

# Natural Language Processing

Arjun Mukherjee<sup>†</sup>

Course webpage:

[http://www.cs.uh.edu/~arjun/courses/nlp\\_ugrad](http://www.cs.uh.edu/~arjun/courses/nlp_ugrad)

---

<sup>†</sup> Contains contents from [Manning and Schütze, 1999] , and various other sources. Referenced in place.

# Goals of the field

- Computers would be a lot more useful if they could
  - Manage emails/social contacts
  - Extract information from archival/library texts
  - Take complex commands
  - Teach
  - ...
- **Q: How to achieve this goal?**
- **A: Need technology that can understand, derive meaning, process, and generate human (natural) language – Natural Language Processing**

# NLP Applications

- Computers would be a lot more useful if they could
  - Speech recognition (e.g., Apple Siri)
  - Language translation (Google translate)
  - Sentiment analysis, recommendations (e.g., Amazon.com)
  - Snippet generation, summarization (e.g., Google news)
  - Write, comprehend and converse intelligent dialogue (e.g., BERT, ChatGPT)...

# Goals of this course

- Understand NLP problems, tasks
- Learn algorithms and techniques to solve NLP tasks
- Connection between NLP/Text mining and Statistics
- **At the end you should be able to:**
  - Understand research papers in the field
  - Use skills learnt  $\Rightarrow$  solve real world problems
  - Develop ownership of formal/statistical models
  - Get a job at a company doing NLP/text mining (Google, Microsoft, Facebook, Yahoo!, IBM, etc.)

# Course Outline

- Introduction to NLP
- Text Retrieval
- Mathematical/Statistical Foundations
- Words: Collocations & N-gram models
- Markov Models: POS Tagging, Chunking
- Grammar and Parsing
- Text Categorization
- Sentiment Analysis and Opinion Spam
- Clustering, Topic Models, Neural Models

# Grading [Tentative]

- Homeworks (theory): 30%
- Project (Programming): 40%
- Final: 25%
- Class participation: 5%
- Bonus/Extra credits to score higher (Maybe? Depending upon student demand):
- Research project: 30%
- **NOTE: Focus is on Learning.** Grading will be curved/prorated to class performance.

# Resources

- Course materials (lecture notes/slides, online resources/offline copies, etc.):
  - Download from this location:  
[http://www2.cs.uh.edu/~arjun/courses/nlp/course\\_materials.7z](http://www2.cs.uh.edu/~arjun/courses/nlp/course_materials.7z)
  - Password: cxAmb6 **Please do not re-post on the web**
  - **IMP:** These resources, along with sections in books (under Required Readings) should be used for preparing for this course.
- Books are available online (e.g., used copies @ Amazon, online pdfs). Some scans of chapters are also included (to help you get started until your book comes!)

# Resources

- Books:
  - **SI:** Statistical Inference, Casella and Berger. Cengage Learning; 2nd edition.
  - **SLP:** Speech and Language Processing, Jurafsky and Martin
  - **FSNLP:** Foundations of Statistical Natural Language Processing, Chris Manning and Heinrich Schütze. MIT Press. Cambridge, MA: May 1999.
  - **WDM:** Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Bing Liu; Springer, 1st Edition.
- All these are available from Amazon or even online
- **All required reading mentioned in [course webpage](#) + lecture slides/notes are essential. Expect homework/exam questions from them!**
- **Suggested/recommended reading references under each topics will be useful for gaining deeper understanding**



# Resources

- NLP conferences and papers
  - ACL
  - EMNLP
  - NAACL
  - All papers are (**freely**) available in ACL Anthology (<http://aclweb.org/anthology-new/>)

# A recipe for NLP research!

## To Write A Typical Paper

Needs to be a real-world problem

Needs deep thinking

knowledge of existing work

- Need some of these ingredients:

- A domain of inquiry    *Scientific or engineering question*
- A task    *Input & output representations, evaluation metric*
- Resources    *Corpora, annotations, dictionaries, ...*
- A method for training & testing    *Derived from a model?*
- An algorithm
- Analysis of results    *Comparison to baselines & other systems, significance testing, learning curves, ablation analysis, error analysis*

# Why is NLP hard?

- Ambiguity

The screenshot shows a Google search for "jaguar". The search bar at the top contains the word "jaguar". Below the search bar, there are tabs for "Web", "Images", "News", "Videos", "Shopping", "More", and "Search tools". The "Web" tab is selected. The search results show "About 80,300,000 results (0.37 seconds)".

