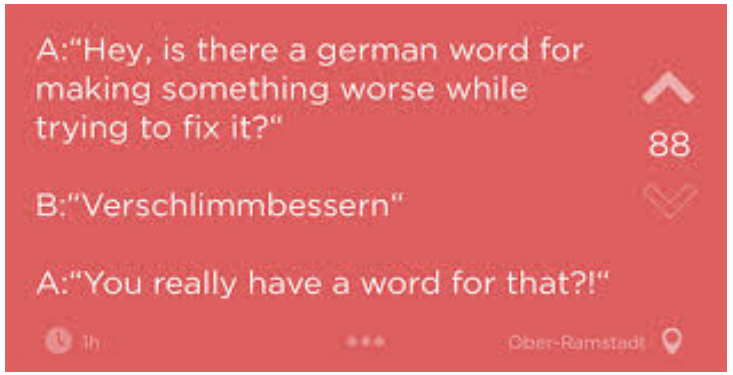
DSLA
DATA SCIENCE LEARNING ACADEMY

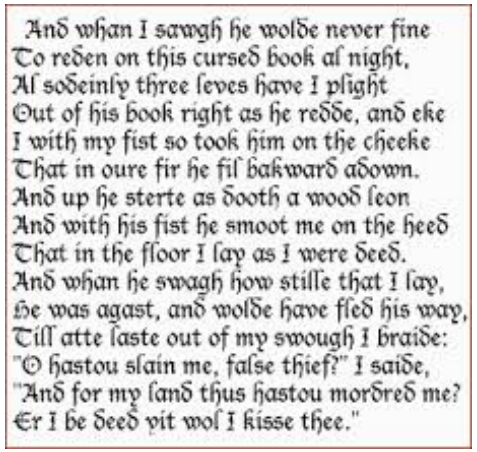
Introduction to NLP

DSLA COURSE

ROHIT PADEBETTU

Types of Language





Artificial Language

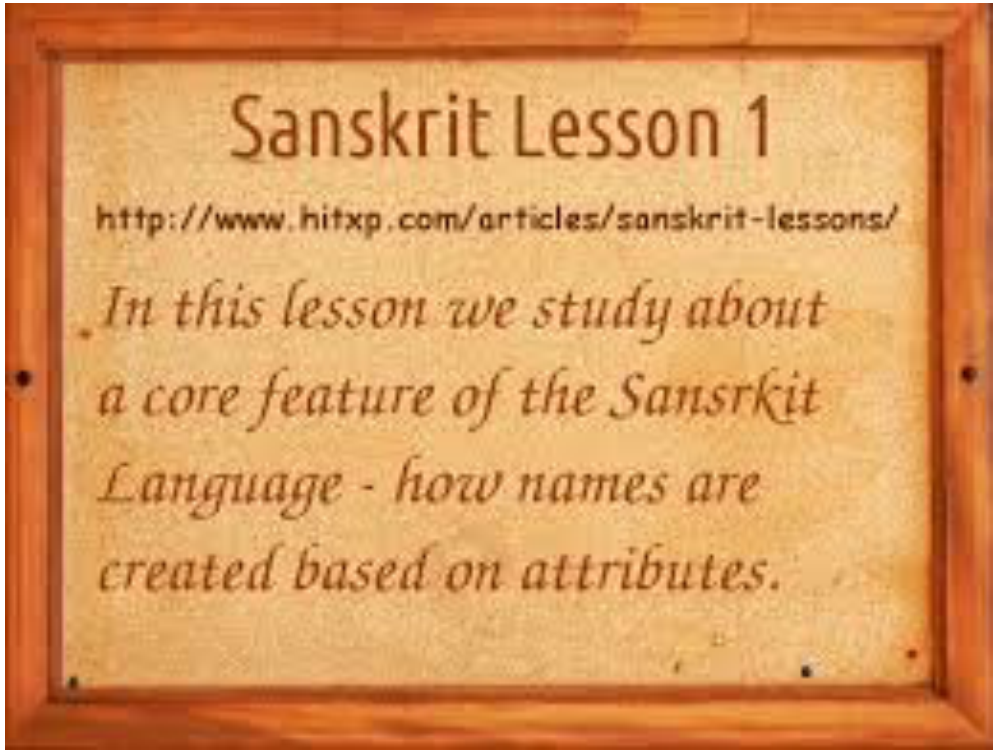
```
def add5(x):
   return x+5

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

```
      IS-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

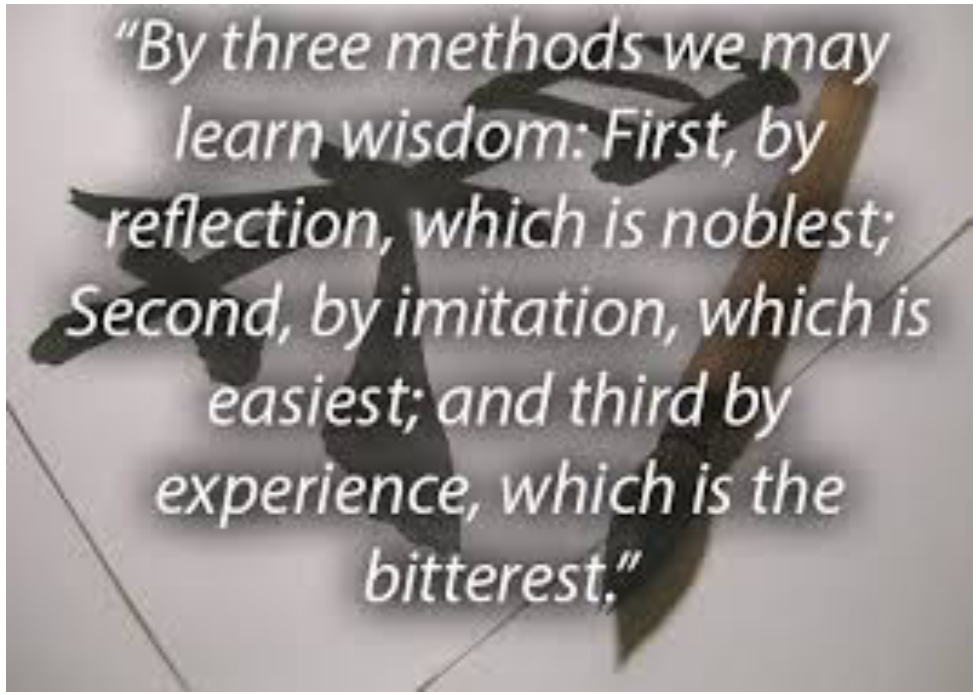
What is a Text?

