



Introduction to NLP

DSLA COURSE

ROHIT PADEBETTU

Types of Language

A:"Hey, is there a german word for making something worse while trying to fix it?"

B:"Verschlimmbessern"

A:"You really have a word for that?!"

1h 88 Ober-Ramstadt

Morning
おはようございます
ohayou gozaimasu
Good Morning

Afternoon
こんには
konnichi wa
Good Afternoon

Evening
こんばんは
konban wa
Good Evening
おやすみなさい
oyasumi nasai
Good Night

Conversation Starters Saying Goodbye

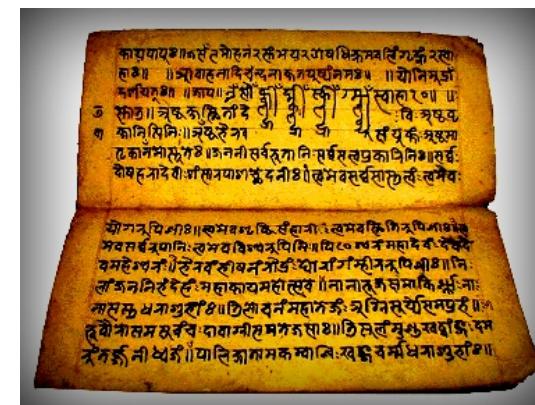
元気ですか?
genki desu ka
How are you?

じゃあ、またね。
jaa, mata ne
See you later.

さようなら
sayounara
Good-bye
気をつけて
ki o tsukete
Take care.
失礼します
shitsurei shimasu
Excuse me.

バイバイ bai bai Bye bye

And whan I swagh he wolde never fine
To reden on this cursed booke al night,
Al sodeinly three leves have I plight
Out of his booke right as he redde, and eke
I with my fist so took him on the cheeke
That in oure fir he fel bakward adown.
And up he sterte as dooth a wood leon
And with his fist he smoot me on the heed
That in the floor I lay as I were dead.
And whan he swagh how stille that I say,
he was agast, and wolde have fled his way,
Till atte laste out of my swough I braide:
"O hastou slain me, false thief?" I saide,
"And for my land thus hastou mordred me?
Er I be deed yit wol I kisse thee."



www.mandmx.com

copyright 2008



Artificial Language

```

def add5(x):
    return x+5

def dotwrite(ast):
    nodename = getNodename()
    label=symbol.sym_name.get(int(ast[0]),ast[0])
    print '  ' + ts + ' ' + (nodename, label)
    if isinstance(ast[1], str):
        if ast[1].strip():
            print ts + ts + ast[1]
        else:
            print ts + ts + ''
    else:
        print ts + ts + ']'
        children = []
        for n, childenumerate(ast[1:]):
            children.append(dotwrite(child))
        print ts + ts + ' -> ' + nodename
        for n in children:
            print ts + ts + ' ' + name,

```

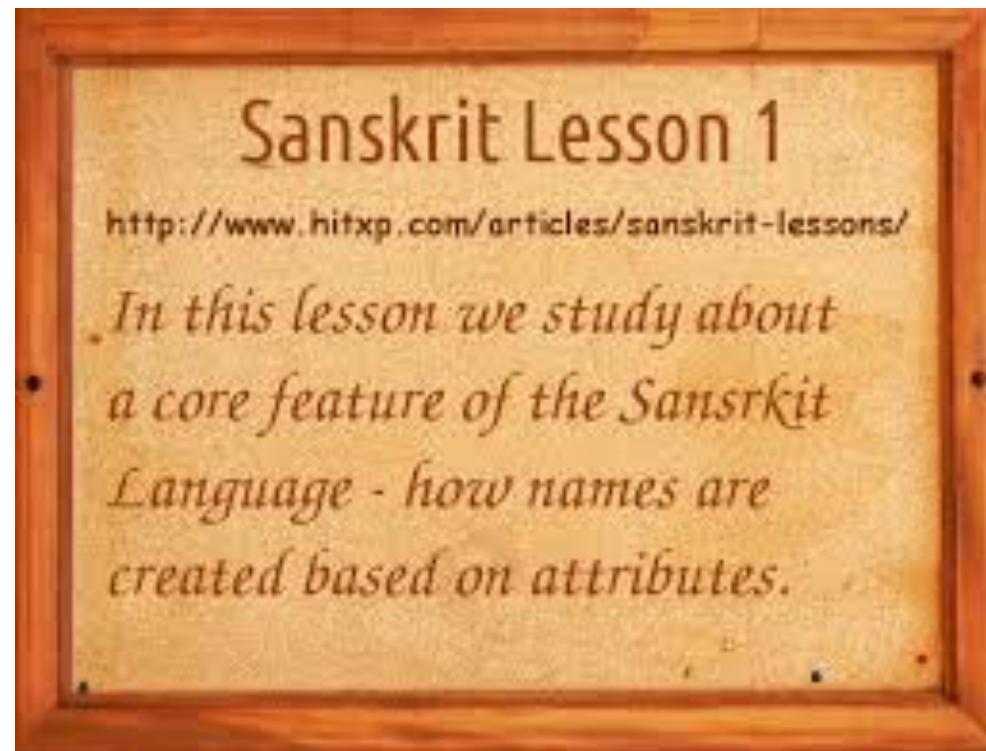
```

          TS-BASIC      program   0
ok
LIST
100  PROGRAM "Prim.bas"
110  FOR X=1 TO 500
120  IF PRIM(X) THEN PRINT X;
130  NEXT
140  DEF PRIM(N)
150  IF N<-1 OR INT(N)<>N THEN
160  LET PRIM=0
170  ELSE IF N=2 THEN
180  LET PRIM=-1
190  ELSE
200  LET PRIM=-1
210  FOR I=2 TO CEIL(SQR(N))
220  IF MOD(N,I)=0 THEN
230  LET PRIM=0
240  EXIT FOR
250  END IF
260  NEXT
270  END IF
280 END DEF
ok
START
2

```

Natural Language Processing

Natural language processing (NLP) is the ability of a computer program to understand human speech as it is spoken. **NLP** is a component of artificial intelligence (AI).



What is a Language?

A **Language** is composed of a set of all possible **Texts**



What is a Text?