The first result is "jaguarusa.com - Jaguar® Official Site" with a link to "www.jaguarusa.com/". Below this link, there are four buttons: "Build & Price", "View Offers", "Jaguar F-TYPE", and "Schedule A Test Drive".

The second result is "Jaguar® Houston Site - Jaguar-Houston.com" with a link to "www.jaguar-houston.com/". Below this link, there is a text description: "Take A Jaguar Out For A Spin When You Locate A Houston Jaguar Dealer".

The third result is "Jaguar: Luxury Cars & Sports Cars | Jaguar USA" with a link to "www.jaguarusa.com/". Below this link, there is a text description: "The official home of Jaguar USA. Our luxury cars feature innovative designs along with legendary performance to deliver one of the top sports cars in the ...".

The fourth result is "Jaguar - Wikipedia, the free encyclopedia" with a link to "en.wikipedia.org/wiki/Jaguar". Below this link, there is a text description: "The jaguar Panthera onca, is a big cat, a feline in the Panthera genus, and is the only Panthera species found in the Americas. The jaguar is the third-largest ...".

The fifth result is "Jaguar Cars - Wikipedia, the free encyclopedia" with a link to "en.wikipedia.org/wiki/Jaguar\_Cars". Below this link, there is a text description: "Jaguar Cars is a brand of Jaguar Land Rover, a British multinational car manufacturer headquartered in Whitley, Coventry, England, owned by Tata Motors since ...".

On the right side of the search results, there is a map of Houston, Texas, with several red pins labeled A, B, C, D, E, and F. Below the map is a link "Map for jaguar".

At the bottom right, there is a "Jaguar" knowledge panel. It shows the Jaguar logo, the number of followers on Google+, and a brief description of the company: "Jaguar Cars is a brand of Jaguar Land Rover, a British multinational car manufacturer headquartered in Whitley, Coventry, England, owned by Tata Motors since 2008. Wikipedia". It also lists the founder, CEO, and founders.

# Why is NLP hard?

- Ambiguity
- Structured semantics



# Why is NLP hard?

- Ambiguity
- Structured semantics
- Implied semantics

*My car uses a lot of gas!*



**Negative connotation** although no negative words like poor, bad, waste, etc. are used.



# Why is NLP hard?

- Ambiguity
- Structured semantics
- Implied semantics
- **Fundamentally: a trade-off between linguistics [rich/complex representation]  $\leftrightarrow$  NLP [mathematical modeling]**

Intractable

Tractable

# Why learn NLP?

- Because you can make the world better:  
Build apps like Google Knowledge graph, IBM Watson, Google translate, etc.
- Its fun to deal with real-world NLP problems:
  - (a) Detect lies, fake reviews from language
  - (b) Predict user personalities, gender, emotions from language
- Your future employer will love it



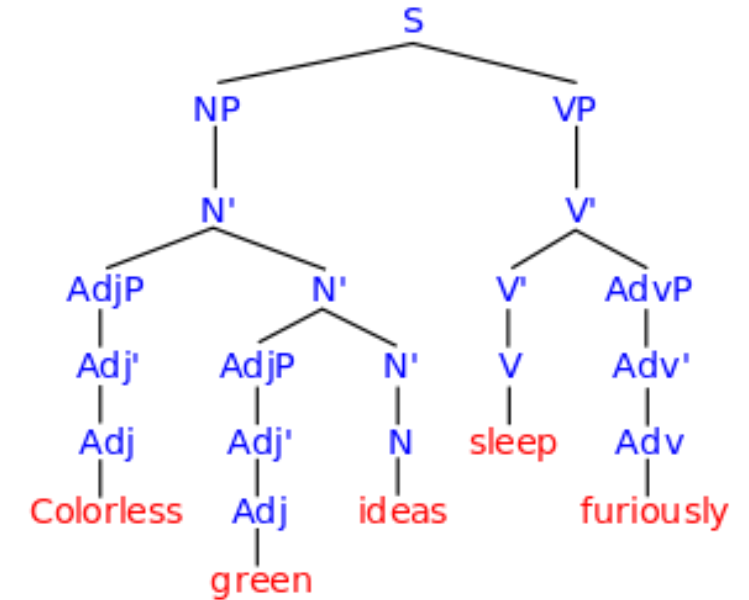
# Basics

- Syntax: grammatical ordering of words
- Semantics: meaning of words, phrases

- Syntax vs. Semantics [Chomsky, 1957]

(1) Colorless green ideas sleep furiously

(2) Furiously sleep ideas green colorless



Neither (1) or (2) has ever occurred in English discourse.