A Text is composed of a sequence of words from Vocabulary

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

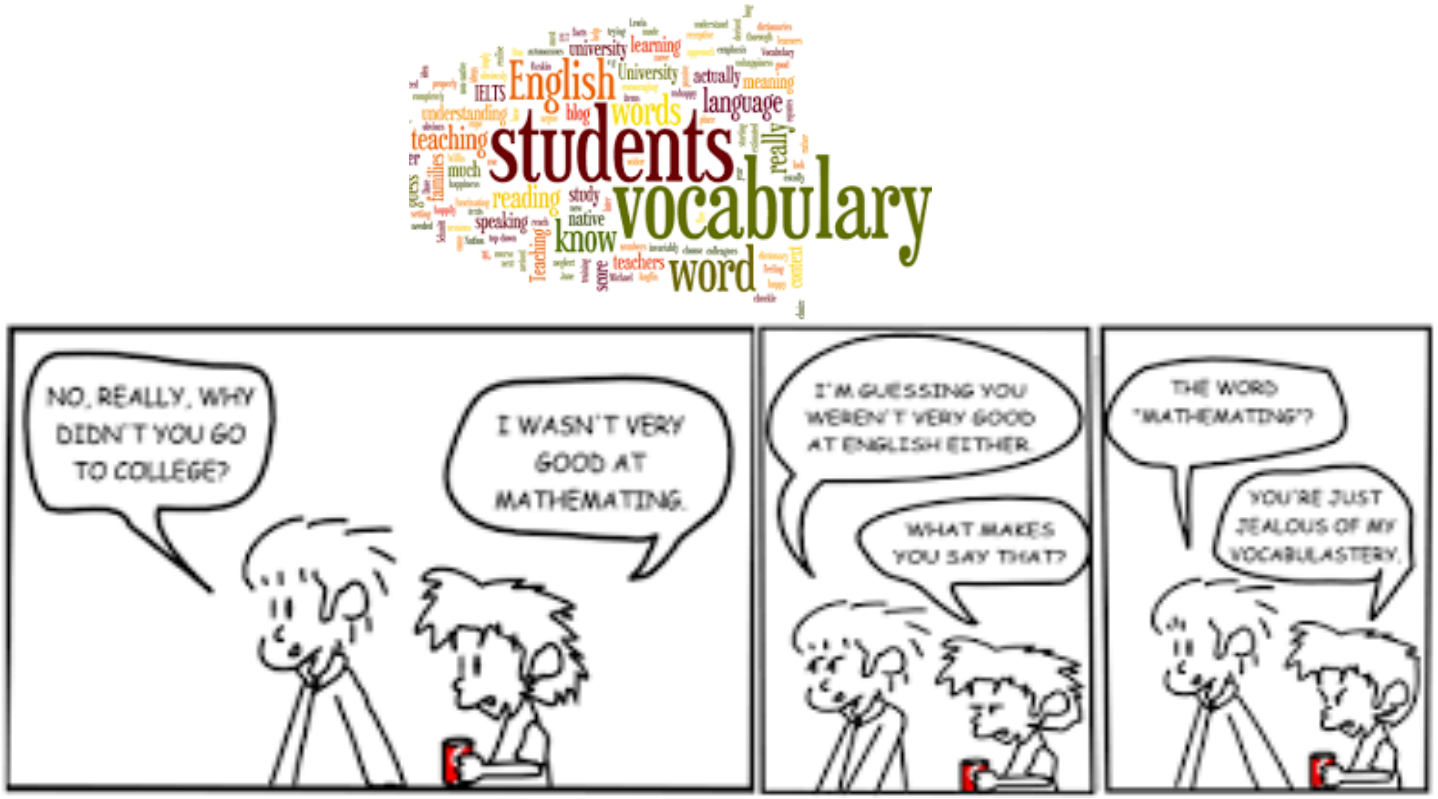
Om, Asato Maa Sad Gamaya;
Tamaso Maa Gyotira Gamaya;
Mrityora Maa Amritam Gamaya

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



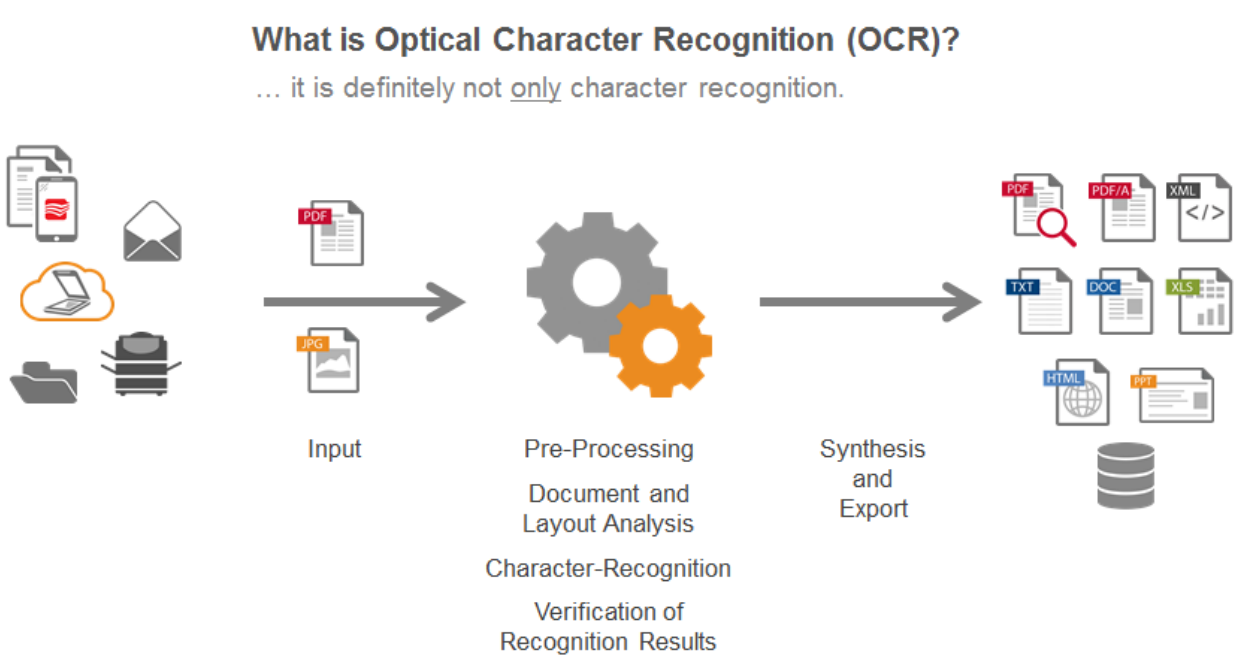
What is a Vocabulary?

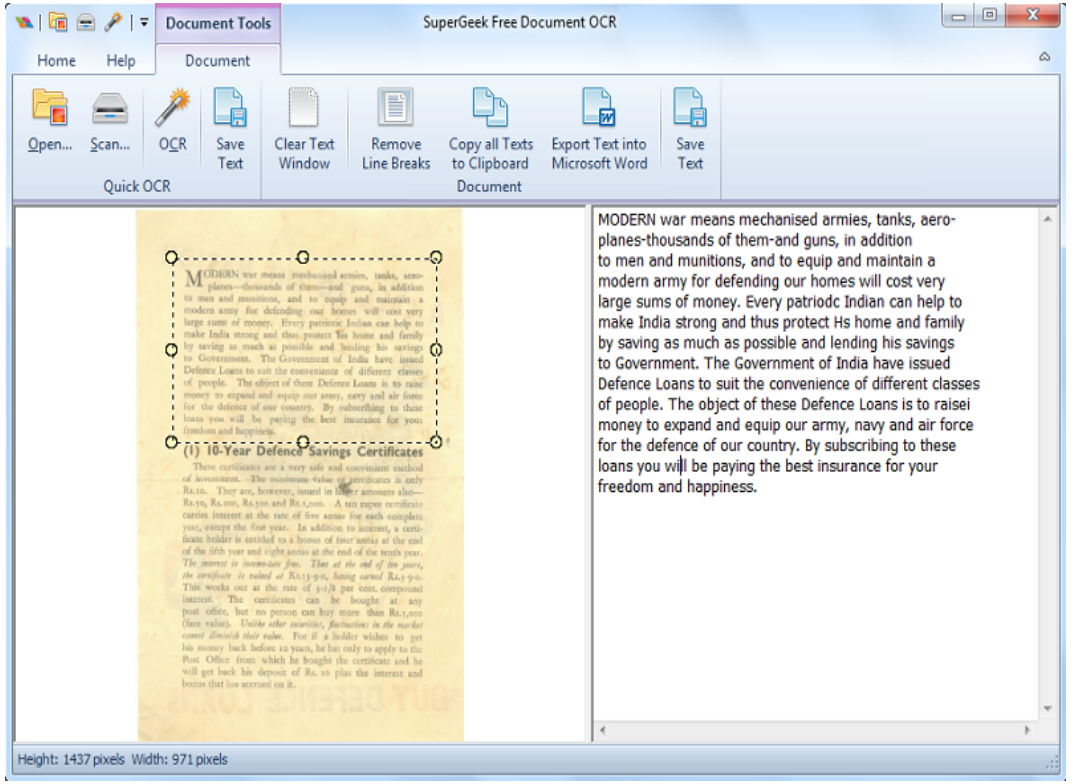
A Vocabulary is composed of a set of words



