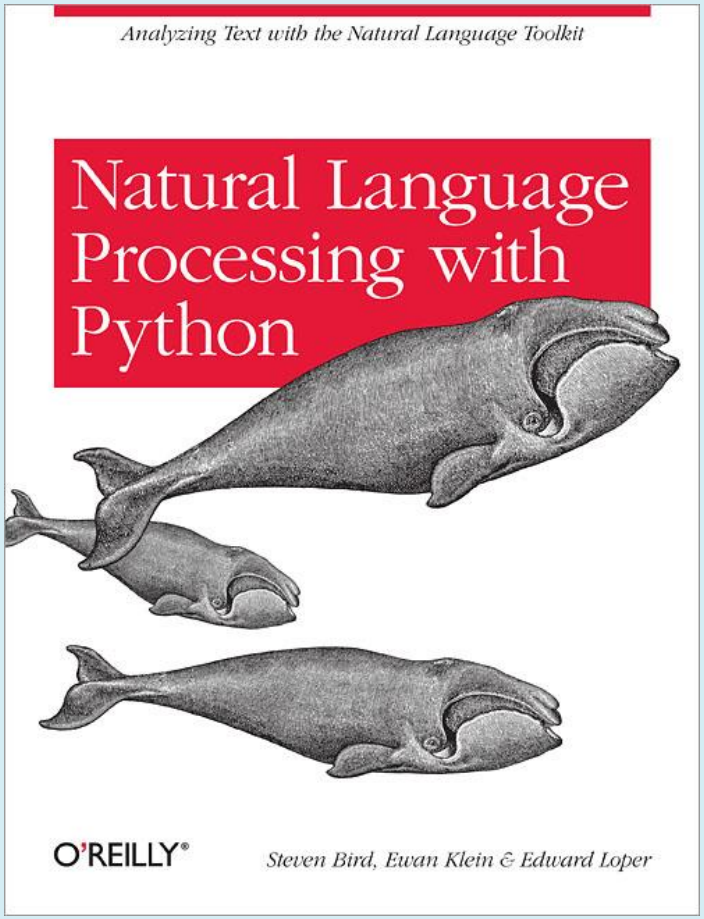
- Create language models from petabytes of text data in different languages using Hadoop/Hive
- Work closely with product teams to implement algorithms that power user and developer-facing products
- Be responsible for measuring and optimizing the quality of your algorithms

Requirements

- Strong desire to build beautiful, expressive products that delight users in any language
- M.S. or Ph.D. in Computer Science, Machine Learning or NLP
- Industry experience preferred
- Experience with Hadoop/Hbase/Pig or Mapreduce/Sawzall/Bigtable a plus
- Experience with scripting languages such as Perl, Python, PHP, and shell scripts
- Fluency in at least 2 natural languages is a plus

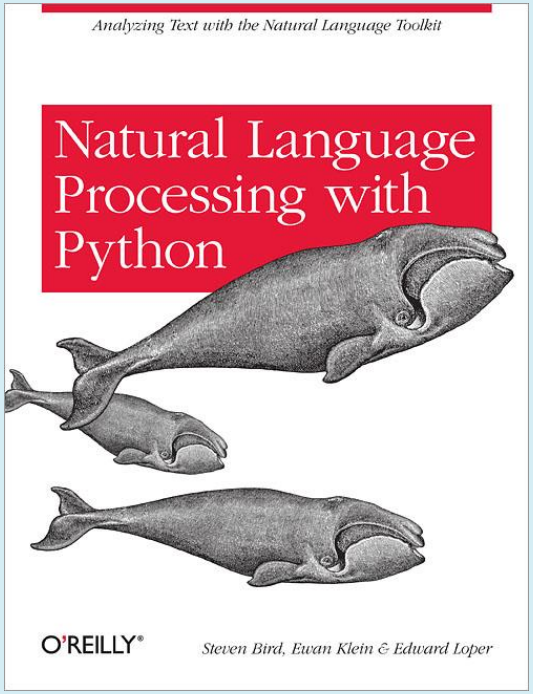
Goals of the Class

- Learn basics of practical NLP
- Start with reading in some language data from a file or the web, end with producing a "representation of its meaning"
- Learn about different NLP applications
- Using NLTK (Natural Language Processing with Python)
- Understand the methods and challenges of the NLP field



Expectations

- You know (some) Python.
- Using NLTK (Natural Language Processing with Python)
  - Goes step by step, but we do not spend time on python itself in class
- Final project: Q&A
  - Being able to find the answers and then answer questions from stories and other types of text using NLP methods
- Simplified for this class, but not that different than the real world



Focus of this Class: Apply NLP to narrative language

- Lots of types of NLP data have narrative structure: News, Stories, Weblogs, posts on Facebook
- Focus on **understanding** language (as opposed to generating it)
- Homework on distributional properties of language, language models, patterns and structures in language, sentiment analysis, text classification
- Last 3 HWs: Project focused on Question Answering:
  - Good example of application
  - Illustrates challenges
  - Only statistical methods is not enough: Impossible to provide a good answer to a WHY question without deep understanding of text

Types of Narrative Data

A Crow was sitting on a branch of a tree with a piece of cheese in her beak when a Fox observed her and set his wits to work to discover some way of getting the cheese. Coming and standing under the tree he looked up and said, "What a noble bird I see above me! Her beauty is without equal, the hue of her plumage exquisite. If only her voice is as sweet as her looks are fair, she ought without doubt to be Queen of the Birds."

The Crow was hugely flattered by this, and just to show the Fox that she could sing she gave a loud caw. Down came the cheese, of course, and the Fox, snatching it up, said, "You have a voice, madam, I see: what you want is wits."