Statistically, both are equally remote and nonsensical in English. Yet, (1) is grammatical.



# Basics

- Syntax: grammatical ordering of words
- Semantics: meaning of words, phrases
- Tokens and tokenization: Given a character sequence and a defined document unit (e.g. word), tokenization is the task of chopping it up into pieces, called *tokens*.

Example from [Manning et al., 2008]

Input: Friends, Romans, Countrymen, lend me your ears;

Output: 

Friends	Romans	Countrymen	lend	me	your	ears
---------	--------	------------	------	----	------	------

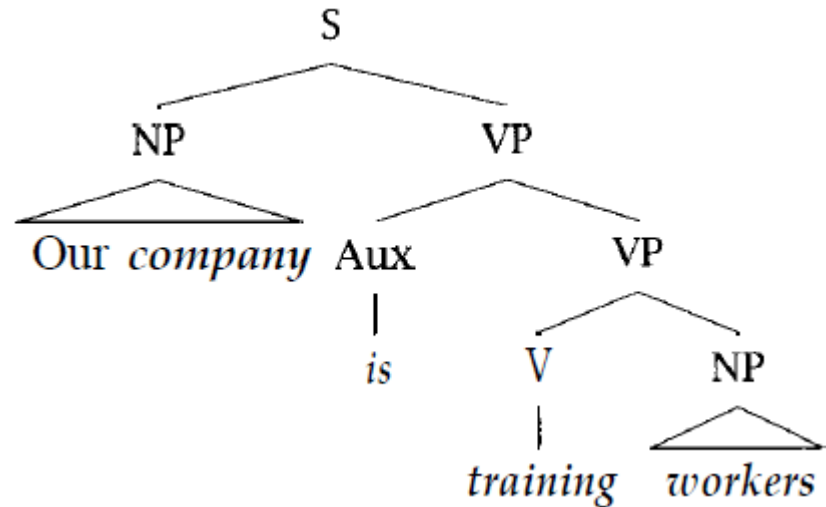
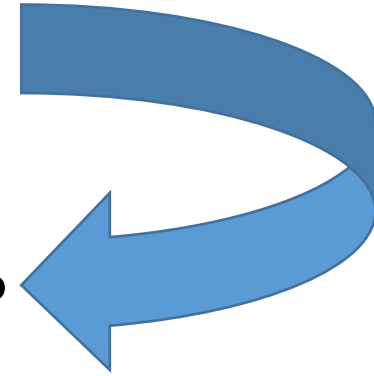
- Corpus/corpora: Document collection (e.g., news archives)

# Language: A probabilistic phenomenon

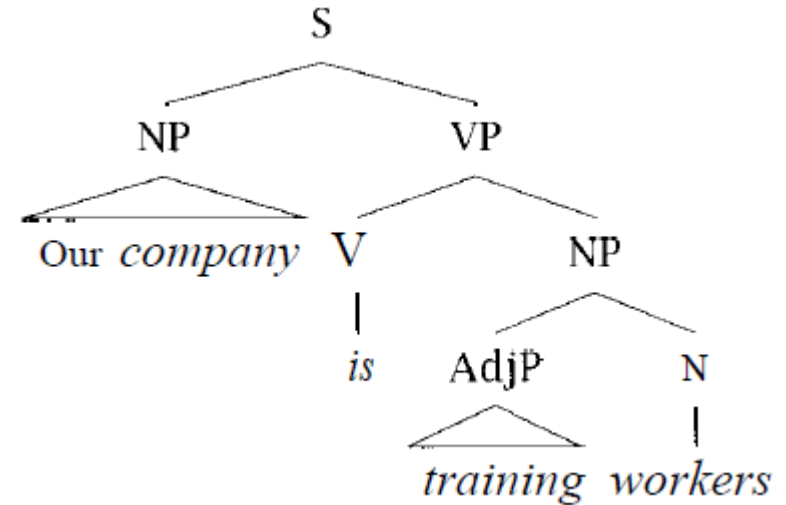
- Human cognition/reasoning is probabilistic