Applications of NLP

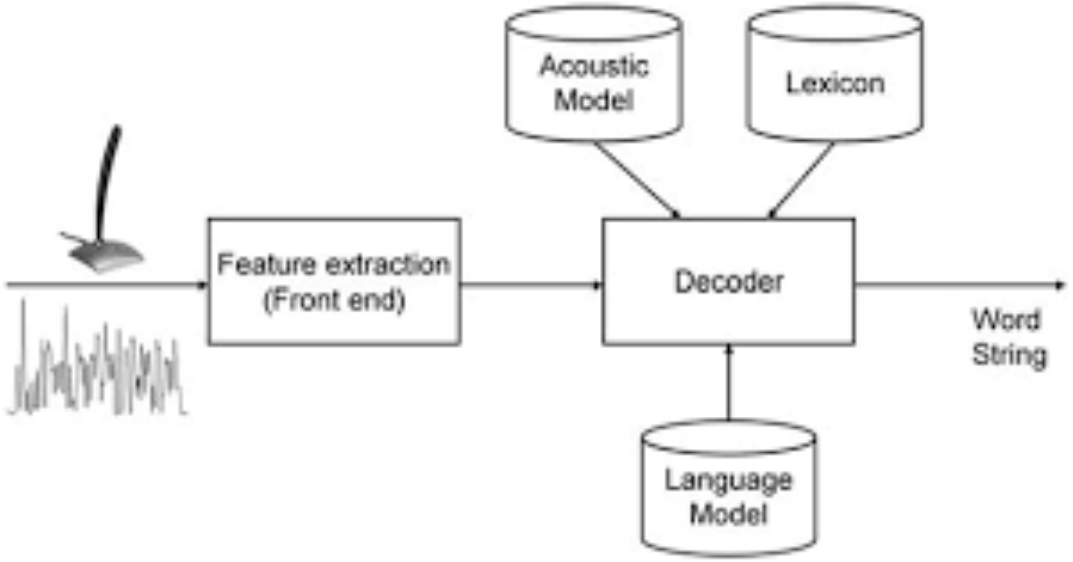
Optical Character Recognition





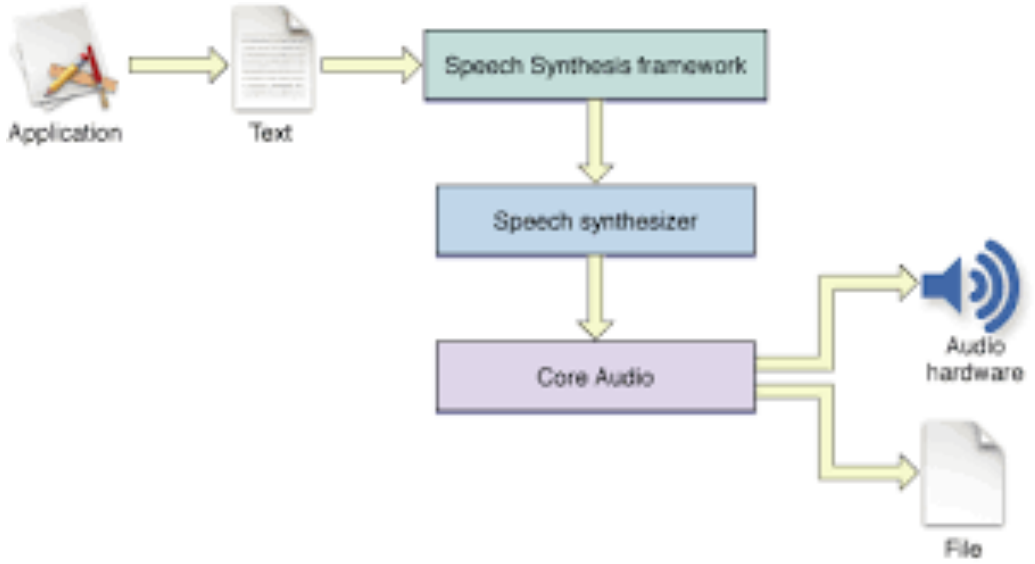
Applications of NLP

Speech Recognition



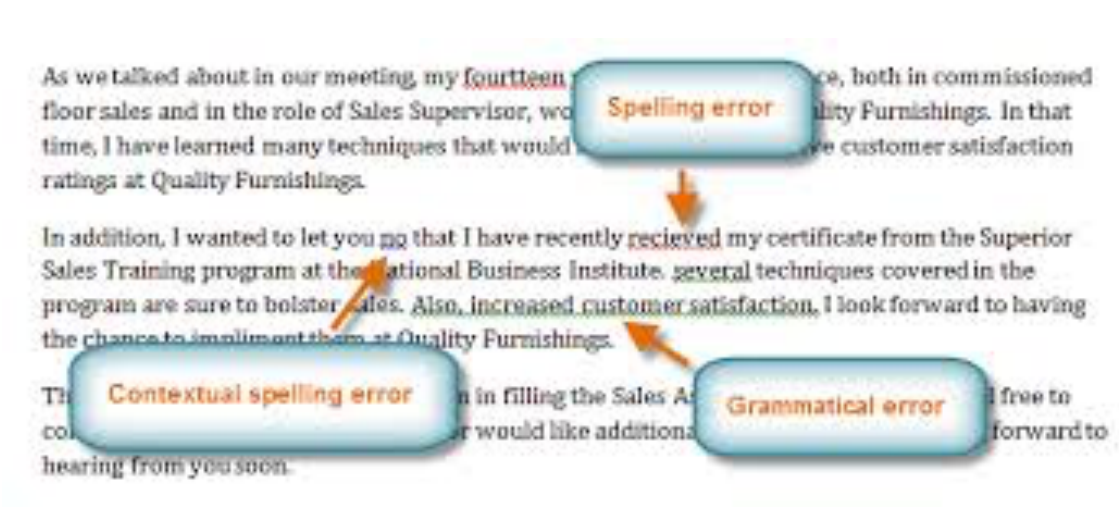
Applications of NLP

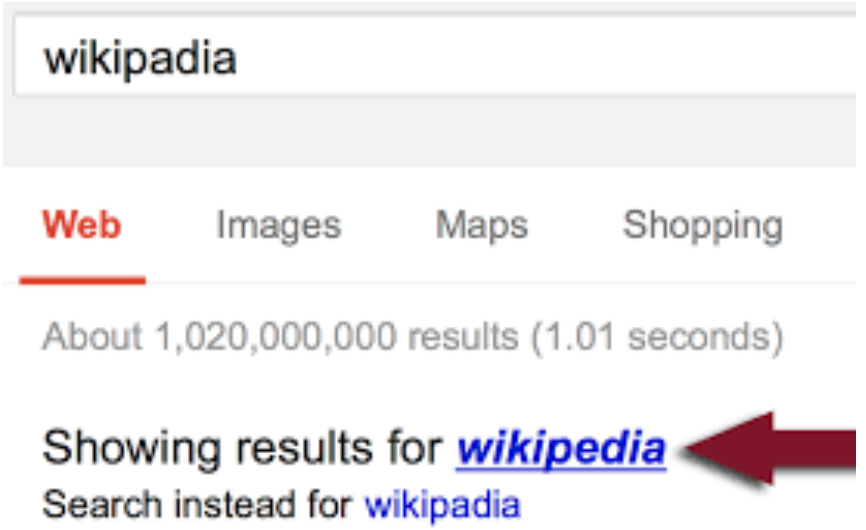
Speech Synthesis & Spoken dialog systems