Answering Questions

- Where was the Crow sitting?
- What did the Crow have?
- Why did the Fox want the cheese?
- Did the Fox deliberately flatter the Crow?
- Why did the Fox flatter the Crow?
- Why did the Fox say the Crow needs wits?
- Is the Crow a noble bird?

Narrative Structure in News

Mob violence March 22 sparked by the death of a 14-year-old stabbed on a TARC bus earlier this month prompted a spate of meetings with city officials and community leaders this week.

Just after 7 p.m. on March 22 a band of teenagers who had been at the waterfront attacked a 13-year-old girl because they wanted her sneakers, police said. A 40-year-old man came to her defense and was pummeled by the mob.

Within two hours, 31 people called to report trouble, Chief Steve Conrad said. Groups of teenagers roved the area, looting a store, vandalizing cars and assaulting passers-by. The police department counted 20 criminal incidents within hours downtown.

Answering Questions

- Who was attacked?
- Who was the attacker?
- What time did the attack happen?
- Why was the girl attacked?
- What did the man do?
- Was the man a hero?
- Why did the man get pummeled by the mob?

Blogs Corpus: Protests, Doctor visits etc

Millions of stories about everyday life told by ordinary people

Topic Browser

Browse & The Topic Collections

Topic	Creator	Created On	Stories	Relevance
Car Accident and Close Calls	Lena	11/06/2014	21	
Christmas	Lyn	08/19/2014	100	
Collection of Negative Emotion Sto	Lena	11/06/2014	368	
Collection of Positive Emotion Stori	Lena	11/06/2014	344	
Collection of Sad Tag stories	Lyn	11/05/2014	128	
Divorce	Lena	10/29/2014	8	
Easter Hunting Eggs and Baskets	Lyn	08/20/2014	19	
Failed Crushes	Casey	08/11/2014	14	
Family Reunion	Lena	11/04/2014	7	
Farewell Parties	Lena	11/02/2014	2	
Freighthopping	kevin	08/27/2014	3	

BLOGS Corpus: Funny & Serious story

This is one of those times I wish I had a digital camera. We keep a large stainless steel bowl of water outside on the back deck for Benjamin to drink out of when he's playing outside. His bowl has become a very popular site. Throughout the day, many birds drink out of it and bathe in it. The birds literally line up on the railing and wait their turn. Squirrels also come to drink out of it. The craziest squirrel just came by- he was literally jumping in fright at what I believe was his own reflection in the bowl. He was startled so much at one point that he leap in the air and fell off the deck. But not quite, I saw his one little paw hanging on! After a moment or two his paw slipped and he tumbled down a few feet. But oh, if you could have seen the look on his startled face and how he jumped back each time he caught his reflection in the bowl!

Structure of the Class
CMPS 143

Class Web Page and Resources

- **Syllabus:**
  - https://cmps143-spring1601.courses.soe.ucsc.edu/node/3
  - Will be updated every week

- eCommons & Piazza
  - Assignments and discussions

Book and other Online Resources

Book http://www.nltk.org/book/

Natural Language Processing with Python

– Analyzing Text with the Natural Language Toolkit

Steven Bird, Ewan Klein, and Edward Loper

The NLTK book is currently being updated for Python 3 and NLTK 3. This is work in progress; chapters that still need to be updated are indicated. The first edition of the book, published by O'Reilly, is available at http://nltk.org/book_1ed/. A second edition of the book is anticipated in early 2016.

Grading Policy

- Attendance: 5%
- Homeworks and discussion in class: 45%
- Project (assignments that include project, and final presentation of project during Finals slot): 25%
- Midterm: 25%
- Final: 25%
- Homework Delivery: Turn it in on eCommons assignments. Please include any code, files, and written documents in a zip file. Written documents should be plain text or PDF only. Multiple uploads (to overwrite) are enabled. Late HW accepted until noon the next day with a 10% penalty.

GROUND RULES:

Do your homework (not at the last minute)!

Go to Section whenever you need!

Ask questions, talk in class!

Slow me down when I go too fast!

Important Information

- I expect you to attend class, and let me know if you need to miss class for any reason.
- You will need to use the Tutor lab sections
- Weekly homework. After the midterm these start ramping up to being phases of your project development and testing
  - You can work with a partner on the project or on your own
  - Project involves ‘competition’ between teams
- First several chapters of book are simple methods
- We will go through the first four chapters very quickly
- Don’t let yourself get behind in the first two weeks