**Why?** Because our world has uncertainties

- Modeling language using probability : Statistical NLP
- Deals with ambiguity. Parse trees for: “Our company is training workers”



**VS**

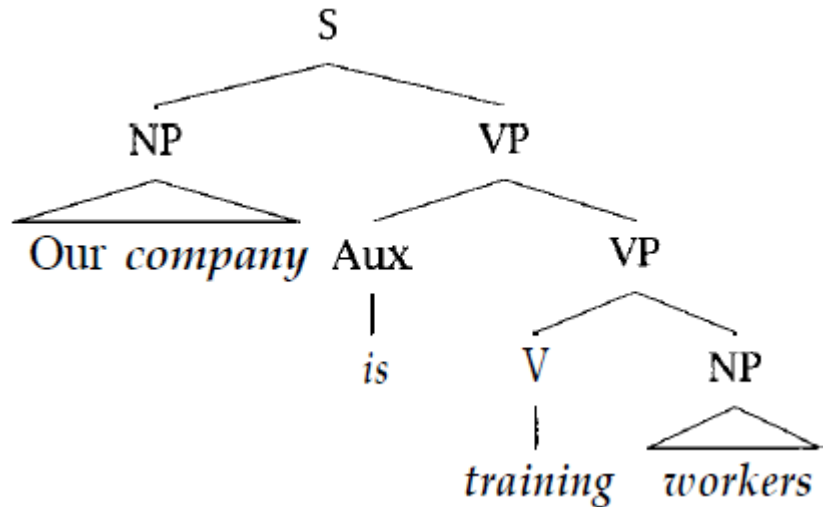
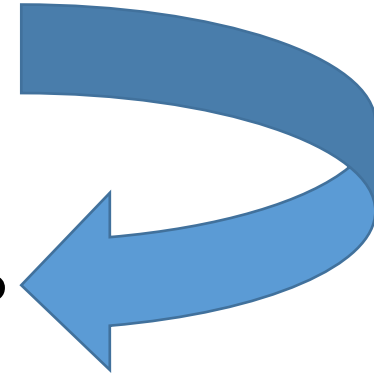


# Language: A probabilistic phenomenon

- Human cognition/reasoning is probabilistic

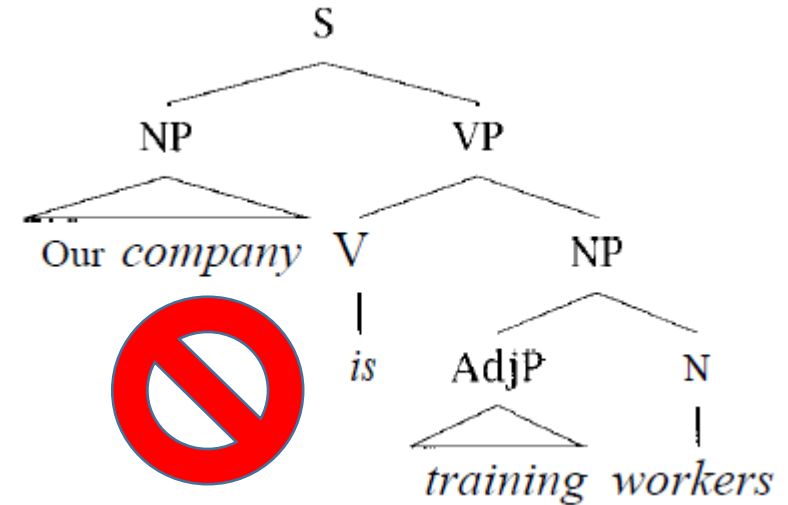
**Why?** Because our world has uncertainties

- Modeling language using probability : Statistical NLP
- Deals with ambiguity. Parse trees for: “Our company is training workers”



“is training” ⇒ main verb

**VS**



training *modifies* workers.

# Frequency of Word Occurrence

- Q: How do words appear in texts (statistically)?

Word	Freq.	Use
the	3332	determiner (article)
and	2972	conjunction
a	1775	determiner
to	1725	preposition, verbal infinitive marker
of	1440	preposition
was	1161	auxiliary verb
it	1027	(personal/expletive) pronoun
in	906	preposition
that	877	complementizer, demonstrative
he	877	(personal) pronoun
I	783	(personal) pronoun
his	772	(possessive) pronoun
you	686	(personal) pronoun
Tom	679	proper noun
with	642	preposition

Table 1.1 Common words in *Tom Sawyer*.