Applications of NLP

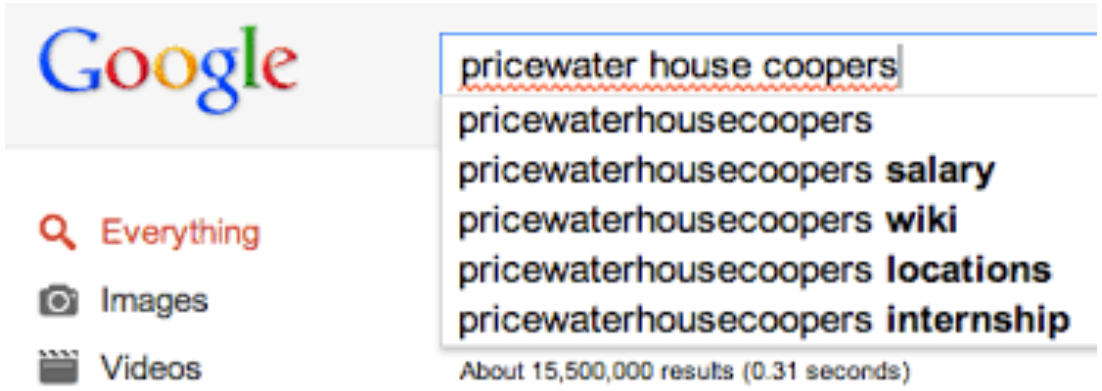
Spelling and Grammar Check





Applications of NLP

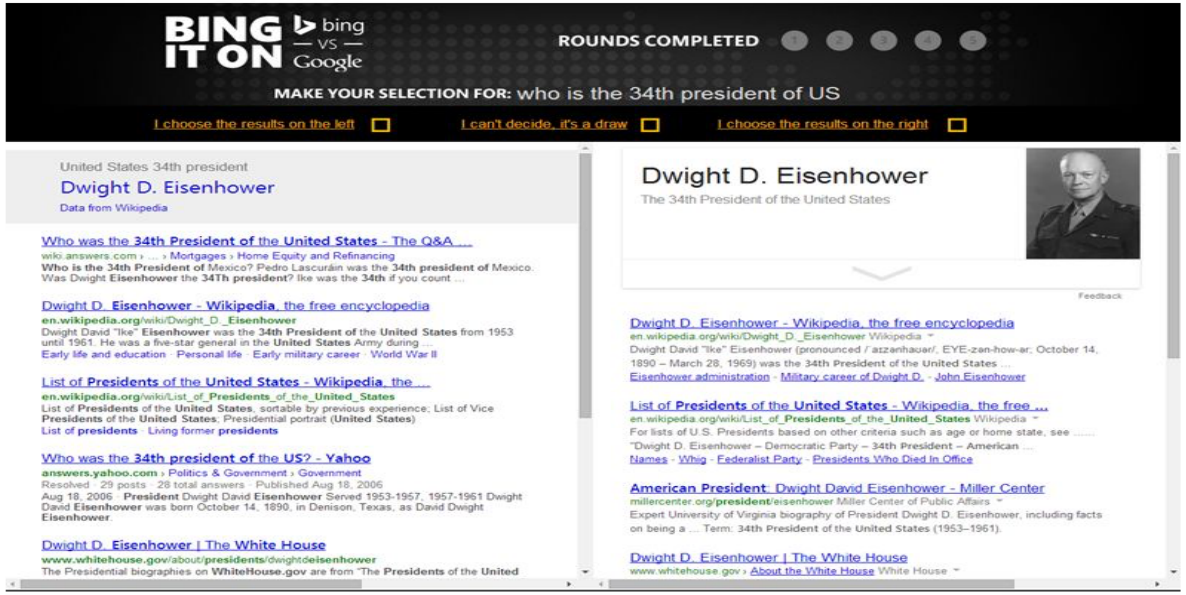
Word Prediction



Applications of NLP

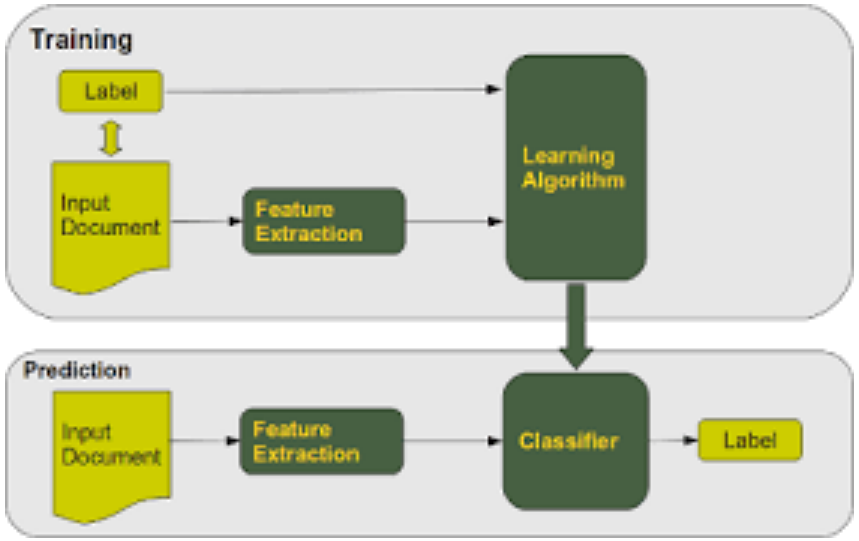
Information Retrieval

Bing v.s. Google?



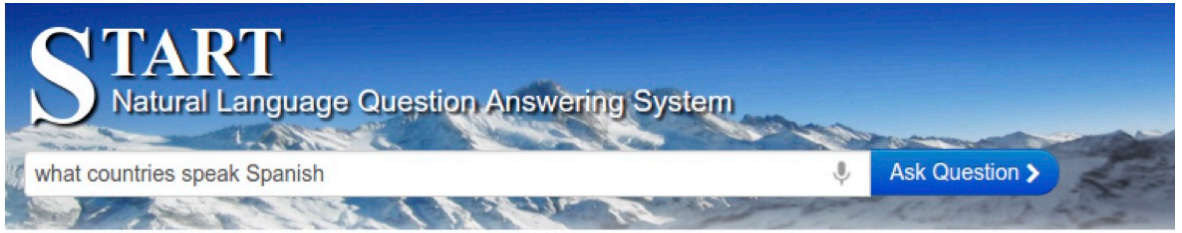
Applications of NLP

Document Categorization



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.

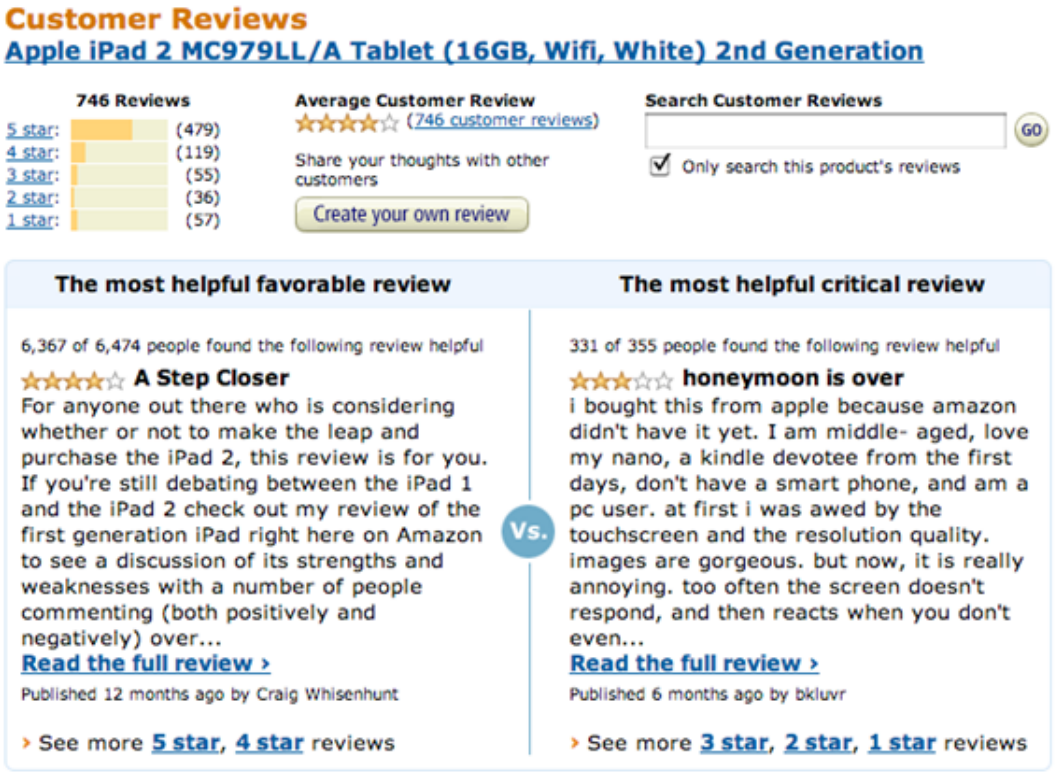
The language Castilian Spanish is spoken in Spain.