A **Text** is composed of a sequence of words from **Vocabulary**



ॐ असतो मा सद् गमय।
तमसो मा ज्योतिर्गमय।
मृत्योर्मातुःमृतम् गमय ॥

Om, Asato Maa Sad Gamaya;
Tamaso Maa Gytira Gamaya;
Mrityora Maa Amritam Gamaya

*Lead me from the unreal to the real;
from darkness (ignorance) to light (knowledge);
and from death to immortality.*

*"By three methods we may
learn wisdom: First, by
reflection, which is noblest;
Second, by imitation, which is
easiest; and third by
experience, which is the
bitterest."*

What is a Vocabulary?

A **Vocabulary** is composed of a set of words

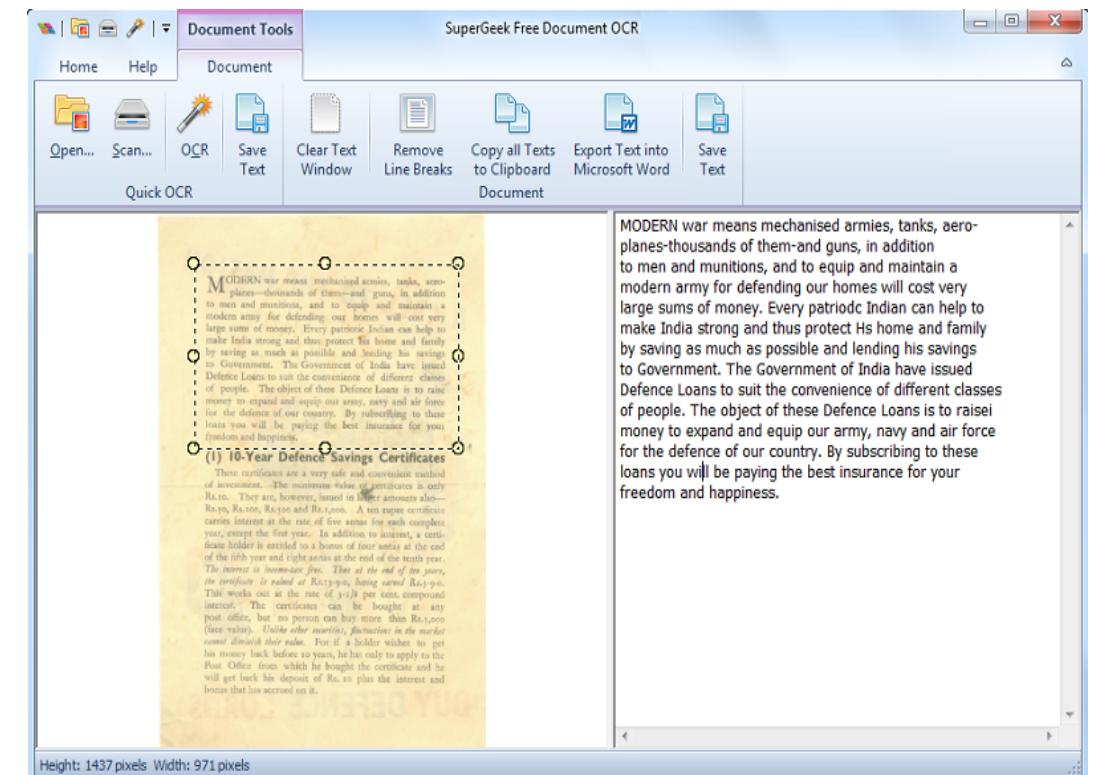
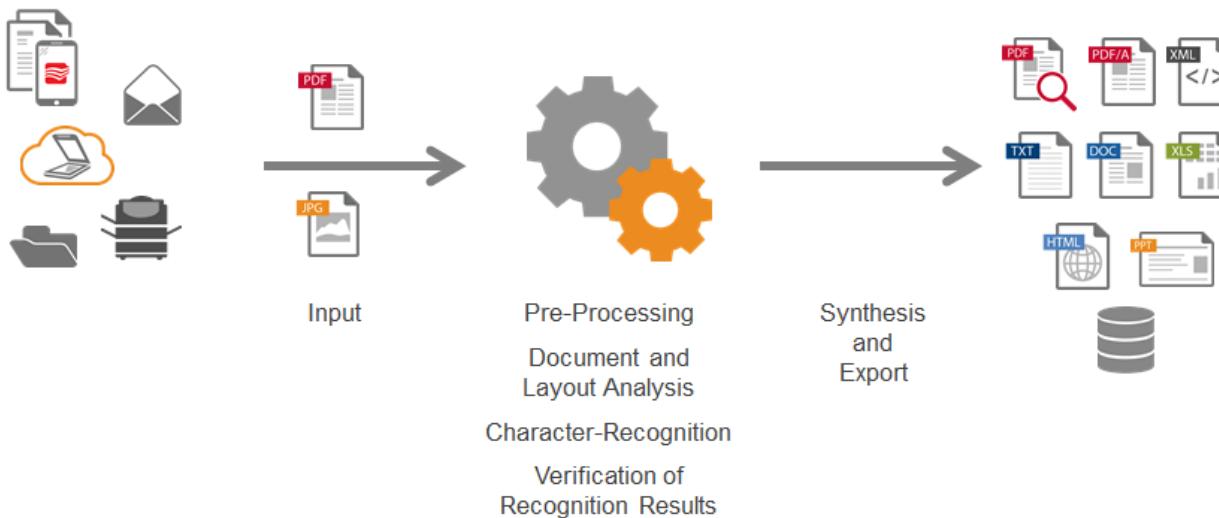