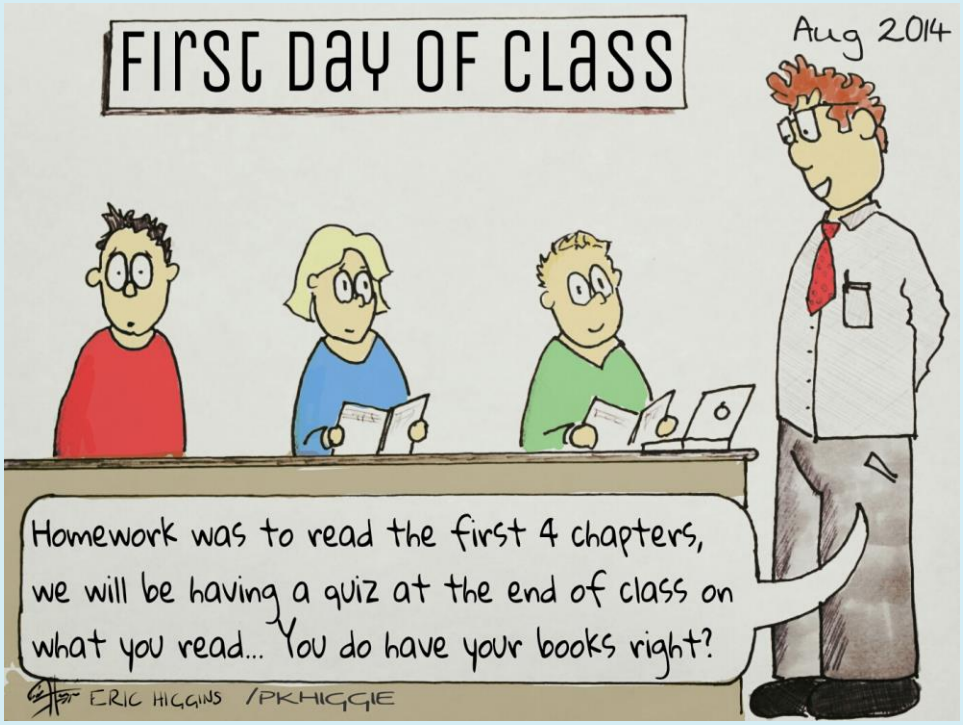
Important Information

- Special Accommodations:
  - *If you have special needs, we will accommodate you. The* **Disability Resource Center** *offers services that are confidential and free of charge.*
  - http://drc.ucsc.edu/
  - https://intranet.soe.ucsc.edu/DisabilityResource

1. Students contact the DRC to determine their eligibility
2. Students then notify their instructor and provide their Accommodation Authorization form.
3. Please note that it is the student's responsibility to contact the instructor about their accommodations.
4. Students should submit their requests to faculty no later than 7 days before a regular exam and 14 days before a final exam.

Main Goals for Week 1

HW 0: Entry Quiz
GET PYTHON AND NLTK INSTALLED
READING IN TEXTS
PROCESSING WORDS
TAGGING WORDS
LOOKING AT PATTERNS (BIGRAMS, POS, PHRASES)



Homework 0 already posted!

- **Homework 0:** Entry Quiz
  - **Due in two days: Thursday, March 31, 10:00 AM (beginning of class)**
  - Let us find out what you remember from your discrete math class
  - Doesn't count towards your grade
- **Homework 1:**
  - **Will be posted on Thursday**
  - **Due next Friday (April 8)**
  - First set up your environment, then do some NLP
  - Material covered this week

Goals for Today

WHAT IS NLP?

OVERVIEW OF THE SYLLABUS

INSTALLING PYTHON AND NLTK: Chapter 1 of NLPP

READING IN TEXTS

PROCESSING WORDS

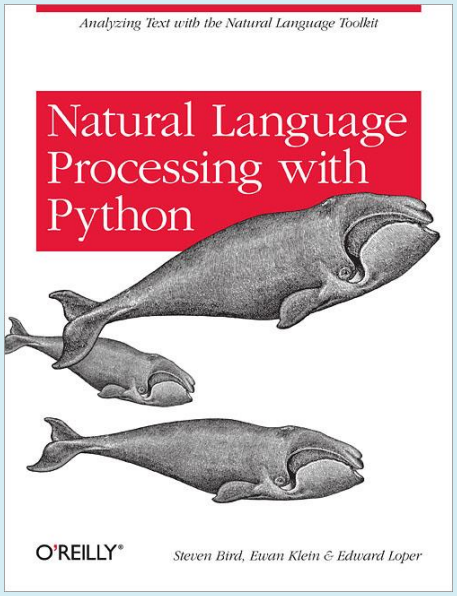
Each week both theory, ideas and practice
Lots of programming

NLTK:
Natural Language Toolkit

NLPP:
Natural Language Processing with Python

NLTK (http://www.nltk.org)

- NLTK is a leading platform for building Python programs to work with human language data.
  - open source Python modules, datasets and tutorials for NLP applications
- Designed for ease of use, code readability and teaching
- Actively developed with dozens of contributors