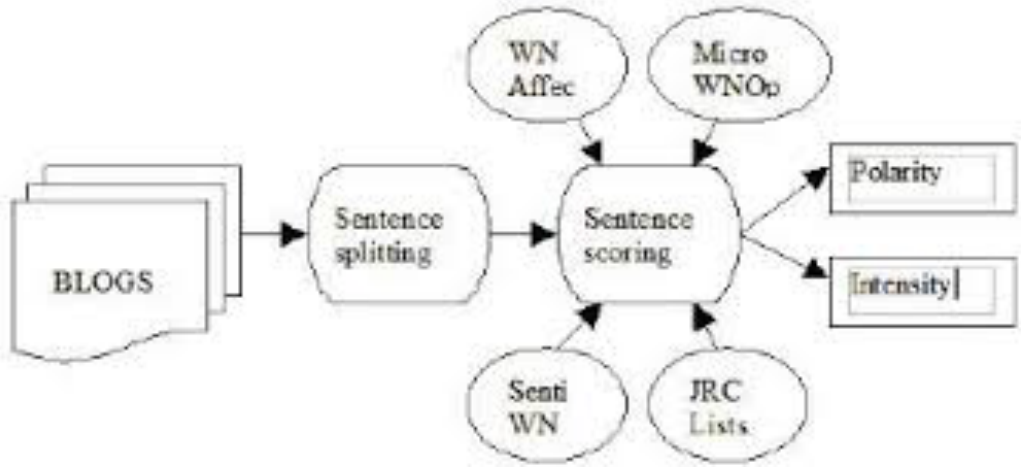
Applications of NLP

Machine Translation

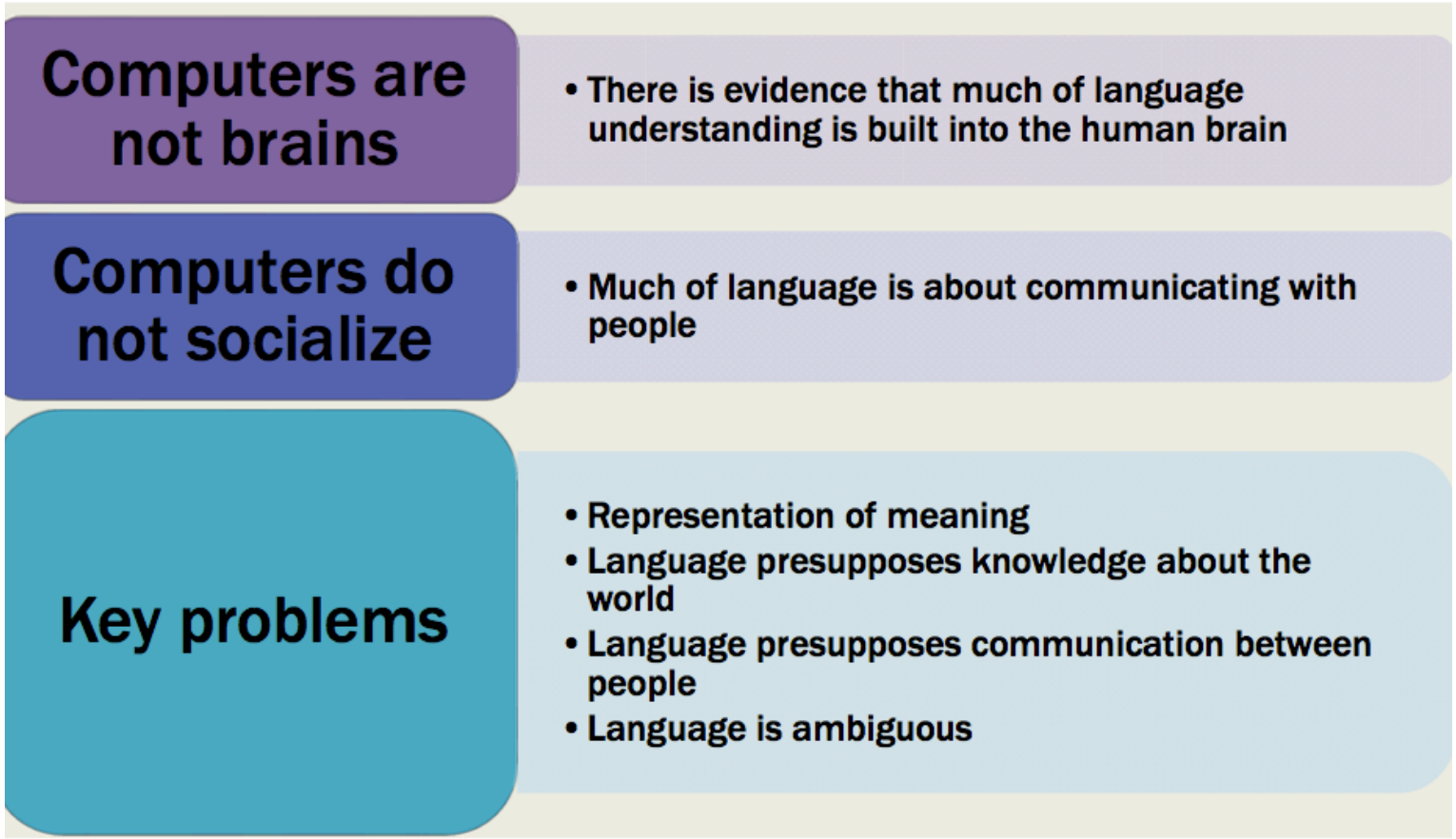
Applications of NLP

Sentiment Analysis

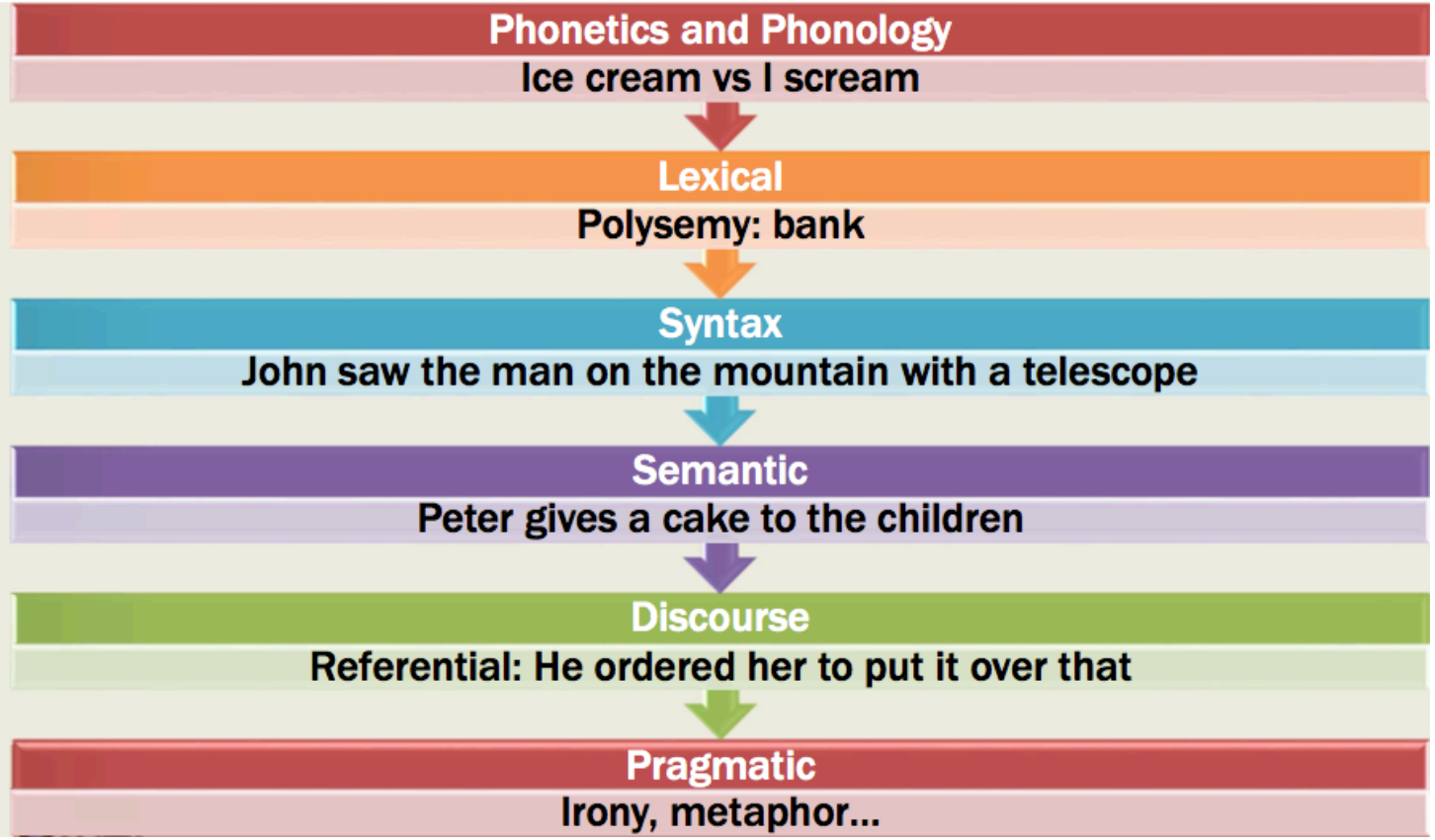




What makes NLP difficult



Ambiguity

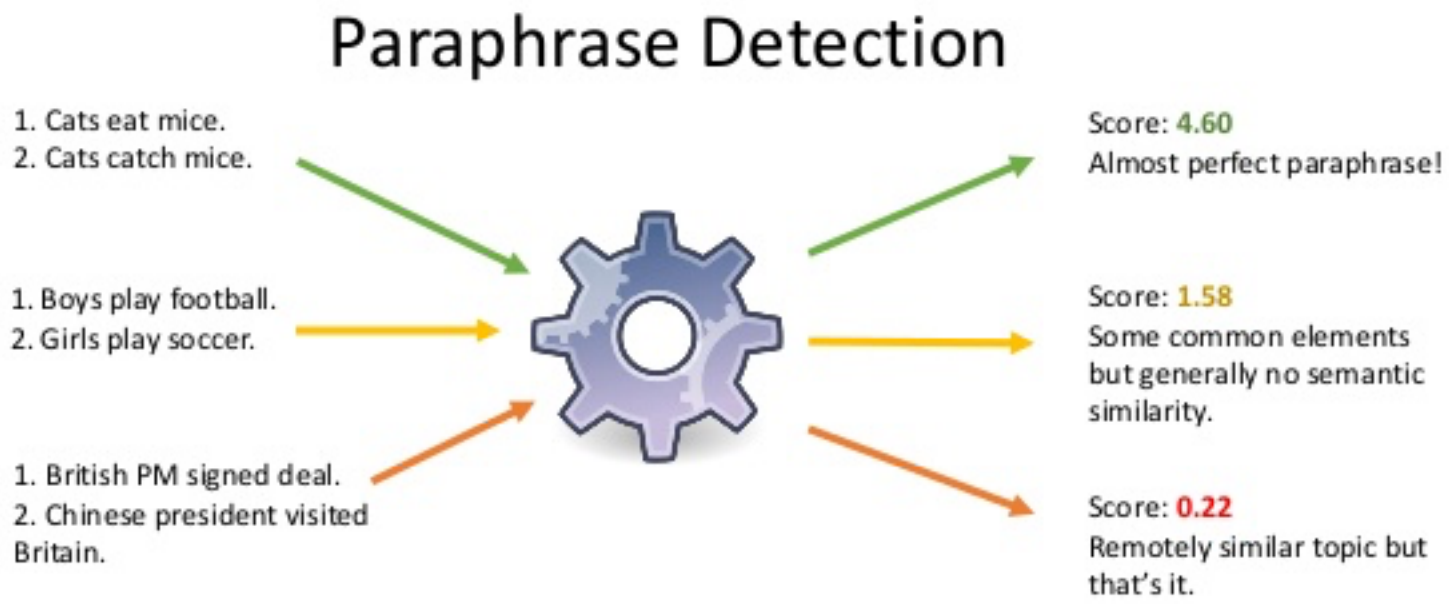

Phonetics and Phonology
Ice cream vs I scream
Lexical
Polysemy: bank
Syntax
John saw the man on the mountain with a telescope
Semantic
Peter gives a cake to the children
Discourse