Applications of NLP

Optical Character Recognition

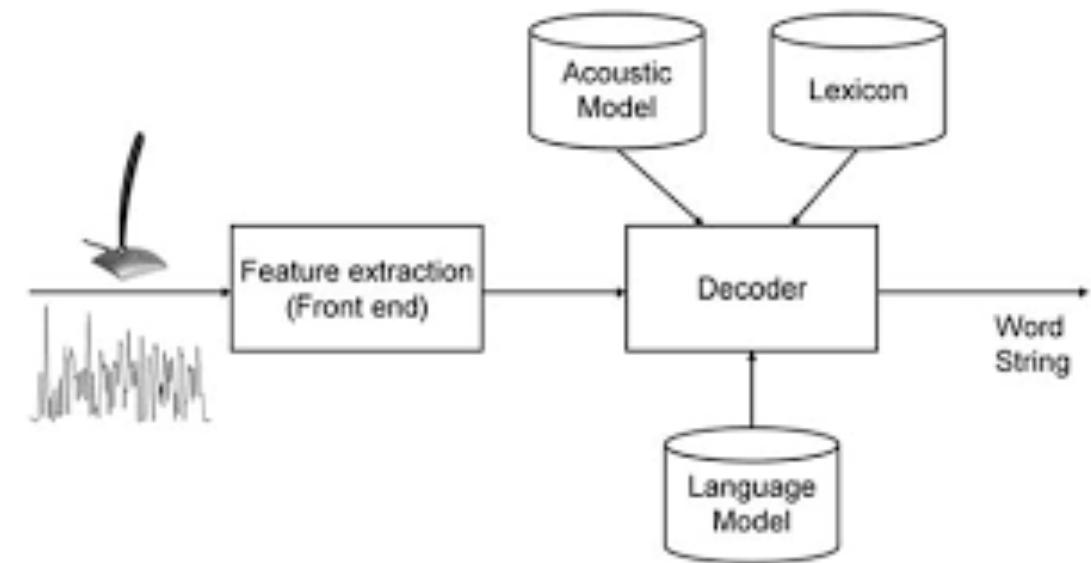
What is Optical Character Recognition (OCR)?

... it is definitely not only character recognition.



Applications of NLP

Speech Recognition



Applications of NLP

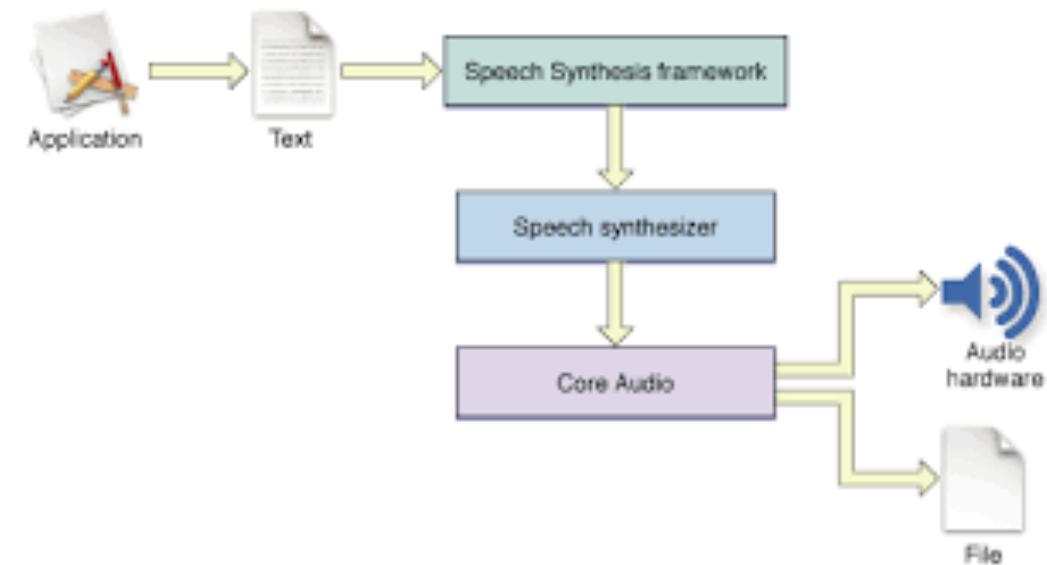
Speech Synthesis & Spoken dialog systems



“ What's the best computer ”
The Apple Macintosh is my favorite computer.

“ What's the best tablet ”
The Apple iPad. Need you ask?

“ What's the best phone ”
Wait... there are other phones?



Applications of NLP

Spelling and Grammar Check

As we talked about in our meeting, my fourteen, floor sales and in the role of Sales Supervisor, wo time, I have learned many techniques that would ratings at Quality Furnishings.

Spelling error

In addition, I wanted to let you pg that I have recently recieived my certificate from the Superior Sales Training program at the National Business Institute. several techniques covered in the program are sure to bolster sales. Also, increased customer satisfaction, I look forward to having the chance to implement them at Quality Furnishings.

Contextual spelling error

Grammatical error

The in filling the Sales As co. I free to forward to hearing from you soon.

wikipadia

Web

Images

Maps

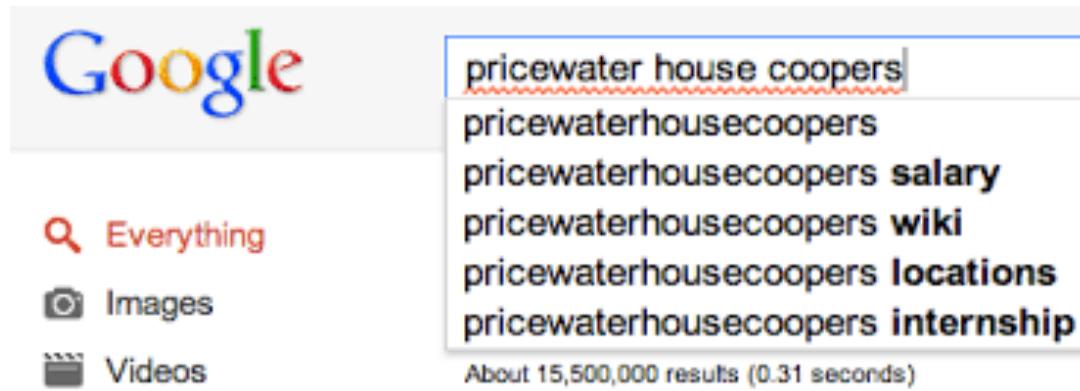
Shopping

About 1,020,000,000 results (1.01 seconds)

Showing results for [wikipedia](#) ←
Search instead for [wikipadia](#)

Applications of NLP

Word Prediction



Google

Everything Images Videos

pricewater house coopers

pricewaterhousecoopers
pricewaterhousecoopers salary
pricewaterhousecoopers wiki
pricewaterhousecoopers locations
pricewaterhousecoopers internship