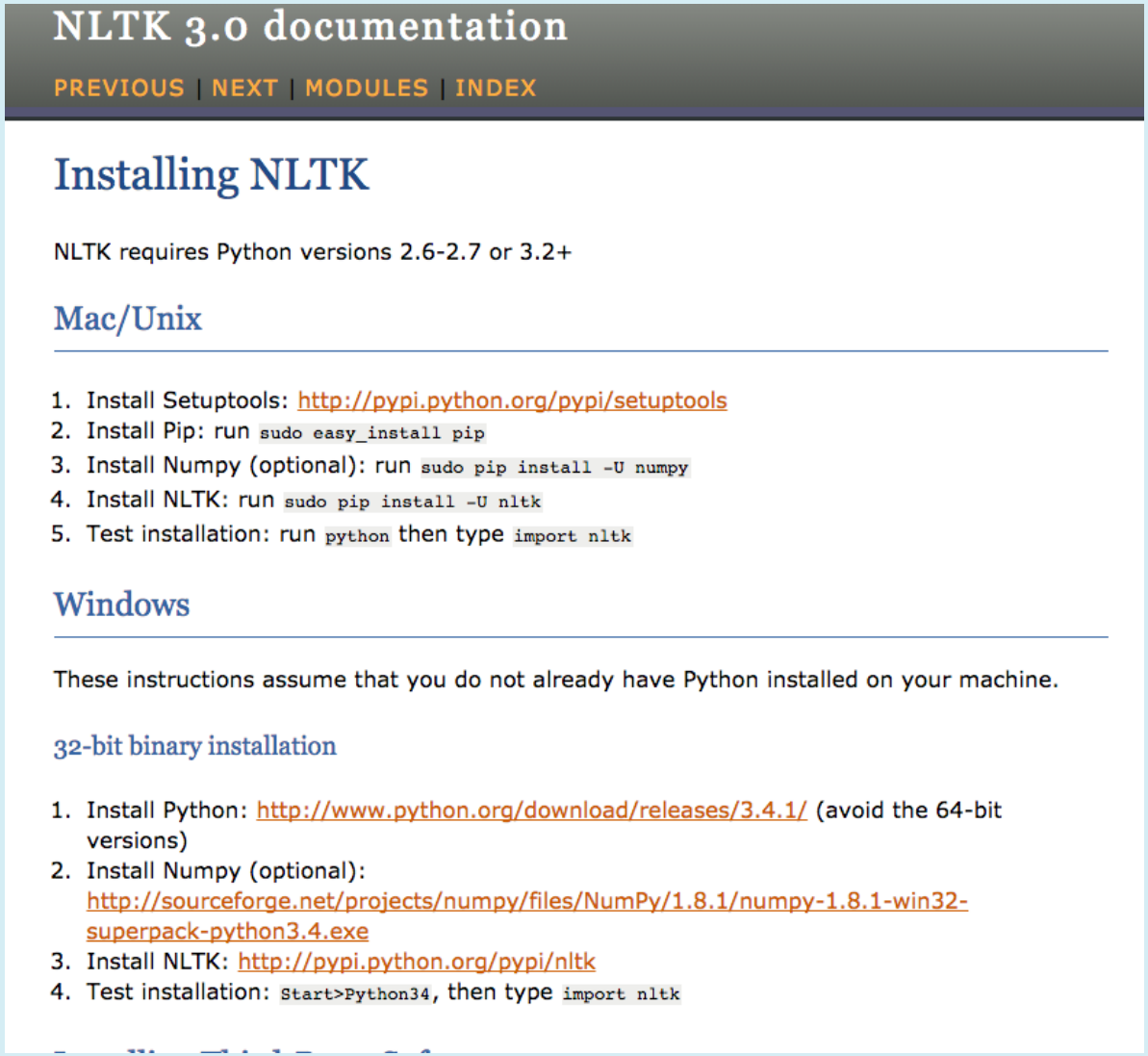
Edward Loper, Ewan Klein, and Steven Bird, Stanford, July 2007

Components

- **Code**:
  - Suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, …
- **Corpora**:
  - Over 50 corpora and lexical resources such as WordNet
- **Documentation**: a 400-page book (NLPP), articles, reviews, API documentation
  - Suitable for linguists, engineers, students, educators, researchers, and industry users

Do it today: Installing Python 3 and NLTK 3

- http://www.nltk.org/install



NLTK 3.0 documentation

PREVIOUS | NEXT | MODULES | INDEX

Installing NLTK

NLTK requires Python versions 2.6-2.7 or 3.2+

Mac/Unix

1. Install Setuptools: http://pypi.python.org/pypi/setuptools
2. Install Pip: run `sudo easy_install pip`
3. Install Numpy (optional): run `sudo pip install -U numpy`
4. Install NLTK: run `sudo pip install -U nltk`
5. Test installation: run `python` then type `import nltk`

Windows

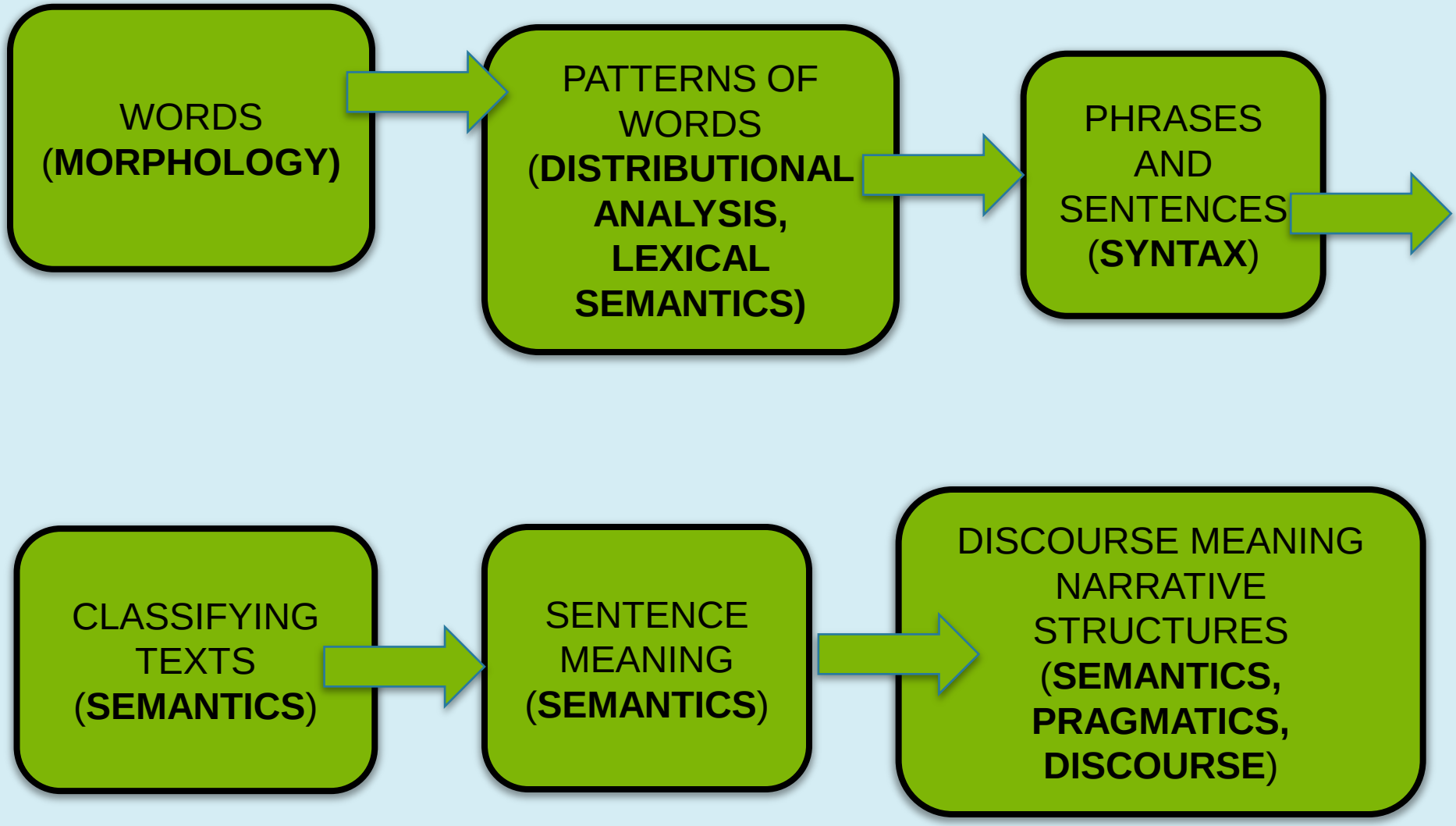
These instructions assume that you do not already have Python installed on your machine.