Referential: He ordered her to put it over that
Pragmatic
Irony, metaphor...

Paraphrasing

Different words/sentences express the same meaning

Book delivery time

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



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**?"

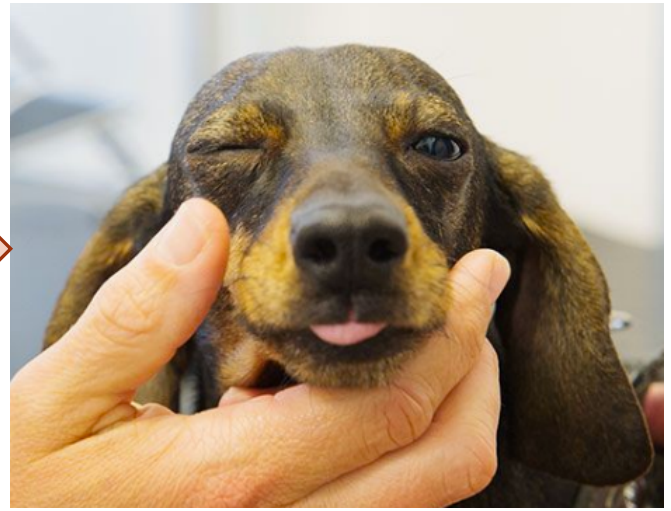
And, you know, it would be all **right** for you to **write** and practice these patterns every day, because you want to master these skills, **right**?