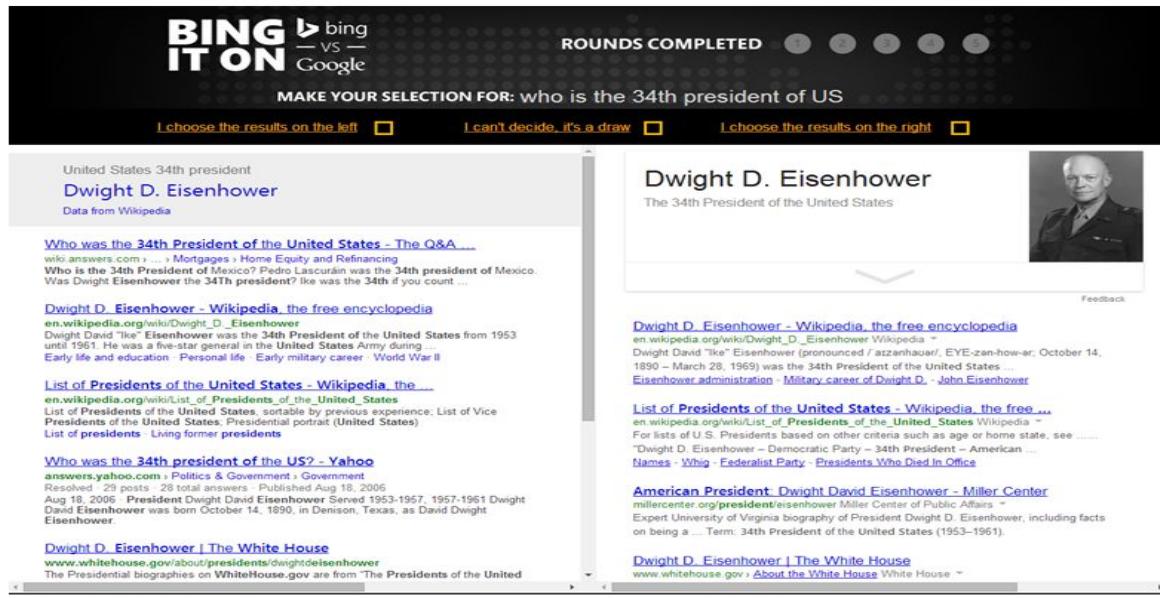
About 15,500,000 results (0.31 seconds)



Applications of NLP

Information Retrieval

Bing v.s. Google?



Applications of NLP

Document Categorization

Publication Types, MeSH Terms

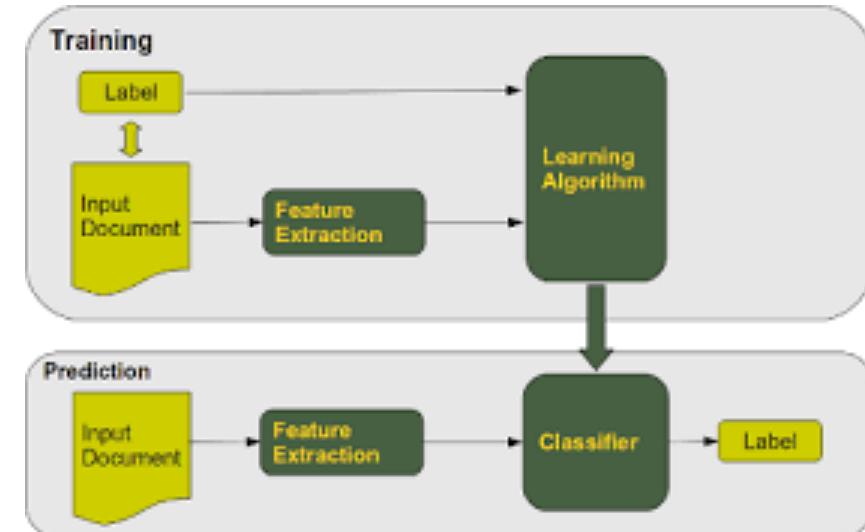
Publication Types
[Evaluation Studies](#)
[Review](#)

MeSH Terms
[Bariatric Surgery/adverse effects*](#)
[Bariatric Surgery/methods](#)
[Bariatric Surgery/rehabilitation](#)
[Bariatric Surgery/statistics & numerical data](#)
[Deficiency Diseases/epidemiology](#)
[Deficiency Disease/etiology*](#)
[Humans](#)
[Models, Biological](#)
[Nutrition Disorders/epidemiology](#)
[Nutrition Disorders/etiology*](#)
[Obesity/epidemiology](#)
[Obesity/surgery](#)
[Postoperative Complications/epidemiology](#)

[LinkOut - more resources](#)

Click on this tab in a Pubmed article and a list of MeSH terms the article has been indexed with will appear. Next to some of the MeSH terms you can see subheadings.

- █ Top Stories
- █ World
- █ U.S.
- █ Business
- █ Elections
- █ Sci/Tech
- █ Sports
- █ Entertainment
- █ Health
- █ **Spotlight**
- █ Most Popular



Applications of NLP

Question Answering and Summarization



==> what countries speak Spanish

The language Spanish is spoken in Argentina, Aruba, Belize, Bolivia, Brazil, Canada, Cayman Islands, Chile, Colombia, Costa Rica, Cuba, Curacao, Dominican Republic, Ecuador, El Salvador, Equatorial Guinea, Falkland Islands (Islas Malvinas), Gibraltar, Guatemala, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru, Puerto Rico, Saint Martin, Sint Maarten, Spain, Switzerland, Trinidad and Tobago, United States, Uruguay, Venezuela, and Virgin Islands.

The language Castilian Spanish is spoken in Spain.

Applications of NLP

Machine Translation



[Web](#) [Images](#) [Videos](#) [Maps](#) [News](#) [Shopping](#) [Mail](#) [more ▾](#)

Google translate

Try a new browser with automatic translation.

[Download Google Chrome](#)

English

Spanish

Translate

Do you want to play a game?

¿Quieres jugar un juego?

 Listen

 Listen

New! Click the words above to view alternate translations. [Dismiss](#)

Applications of NLP

Sentiment Analysis

Customer Reviews

Apple iPad 2 MC979LL/A Tablet (16GB, Wifi, White) 2nd Generation



The most helpful favorable review

6,367 of 6,474 people found the following review helpful

★★★★★ A Step Closer

For anyone out there who is considering whether or not to make the leap and purchase the iPad 2, this review is for you. If you're still debating between the iPad 1 and the iPad 2 check out my review of the first generation iPad right here on Amazon to see a discussion of its strengths and weaknesses with a number of people commenting (both positively and negatively) over...