32-bit binary installation

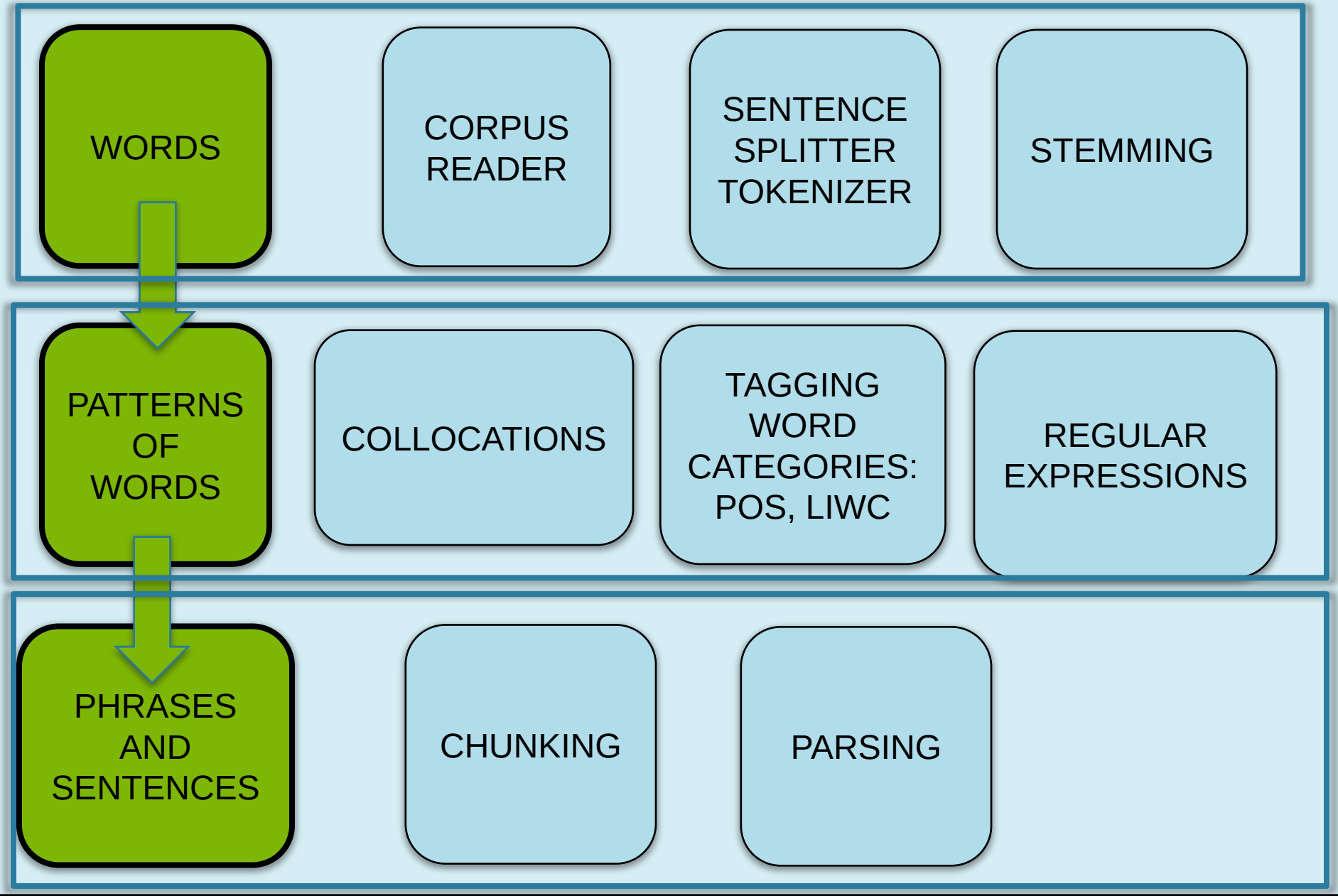
1. Install Python: http://www.python.org/download/releases/3.4.1/ (avoid the 64-bit versions)
2. Install Numpy (optional): http://sourceforge.net/projects/numpy/files/NumPy/1.8.1/numpy-1.8.1-win32-superpack-python3.4.exe
3. Install NLTK: http://pypi.python.org/pypi/nltk
4. Test installation: `Start>Python34`, then type `import nltk`

- Available for Windows, Mac OS X, and Linux.
- Free, open source, community-driven project.