Syntax Ambiguity

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

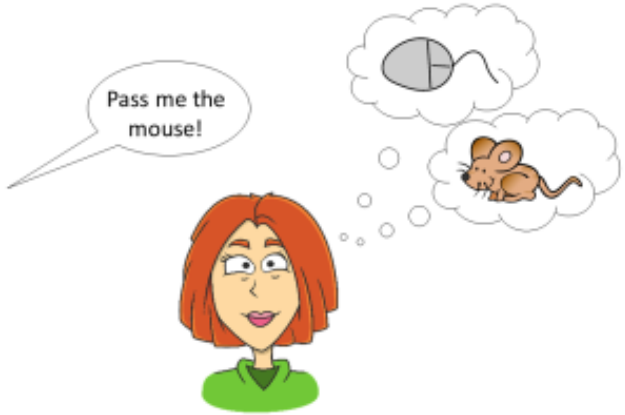
I saw the man with a telescope

Look at the dog with one eye



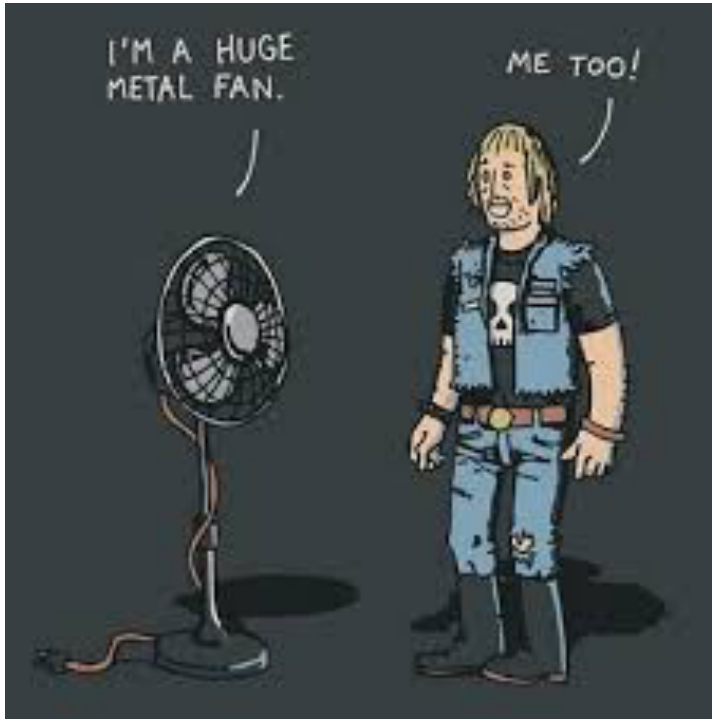
Semantic Ambiguity

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



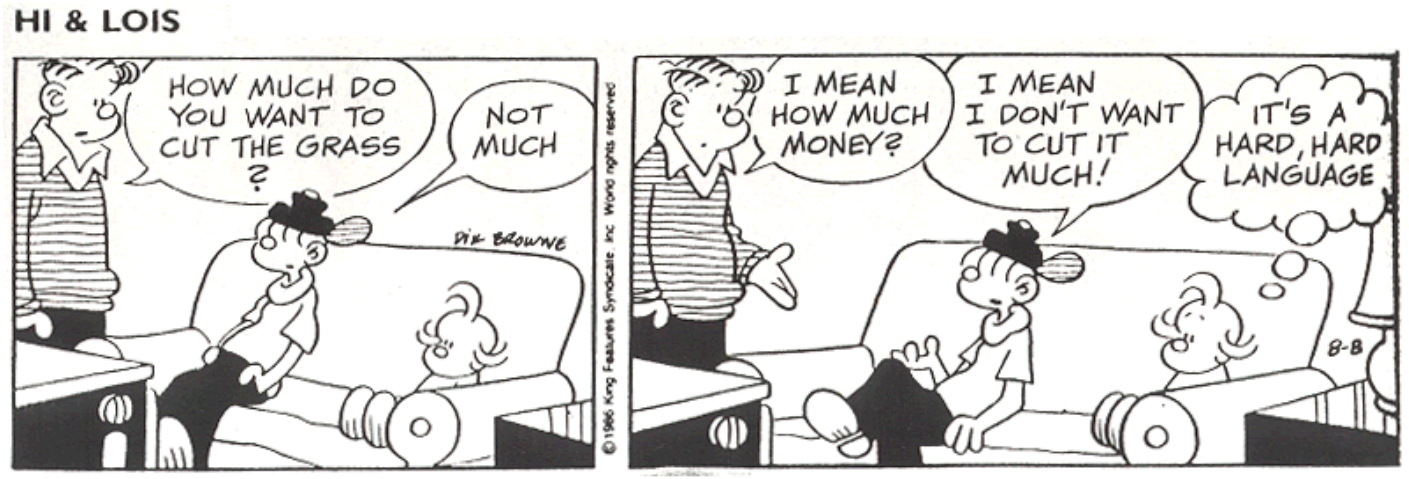
The Astronomer loves the star

- *Star in the sky?*
- *Celebrity?*



Pragmatic Ambiguity

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



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

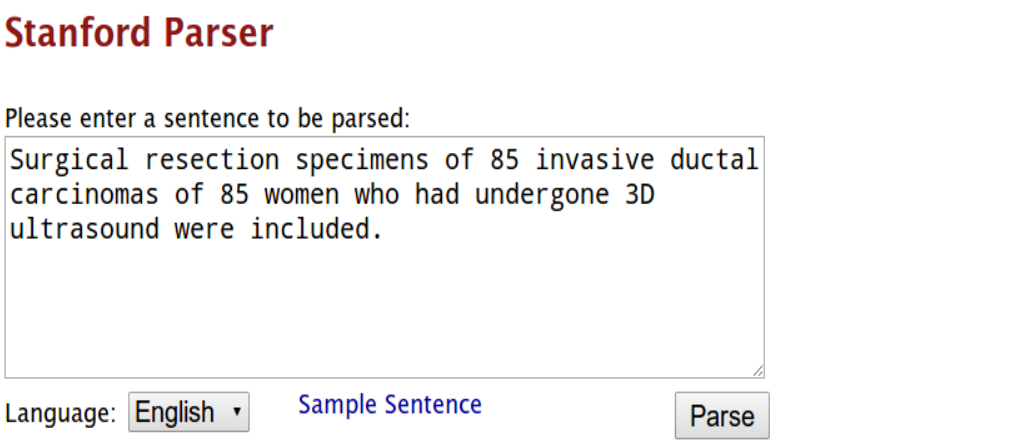
Splitting Sections into Words/Ngrams

NLP - Techniques

Part of Speech tagging

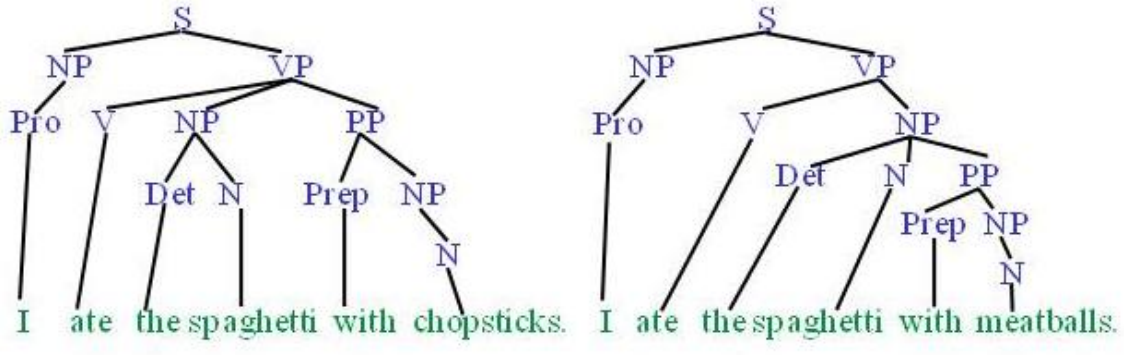
I ate the spaghetti with meatballs.
Pro V Det N Prep N

John saw the saw and decided to take it to the table.
PN V Det N Con V Part V Pro Prep Det N