[Read the full review >](#)

Published 12 months ago by Craig Whisenhunt

[See more 5 star, 4 star reviews](#)

The most helpful critical review

331 of 355 people found the following review helpful

★★★★☆ honeymoon is over

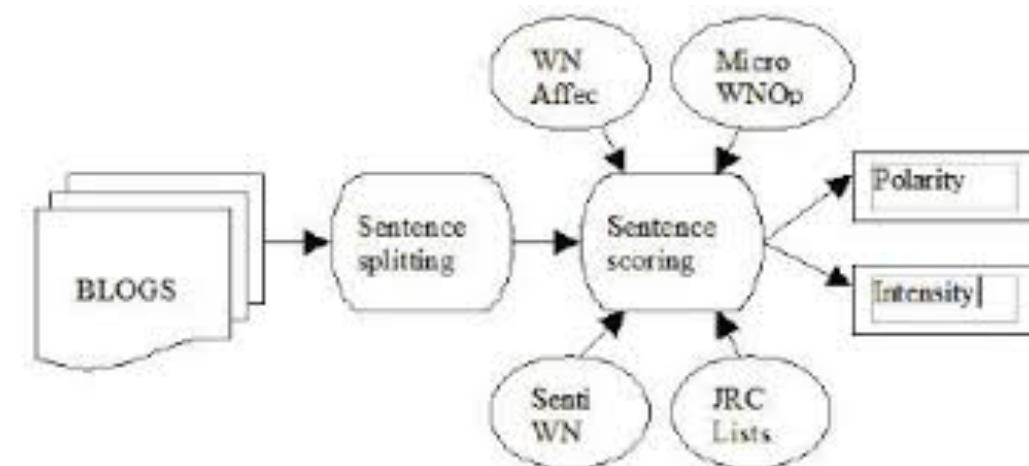
I bought this from apple because amazon didn't have it yet. I am middle-aged, love my nano, a kindle devotee from the first days, don't have a smart phone, and am a pc user. at first i was awed by the touchscreen and the resolution quality. images are gorgeous. but now, it is really annoying. too often the screen doesn't respond, and then reacts when you don't even...

[Read the full review >](#)

Published 6 months ago by bkluvr

[See more 3 star, 2 star, 1 star reviews](#)

Vs.



What makes NLP difficult

**Computers are
not brains**

- There is evidence that much of language understanding is built into the human brain

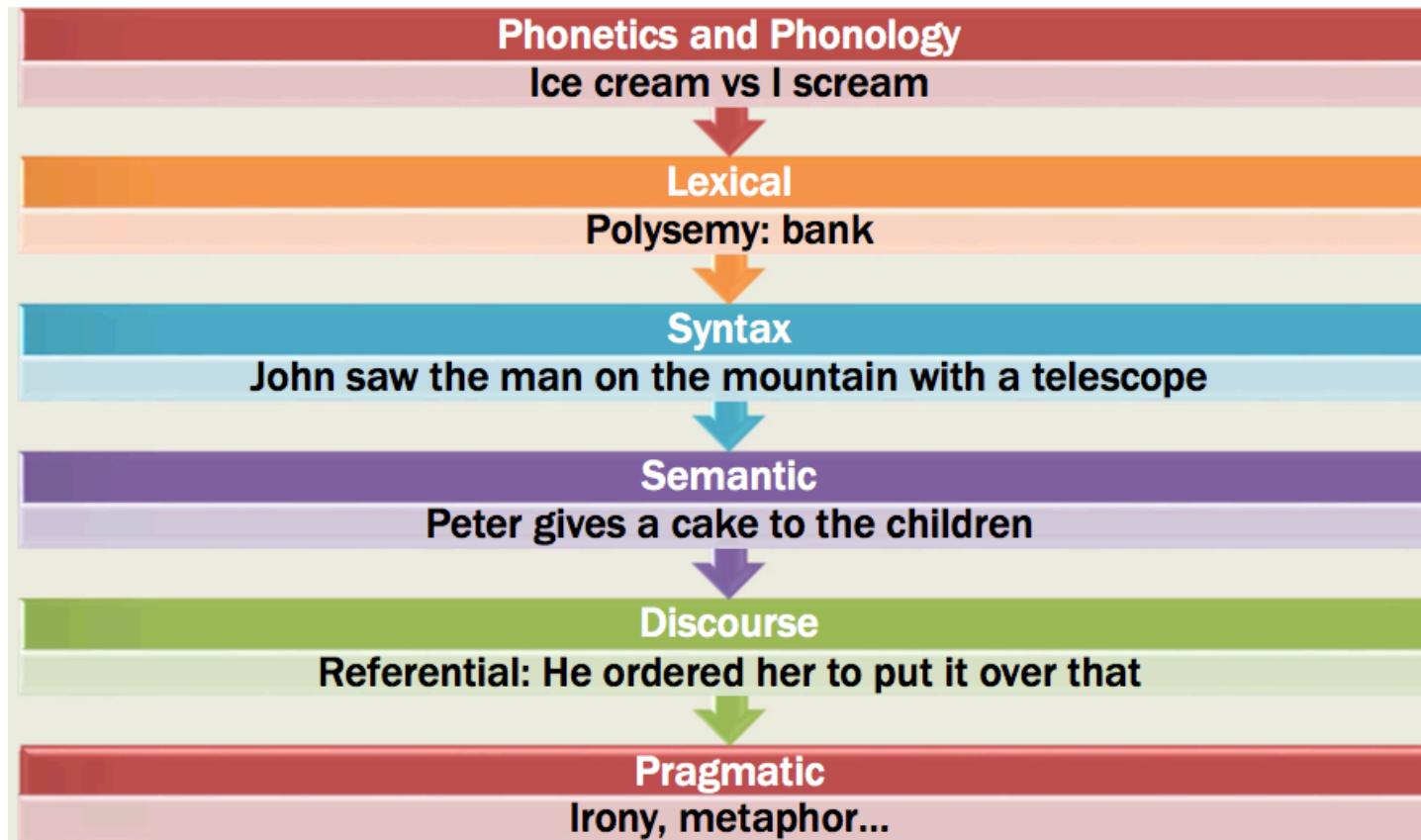
**Computers do
not socialize**

- Much of language is about communicating with people

Key problems

- Representation of meaning
- Language presupposes knowledge about the world
- Language presupposes communication between people
- Language is ambiguous

Ambiguity



Paraphrasing

Different words/sentences express the same meaning

Book delivery time

- *When will I receive my book?*
- *When will my book arrive?*

Paraphrase Detection

1. Cats eat mice.
2. Cats catch mice.

1. Boys play football.
2. Girls play soccer.

1. British PM signed deal.
2. Chinese president visited Britain.



Score: **4.60**
Almost perfect paraphrase!