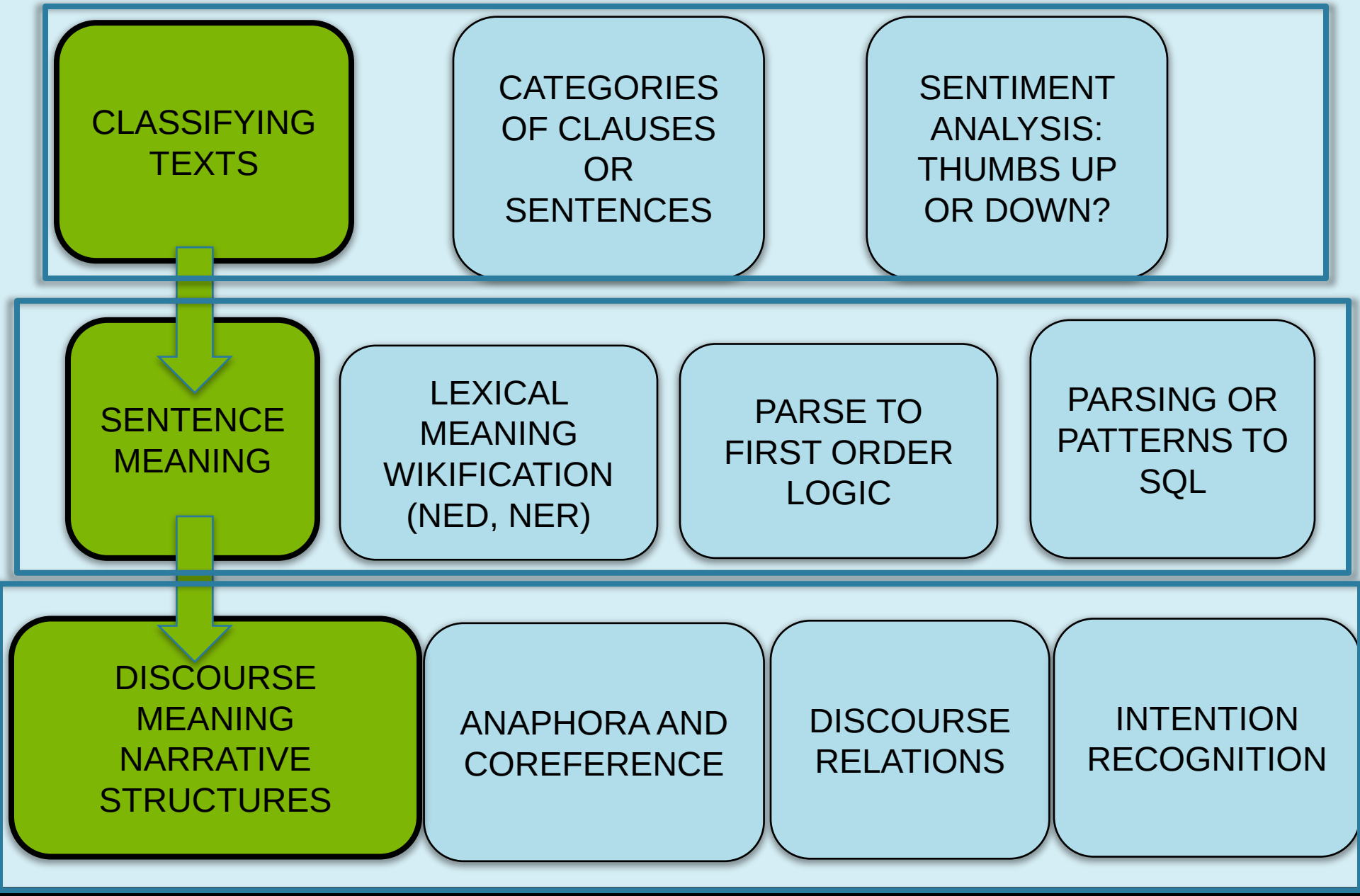
NLP PIPELINE



NLP Architecture



NLP Architecture



Modules

- Tokenizers
- Part-of-speech taggers
- Frequency Distributions
- Language Modeling
- Syntactic analysis
- Text classification
- Lexical resources
- Semantic interpretation

VERY IMPORTANT POINT:
TOOLS FOR NLP WORK MUCH BETTER ON NEWS TEXT!!

FIRST TOPIC:
Words & Word Frequencies & Word Patterns

Type / Token Distinction

- How to use the computer to count the words in a text
  - **token** = any word in the corpus
  - # tokens is an estimate of the corpus size

- The vocabulary of a text is just the *set* of tokens that it uses, since in a set, all duplicates are collapsed together.
  - **type** = unique representatives of the tokens
  - # types is an estimate of the vocabulary size

- Example:
  - *I spoke to the chap who spoke to the child*
  - 10 tokens
  - 7 types
    - *I spoke to the chap who spoke to the child*

Frequency Lists

- Frequency list for the *Gutenberg* corpus
  - Large collection of text from a variety of sources

type	frequency
,	186091
the	133583
and	95442
.	73746
of	71267
...	
zuriel	1
zuyder	1
zuzims	1

Frequency Ranks

- Word counts can get very big.
- Raw frequency lists can be hard to process.