# Frequency of Word Occurrence

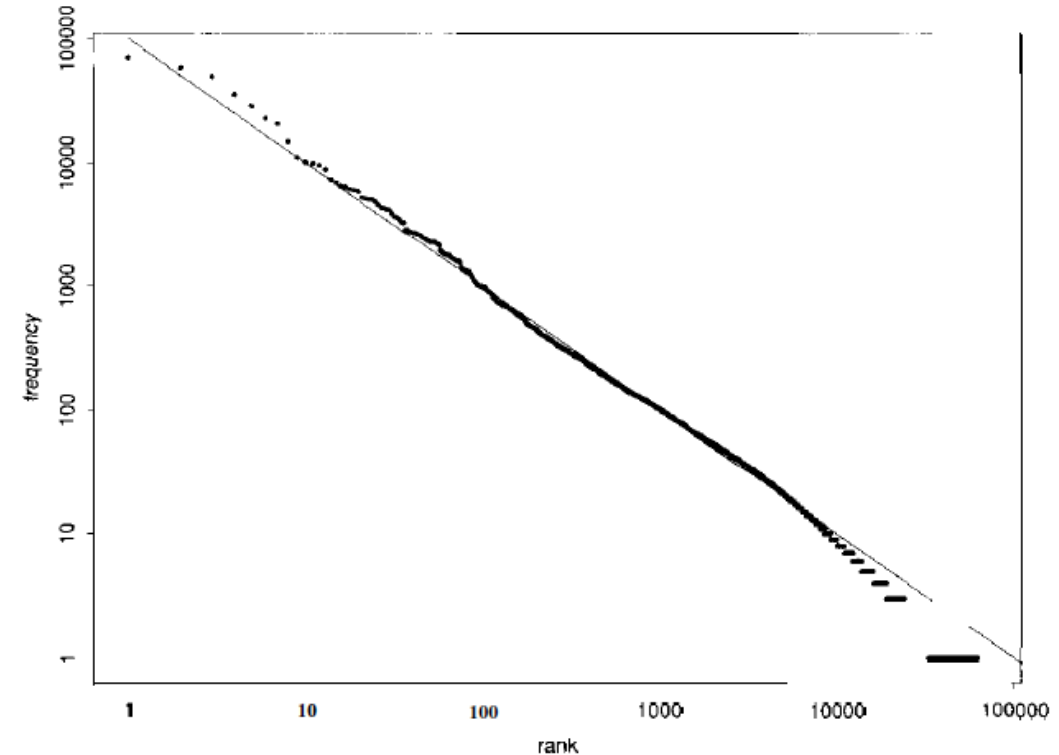
- Q: What is the relation of frequency and rank (statistically)?

Word	Freq. ( $f$ )	Rank ( $r$ )	$f \cdot r$	Word	Freq. ( $f$ )	Rank ( $r$ )	$f \cdot r$
the	3332	1	3332	turned	51	200	10200
and	2972	2	5944	you'll	30	300	9000
a	1775	3	5235	name	21	400	8400
he	877	10	8770	comes	16	500	8000
but	410	20	8400	group	13	600	7800
be	294	30	8820	lead	11	700	7700
there	222	40	8880	friends	10	800	8000
one	172	50	8600	begin	9	900	8100
about	158	60	9480	family	8	1000	8000
more	138	70	9660	brushed	4	2000	8000
never	124	80	9920	sins	2	3000	6000
Oh	116	90	10440	Could	2	4000	8000
two	104	100	10400	Applausive	1	8000	8000

Table 1.3 Empirical evaluation of Zipf's law on *Tom Sawyer*.

# Zipf's law

- We see that  $f \propto \frac{1}{r}$
- Plot of  $\log(f)$  vs.  $\log(r)$
- Mandelbrot generalizes this with more parameters to
$$f = P(r + \rho)^{-B}$$



- Q: Does Zipf's law provide any “deep knowledge” about how language is emitted by human?

## Zipf's law (contd.)

- Q: If we promise bananas to a monkey for typing randomly (& tirelessly!) on a QWERTY keyboard, what do we get?

(a) A monkey software engineer!