Score: **1.58**
Some common elements
but generally no semantic
similarity.

Score: **0.22**
Remotely similar topic but
that's it.

Ambiguity

Same words/sentences express different meaning

Fall

- *The third season of the year*
- *Moving downwards towards the ground or towards a lower position*

The door is open

- *Could be expressing a fact*
- *Could be a request to close the door*
- *Could mean you can speak to me anytime*

Phonetics and Phonology

Two or more words sounding the same, but having different meanings

English examples:

- there – their
- here - hear
- plane – plain
- Hamburger (Citizens of Hamburg) – hamburger (burger, food)
- sea - see
- Friday - fry day
- weekend - weak end
- ice cream - I scream.
- new direction - nude erection
- new day - nude, eh?
- I don't know! - I don't - no!
- but – butt
- Wait - Weight
- psychotherapist - psycho the rapist
- You're unconscious now... - Your unconscious now...
- Your students... - You're students...
- Two - too - to

German examples:

- Du hast Gewehre. (You have got guns.) - Du hasst Gewehre. (You hate guns.)
- Lehrer (teacher) – leerer (emptier)

“I have a **knot** question. Will you **not** progress faster by sailing at 20 knots rather than railing about how you should **not** sail at all? Do you **know knots** or do you **know not?**”

And, you know, it would be all **right** for you to **write** and practice these patterns every day, because you want to master these skills, **right?**

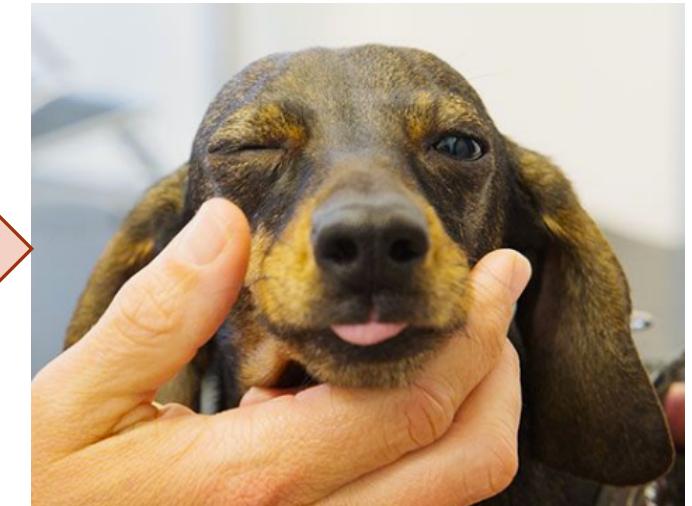
Syntax Ambiguity

In syntactic ambiguity, the same sequence of words is interpreted as having different syntactic structures



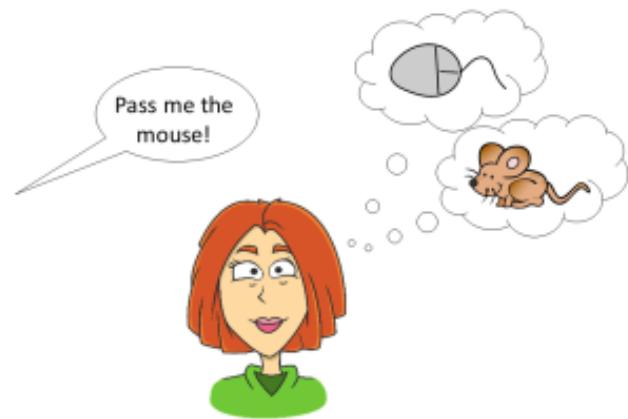
Look at the dog with one eye

I saw the man with a telescope

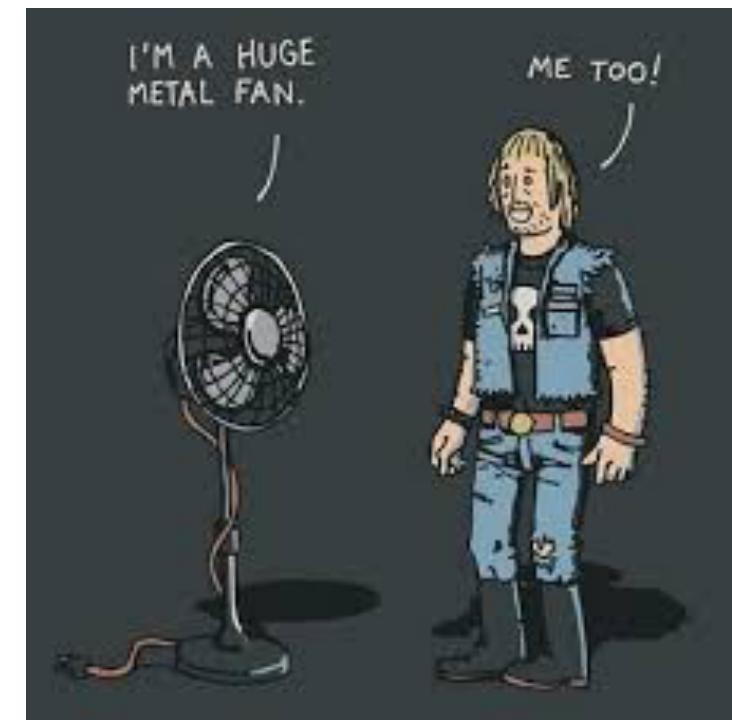


Semantic Ambiguity

In **semantic ambiguity** the structure remains the same, but the individual words are interpreted differently



The Astronomer loves the star
- *Star in the sky?*
- *Celebrity?*



Pragmatic Ambiguity

Pragmatics concerns the overall communicative and social context and its effect on interpretation

HI & LOIS



Can you pass the salt?

- Request to pass the salt ?
- Physical ability to pass the salt ?

Clouseau: Does your dog bite?

Hotel Clerk: No.