- Useful to represent words in terms of rank:
  - count the words
  - sort by frequency (most frequent first)
  - assign a rank to the words:
    - rank 1 = most frequent
    - rank 2 = next most frequent
    - …

Rank/Frequency Profile

- rank 1 goes to the most frequent type
  - all ranks are unique

rank (r)	freq (f)
1	186091
2	133583
3	95442
…	

Note the large differences in frequency from one rank to another

Distribution of Words

- Out of 2,621,613 tokens in the Gutenberg corpus:
  - 608,186 tokens belong to just the 5 most frequent types (the types at ranks 1 -- 5)
  - 23% of our corpus size is made up of only 5 different words!
- Out of 42,339 types:
  - 15,432 are occurring only once (bottom ranks)
  - 5,761 occur only twice
  - …
- Familiar stats like the mean won't tell us very much.
  - it hides huge variations!

Rank and Frequencies of Gutenberg

- Rank 1: 186091
- Rank 2: 133583
- Rank 3: 95442
- Rank 4: 73746

Among top ranks, frequency drops very dramatically (but depends on corpus size)

…

- Rank n-2: 1
- Rank n-1: 1
- Rank n: 1

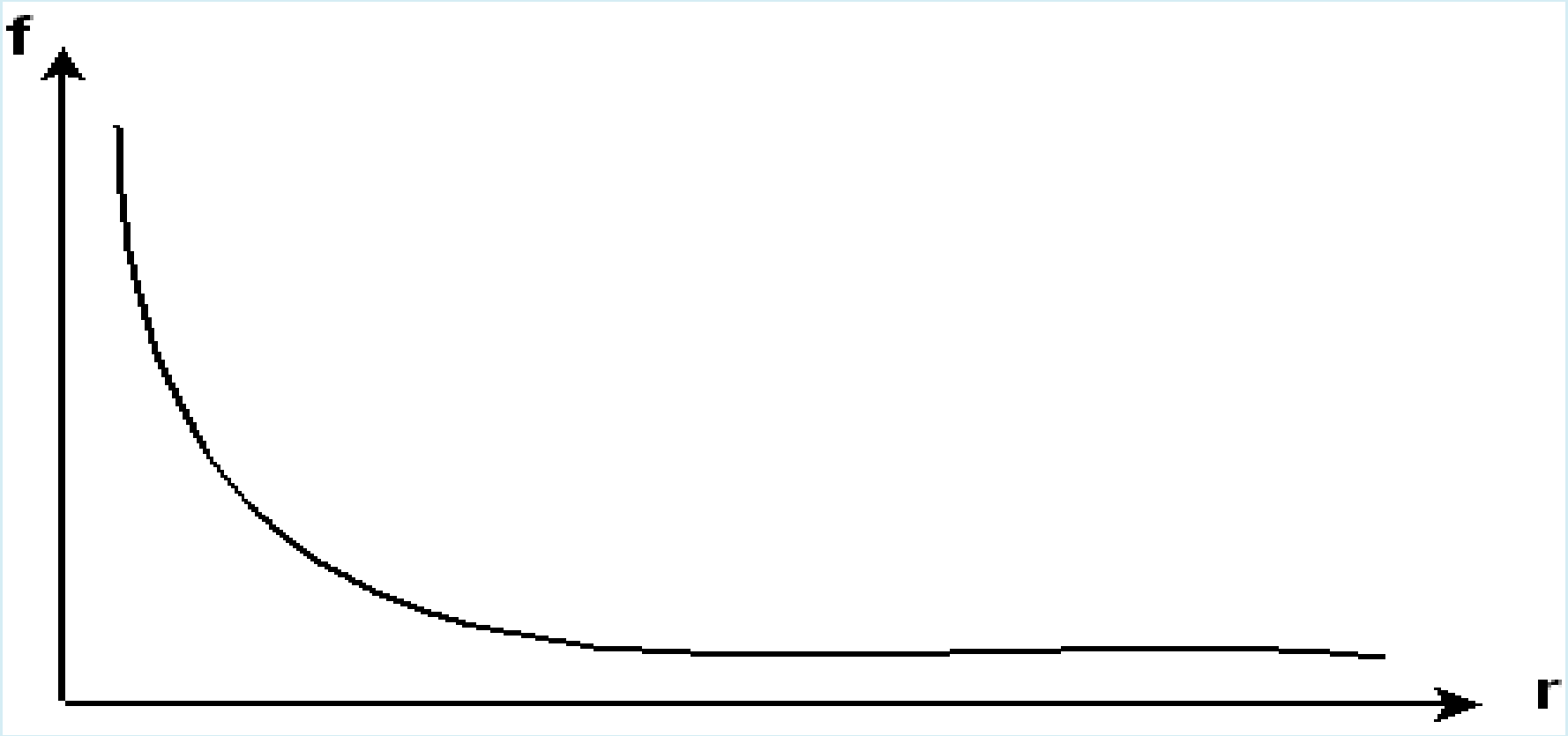
Among bottom ranks, frequency drops very gradually

General Observations

- There are always a few very high-frequency words, and many low-frequency words.
- Among the top ranks, frequency differences are big.
- Among bottom ranks, frequency differences are very small.

Typical Shape of the Rank/Freq Curve

- Frequency decreases very rapidly (exponentially) as rank increases.