NLP - Techniques

Parsing Syntactic Tree

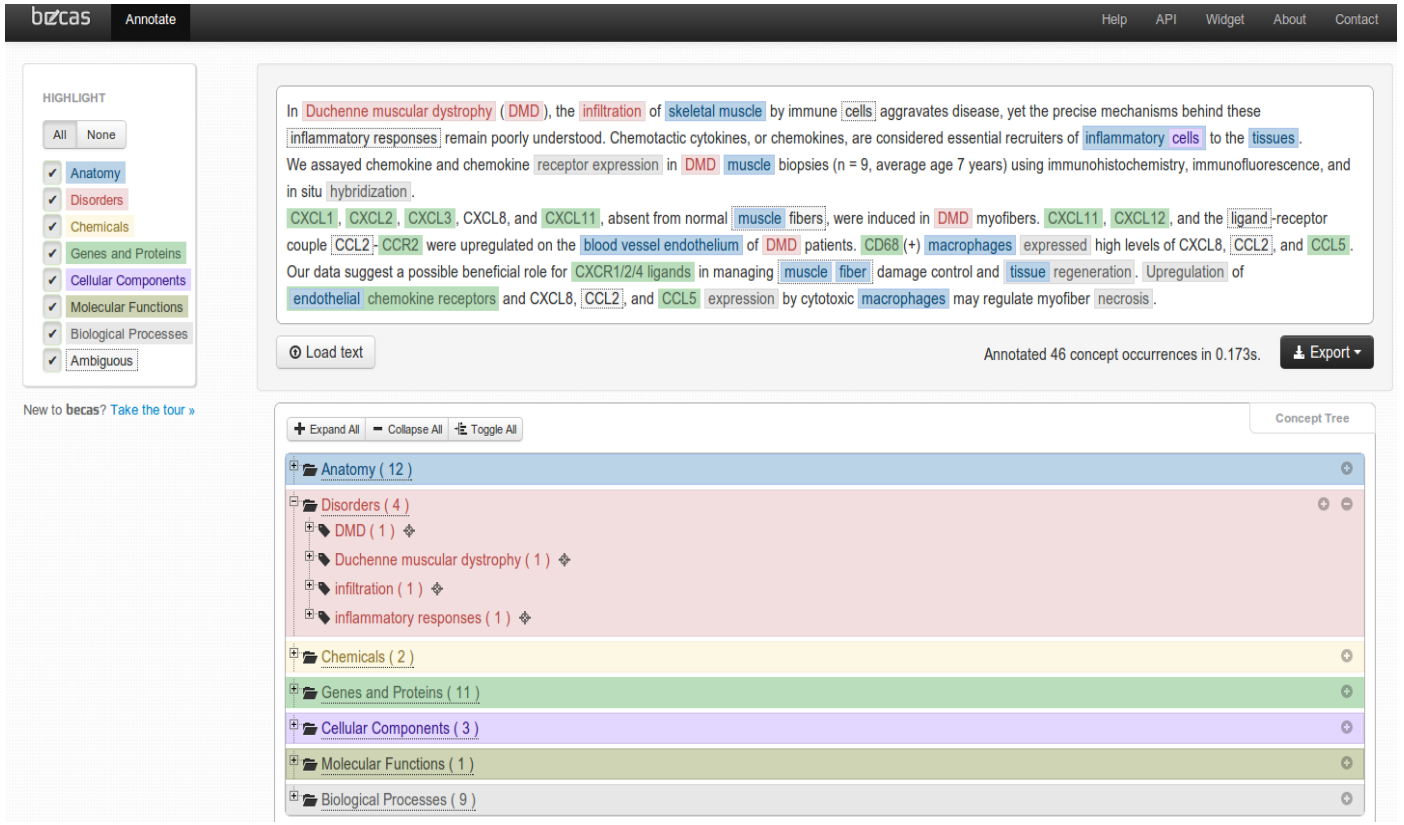


Parse

```
(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



NLP - Techniques

Word Sense Disambiguation

Analysis with definitions(s)

*Bill Gates has developed an interest/***[readiness to give attention]** *in language technology and yesterday aquired 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

DSLA
DATA SCIENCE LEARNING ACADEMY

Break Time

Demo - 1

DSLA
DATA SCIENCE LEARNING ACADEMY

Introduction to Text Analysis using Tidy Text

Demo - 2

DSLA
DATA SCIENCE LEARNING ACADEMY

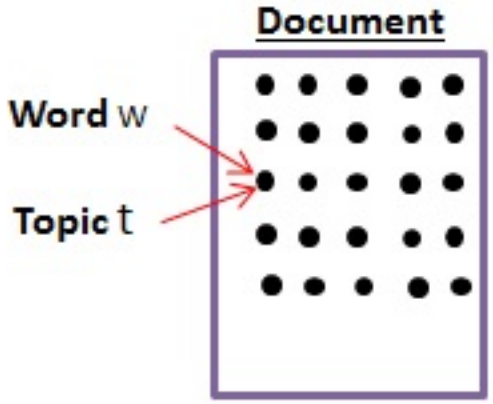
Introduction to Sentiment Analysis

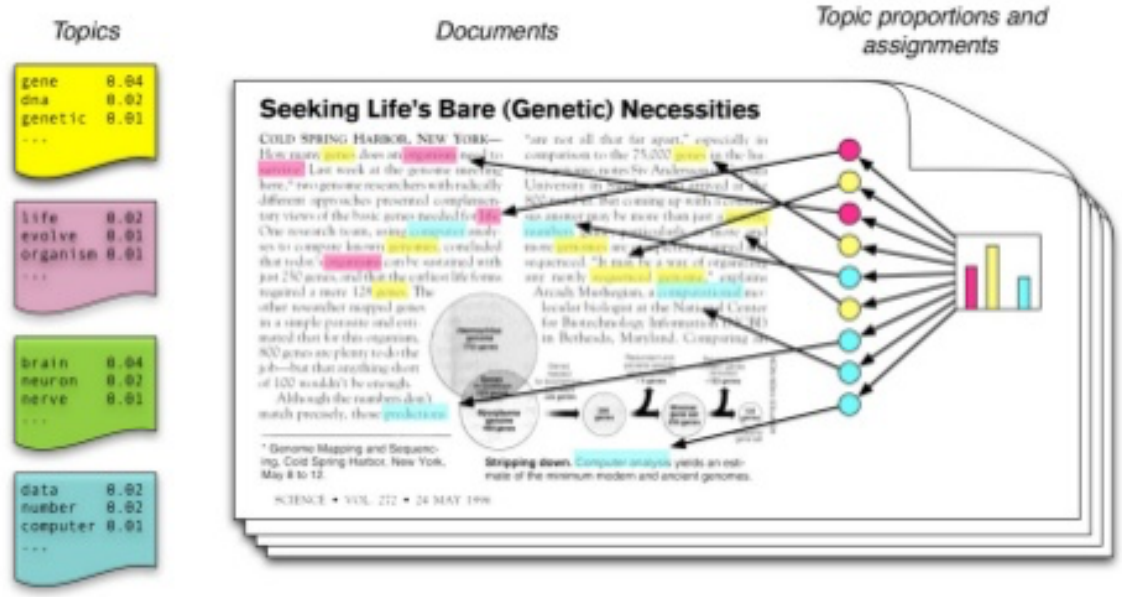
Demo - 3

Twitter Data Analysis

Latent Dirichlet Allocation (LDA)

A Bayesian Unsupervised Learning Model



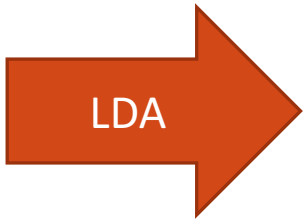


- Each Document is a mixture of Topics
- Each Topic is a mixture of words
- Each word is drawn from one of those topics

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

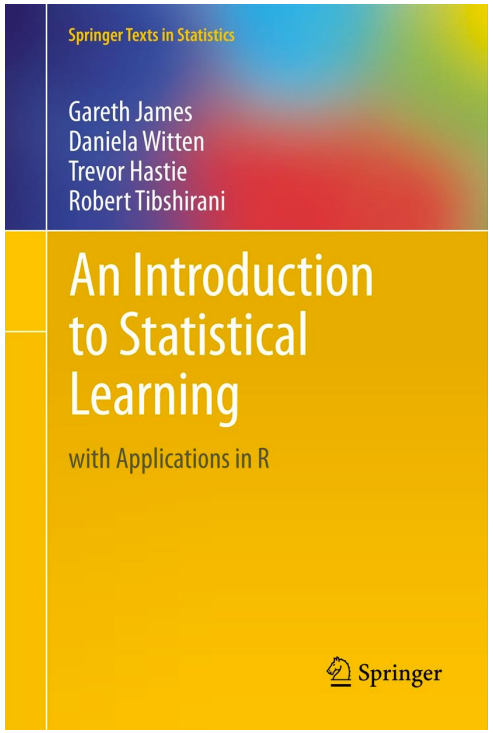
https://shiny.rstudio.com/tutorial/

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

DSLA
DATA SCIENCE LEARNING ACADEMY

Lunch