Clouseau: [bowing down to pet the dog] Nice doggie.
[Dog barks and bites Clouseau in the hand]

Clouseau: I thought you said your dog did not bite!

Hotel Clerk: That is not my dog.

Discourse Ambiguity

Relates to entities previously introduced and processing it requires shared knowledge or world.
Interpretation is carried out using this context

Alice understands that you like your mother, but she ...

- Does **she** refer to Alice?
- Does **she** refer to your mother?

The horse ran up the hill. It was very steep. It soon got tired

- Does **it** refer to hill?
- Does **it** refer to horse?

NLP - Techniques

Splitting

*Splitting Text into
Section*

*Splitting Sections into
Sentences*

*Splitting Sections into
Words/Ngrams*

NLP - Techniques

Part of Speech tagging

I ate the spaghetti with meatballs.

Pro V Det N Prep N

John saw the saw and decided to take it to the table.

PN V Det N Con V Part V Pro Prep Det N

Stanford Parser

Please enter a sentence to be parsed:

Surgical resection specimens of 85 invasive ductal carcinomas of 85 women who had undergone 3D ultrasound were included.

Language: English ▾ Sample Sentence

Parse

Your query

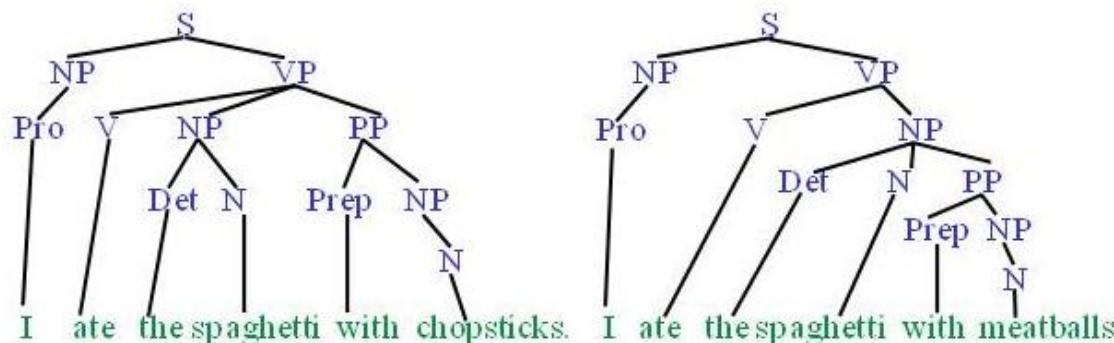
Surgical resection specimens of 85 invasive ductal carcinomas of 85 women who had undergone 3D ultrasound were included.

Tagging

Surgical/NNP resection/NN specimens/NNS of/IN 85/CD invasive/JJ ductal/JJ carcinomas/NNS of/IN 85/CD women/NNS who/WP had/VBD undergone/VBN 3D/CD ultrasound/NN were/VBD included/VBN ./.

NLP - Techniques

Parsing Syntactic Tree



Parse

```
(ROOT
  (S
    (NP
      (NP (NNP Surgical) (NN resection) (NNS specimens))
      (PP (IN of)
        (NP
          (NP (CD 85) (JJ invasive) (JJ ductal) (NNS carcinomas))
          (PP (IN of)
            (NP
              (NP (CD 85) (NNS women))
              (SBAR
                (WHNP (WP who))
                (S
                  (VP (VBD had)
                  (VP (VBN undergone)
                    (NP (CD 3D) (NN ultrasound)))))))))))
        (VP (VBD were)
        (VP (VBN included))))
      (. .)))
```

NLP - Techniques

*Named entity
recognition*

becas Annotate Help API Widget About Contact

HIGHLIGHT

All None

Anatomy

Disorders

Chemicals

Genes and Proteins

Cellular Components

Molecular Functions

Biological Processes

Ambiguous

New to becas? Take the tour »

In Duchenne muscular dystrophy (DMD), the infiltration of skeletal muscle by immune cells aggravates disease, yet the precise mechanisms behind these inflammatory responses remain poorly understood. Chemotactic cytokines, or chemokines, are considered essential recruiters of inflammatory cells to the tissues. We assayed chemokine and chemokine receptor expression in DMD muscle biopsies ($n = 9$, average age 7 years) using immunohistochemistry, immunofluorescence, and in situ hybridization. CXCL1, CXCL2, CXCL3, CXCL8, and CXCL11, absent from normal muscle fibers, were induced in DMD myofibers. CXCL11, CXCL12, and the ligand-receptor couple CCL2-CCR2 were upregulated on the blood vessel endothelium of DMD patients. CD68(+) macrophages expressed high levels of CXCL8, CCL2, and CCL5. Our data suggest a possible beneficial role for CXCR1/2/4 ligands in managing muscle fiber damage control and tissue regeneration. Upregulation of endothelial chemokine receptors and CXCL8, CCL2, and CCL5 expression by cytotoxic macrophages may regulate myofiber necrosis.

Load text Export ▾

Annotated 46 concept occurrences in 0.173s.

Concept Tree

- + Expand All - Collapse All Toggle All
- + **Anatomy (12)**
 - **Disorders (4)**
 - **DMD (1)**
 - **Duchenne muscular dystrophy (1)**
 - **Infiltration (1)**
 - **Inflammatory responses (1)**
 - + **Chemicals (2)**
 - + **Genes and Proteins (11)**
 - + **Cellular Components (3)**
 - + **Molecular Functions (1)**
 - + **Biological Processes (9)**

NLP - Techniques

Word Sense Disambiguation

Analysis with definitions(s)

Bill Gates has developed an interest/[readiness to give attention] in language technology and yesterday acquired a 10 % interest/[a share (in a company, business, etc.)] in Torbjörn Lager 's sense disambiguation technology . Lager will retain a 90 % interest/[a share (in a company, business, etc.)] in the new company , which will be based in Göteborg , Sweden . Last year 's drop in interest/[money paid for the use of money] rates will probably be good for the company . Finally , although all this may sound like an arcane maneuver of little interest/[quality of causing attention to be given] outside Wall Street , it would set off an economical earthquake .

These are the six senses of the noun *interest* according to the LDOCE:

Sense	Definition
1	readiness to give attention
2	quality of causing attention to be given
3	activity, subject, etc., which one gives time and attention to
4	advantage, advancement, or favour
5	a share (in a company, business, etc.)
6	money paid for the use of money

Assignment

For some humor !