(b)  $f \propto \frac{1}{r}$

(c) Some other relation between  $f$  and  $r$

(d) No relation between  $f$  and  $r$  whatsoever

# Miller's Monkey

- For simplicity, assume keyboard has 27 keys (a-z, space) and each key hit is equally likely. Sequence separated by 'space' forms a monkey word.

The probability that a specific monkey word type has length  $i$  is

$$P(i) = (1/27)^i (1/27) = (1/27)^{i+1}. \quad (1)$$

As we can see, the longer the word, the lower its probability – therefore the lower the expected count in the monkey corpus. Let us rank all monkey words by its probability. The number of monkey word-types with length  $i$  is  $26^i$ . The rank  $r_i$  of a word with length  $i$  thus satisfies

$$\sum_{j=1}^{i-1} 26^j < r_i \leq \sum_{j=1}^i 26^j \quad (2)$$

Let us consider the word with rank  $r = \sum_{j=1}^i 26^j$ . The word actually has length  $i$ , but from

$$r = \sum_{j=1}^i 26^j = \frac{26}{25}(26^i - 1), \quad (3)$$



# Miller's Monkey, contd.

we can derive a 'fractional length'  $i'$

$$i' = \frac{\log \left( \frac{25}{26}r + 1 \right)}{\log 26}. \quad (4)$$

The frequency of this word is

$$p(i') = (1/27)^{i'+1} \quad (5)$$

$$= (1/27)^{\frac{\log \left( \frac{25}{26}r + 1 \right)}{\log 26} + 1} \quad (6)$$

$$= (1/27)^{\left( \frac{25}{26}r + 1 \right)^{-\frac{\log 27}{\log 26}}} \quad \text{using the fact } a^{\log b} = b^{\log a} \quad (7)$$

$$\approx 0.04(r + 1.04)^{-1.01}, \quad (8)$$

which fits Mandelbrot's law, and is fairly close to Zipf's law.

In light of the above analysis, Zipf's law may not reflect some deep knowledge of languages. Nonetheless, it still points to an important empirical observation, that almost all words are rare. This is also known as the heavy tail property.

# Common Text Preprocessing

- Removal of stopwords: Words like “the,” “and,” “a,” “to,” “of,” etc. have less information but appear much more frequently than other content/function words.
- SMART is such a stopword list:  
<https://gist.github.com/sebleier/554280>
- Stemming/Lemmatization: Optional (may help increase statistical co-occurrence). Words like looked, looking, looks share the same stem: “look.”
- Porter stemmer: A popular algorithm for stemming.
- Warning: Aggressive stemming can be risky!

# Student Background Questionnaire

- Registered for this course, auditing, not decided yet (?)  
**IMP: Sep 2, is the last day to add this class**
- Taken courses in any of these areas: statistics/probability/AI/machine learning/data mining? If yes, which area(s)?
- What other areas interest you (systems, theory, etc.)?
- Why are you taking this course?
- This course needs Math and Programming, are you willing to work towards it?

# Text Retrieval: Search

- Suppose we are given a large collection of documents (corpus): e.g., Amazon reviews on electronics, Shakespeare's plays, Yahoo news archives, etc.
- **Q: How do we search of reviews/plays/articles containing certain words?**
- **Q: How do we perform the simple NLP task of text retrieval?**
- **Q: How to build a search engine for retrieving documents with specific queries e.g., “iphone wifi issues”, “romeo and juliet”, “syrian uprising obama” ?**

# Boolean Text Retrieval

- We refer to [slides by H. Schütze](#).

Introduction to Information Retrieval  
<http://informationretrieval.org>

IIR 1: Boolean Retrieval

Hinrich Schütze

Center for Information and Language Processing, University of Munich

2014-04-09

# Homework 1

- Due date: 9/3/2014.
- Required reading (from course webpage):

Topic(s)	Resources: Readings, Slides, Lecture notes, Papers, Pointers to useful materials, etc.
<b>Introduction</b> Course administrivia, plan, goals, NLP Resources Language as a probabilistic phenomenon, Zipf's law Word collocations and text retrieval	<b>Required readings:</b> Lecture notes/slides Chapter 1 FSNLP (Sections 1.2.3, 1.4, 1.4.1, 1.4.2, 1.4.3, 1.4.4) <a href="#">Boolean retrieval slides by H.Schutze</a> <a href="#">Boolean retrieval [Manning et al., 2008] (upto section 1.4)</a>

- Understand the concepts of text retrieval well as this will help in your mini project.