THE LONG TAIL

Getting some data: NLPP Ch. 1

```
>>> from nltk.book import *
*** Introductory Examples for the NLTK Book ***
Loading text1, ..., text9 and sent1, ..., sent9
Type the name of the text or sentence to view it.
Type: 'texts()' or 'sents()' to list the materials.
text1: Moby Dick by Herman Melville 1851
text2: Sense and Sensibility by Jane Austen 1811
text3: The Book of Genesis
text4: Inaugural Address Corpus
text5: Chat Corpus
text6: Monty Python and the Holy Grail
text7: Wall Street Journal
text8: Personals Corpus
text9: The Man Who Was Thursday by G . K . Chesterton 1908
>>> text1
<Text: Moby Dick by Herman Melville 1851>
>>> text1.concordance("monstrous")
Displaying 11 of 11 matches:
ong the former , one was of a most monstrous size . ... This came towards us ,
ON OF THE PSALMS . " Touching that monstrous bulk of the whale or ork we have r
ll over with a heathenish array of monstrous clubs and spears . Some were thick
d as you gazed , and wondered what monstrous cannibal and savage could ever hav
that has survived the flood ; most monstrous and most mountainous ! That Himmal
they might scout at Moby Dick as a monstrous fable , or still worse and more de
```

Tools for counting in NLTK

```
>>> len(text3)
44764
>>>
```

- Genesis has 44,764 words and punctuation symbols, or "tokens."

```
>>> len(set(text3))
2789
>>>
```

- Discover the size of the vocabulary indirectly, by asking for the number of items in the set, and use **len** to obtain this number

```
>>> len(set(text3)) / len(text3)
0.06230453042623537
>>>
```

- Calculate a measure of the lexical richness of the text: number of distinct words is just 6% of the total number of words

```
>>> text3.count("smote")
5
>>> 100 * text4.count('a') / len(text4)
1.4643016433938312
>>>
```

- Count how often a word occurs in a text, and compute what percentage of the text is taken up by a specific word.

Tools for counting in NLTK

- What makes a text distinct?
- Automatically identify the words of a text that are most informative about the topic and genre of the text
- **Frequency Distribution:** the frequency of each vocabulary item in the text.

```
>>> fdist1 = FreqDist(text1) ❶
>>> print(fdist1) ❷
<FreqDist with 19317 samples and 260819 outcomes>
>>> fdist1.most_common(50) ❸
[(',', 18713), ('the', 13721), ('.', 6862), ('of', 6536), ('and', 6024),
('a', 4569), ('to', 4542), (';', 4072), ('in', 3916), ('that', 2982),
("'", 2684), ('-', 2552), ('his', 2459), ('it', 2209), ('I', 2124),
('s', 1739), ('is', 1695), ('he', 1661), ('with', 1659), ('was', 1632),
('as', 1620), ('"', 1478), ('all', 1462), ('for', 1414), ('this', 1280),
('!', 1269), ('at', 1231), ('by', 1137), ('but', 1113), ('not', 1103),
('--', 1070), ('him', 1058), ('from', 1052), ('be', 1030), ('on', 1005),
('so', 918), ('whale', 906), ('one', 889), ('you', 841), ('had', 767),
('have', 760), ('there', 715), ('But', 705), ('or', 697), ('were', 680),
('now', 646), ('which', 640), ('?', 637), ('me', 627), ('like', 624)]
>>> fdist1['whale']
906
>>>
```

Ngrams & Counting Other Things

- Can basically count anything with `FreqDist`
- **N-grams:** sequences of *n* consecutive words e.g., "more is said than done"
  - Unigrams: "more", "is", "said", "than", "done"
  - Bigrams: "more is", "is said", "said than", "than done"
  - Trigrams: "more is said", "is said than", "said than done"
  - …
- Used a lot in NLP applications
  - Language models (next week)
  - Collocation (next)
  - Language Identification
  - Machine Translation

Conditional Counts

- We can also get some more interesting information by using a `ConditionalFreqDist`
- How often have I seen $word_2$ given that $word_1$ immediately preceded it?
  - *fox* is seen exactly twice after having seen *the*

```
>>> import nltk
>>> fables_text = open('cmps143/fables/TheFoxAndTheCrow.txt').read()
>>> sentences = nltk.sent_tokenize(fables_text)
>>> words = [nltk.word_tokenize(sentence) for sentence in sentences]
>>> flat_words = [word.lower() for sentence in words for word in sentence]
>>> bigrams = nltk.bigrams(flat_words)
>>> bgcdist = nltk.ConditionalFreqDist(bigrams)
>>> bgcdist.tabulate(conditions=["the", "fox", "and", "the", "crow"])
        ,  birds  cheese  crow   fox   hue  just  observed  said   set  standing  that   the  tree   was
 the    0      1       2     1     2     1     0         0     0     0         0     0     1     0
 fox    1      0       0     0     0     0     0         1     0     0         0     1     0     0
 and    0      0       0     0     0     0     1         0     1     1         1     0     1     0
 the    0      1       2     1     2     1     0         0     0     0         0     0     1     0
crow    0      0       0     0     0     0     0         0     0     0         0     0     0     2
```

To do

- Read Chapter 1, Sections 1-4:
  http://www.nltk.org/book/ch01.html
- Install NLTK 3 (and all the requirements):
  http://www.nltk.org/install.html
- Go to lab sections and get help
  - Tuesday-Thursday 4-6 pm in Soc Sci I Mac Lab – Room 135
- Complete HW0 (quiz)
  - Return it on Thursday, March 31, beginning of class

Next…

More of NLTK…

- Using your own data in NLTK
- Splitting sentences
- Tokenization
- N-grams and conditional counts
- Collocations
- Stemming
- POS tagging
- Lexical Resources: WordNet