What type of ambiguity?

1. Groucho Marx: One morning I shot an elephant in my pajamas. How he got into my pajamas, I'll never know.
2. She criticized my apartment, so I knocked her flat.
3. Noah took all of the animals on the ark in pairs. Except the worms, they came in apples.
4. Policeman to little boy: "We are looking for a thief with a bicycle." Little boy: "Wouldn't you be better using your eyes."
5. Why is the teacher wearing sun-glasses. Because the class is so bright.

Natural Language Processing

Break Time

Demo - 1

*Introduction to Text Analysis using
Tidy Text*

Demo - 2

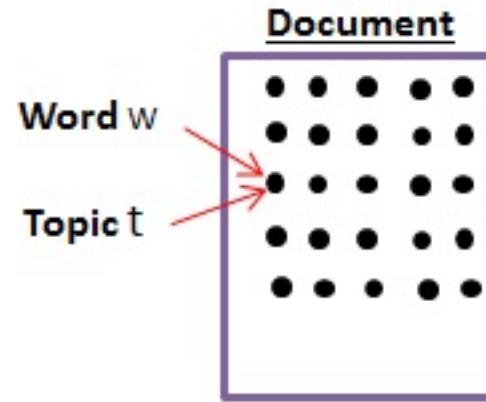
Introduction to Sentiment Analysis

Demo - 3

Twitter Data Analysis

Latent Dirichlet Allocation (LDA)

A Bayesian Unsupervised Learning Model



Topics	
gene	0.04
dna	0.02
genetic	0.01
...	...
life	0.02
evolve	0.01
organise	0.01
...	...
brain	0.04
neuron	0.02
nerve	0.01
...	...
data	0.02
number	0.02
computer	0.01
...	...

Documents

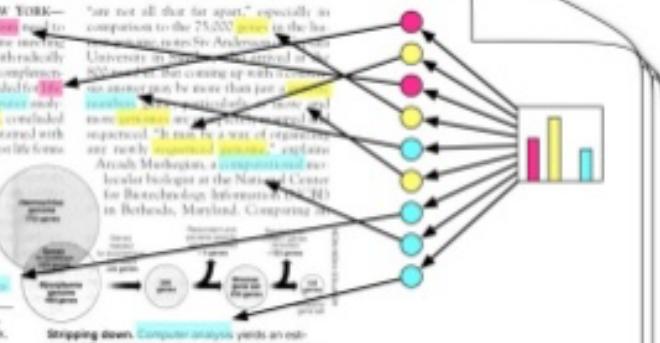
Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many *genes* does an *organism* need to *survive*? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using *biochemical analysis* to compare known *genomes*, concluded that only 300 genes can be satisfied with just 250 acres, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a single parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, these predictions

* Genome Mapping and Sequencing. Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

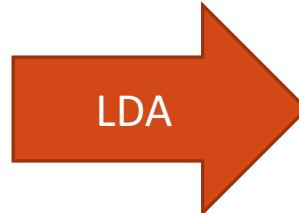


- Each Document is a mixture of Topics
- Each Topic is a mixture of words
- Each word is drawn from one of those topics

Latent Dirichlet Allocation (LDA)

It's a way of automatically discovering **topics** sentences contain

1. I like to eat broccoli and bananas.
2. I ate a banana and spinach smoothie for breakfast.
3. Chinchillas and kittens are cute.
4. My sister adopted a kitten yesterday.
5. Look at this cute hamster munching on a piece of broccoli



Sentences 1 and 2: 100% Topic A

Sentences 3 and 4: 100% Topic B

Sentence 5: 60% Topic A, 40% Topic B

Topic A: 30% broccoli, 15% bananas, 10% breakfast, 10% munching, ... (**topic A could be about food**)

Topic B: 20% chinchillas, 20% kittens, 20% cute, 15% hamster, ... (**topic B could be about cute animals**)

Demo - 4

Topic Modeling using LDA

Course Assignments

Programming Assignments

Reading Assignments

Presentation Assignments

Technical Skills Assignments

Writing Assignments

Twitter Case

Twitter User Analysis

Build a set of tools to later (next week) use to build a Shiny App

Tools:

1. Build an LDA model using tweets downloaded from a few distinct topics
2. Given a twitter user (fairly famous users)
 1. Based on his tweets, guess the topics he is interested in and compare it to his description
 2. Do a sentiment analysis of the timeline of his/her tweets and present it
 3. Display a word cloud or bag of words of what the twitter user uses
 4. Find and compare his top 5 best followers
3. Given a tweet, estimate the top 3 topics it might be classified as and show it graphically

Technical Assignment

Teach yourself Shiny

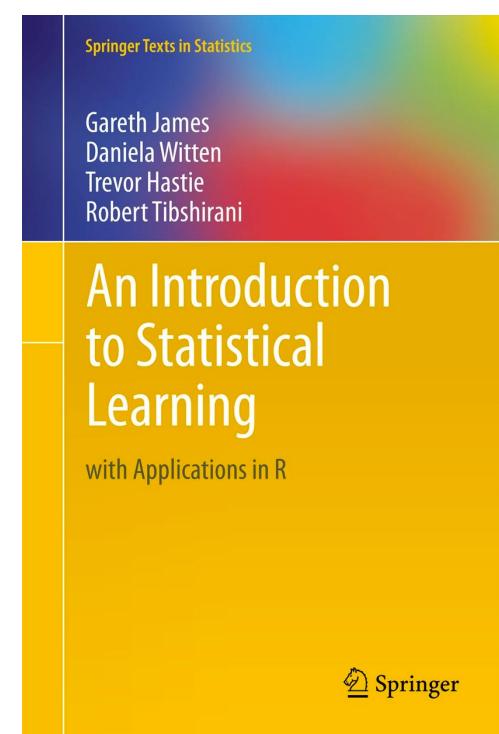
<https://shiny.rstudio.com/tutorial/>

Programming Assignment

Code Submission on Github for Twitter Case

Reading Assignment

Read Chapter 6: *Linear Model Selection & Regularization*



Writing Assignment

*Submit by Saturday
Written Report (not to exceed 15 pages) on Twitter Case*

Presentation Assignment

By Saturday Submit

Your Presentations on Twitter Case

1. Technical Presentation
2. Business Presentation (Not to exceed 5 slides)

Natural Language Processing

Lunch