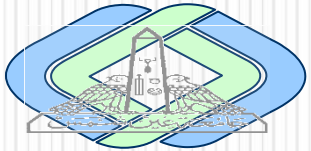
Faculty of Computer
and Information Sciences
Ain Shams University



كلية الحاسبات والمعلومات
جامعة عين شمس

Data Science

Code:

Instructor

Prof. Dr. Abeer M. Mahmoud

Professor of computer Science-faculty of Computer and Information Sciences- Ain Shams University

Data Science and Big Data Analytics v2

Topics : Data Science and Big Data Analytics Course

Introduction to Big Data Analytics + Data Analytics Lifecycle	Review of Basic Data Analytic Methods Using R	Advanced Analytics – Theory and Methods	Advanced Analytics - Technology and Tools	The Endgame, or Putting it All Together + Final Lab on Big Data Analytics
Big Data Overview State of the Practice in Analytics The Data Scientist Big Data Analytics in Industry Verticals Data Analytics Lifecycle	Using R to Look at Data - Introduction to R Analyzing and Exploring the Data Statistics for Model Building and Evaluation	K-means Clustering Association Rules Linear Regression Logistic Regression Naive Bayesian Classifier Decision Trees Time Series Analysis Text Analysis	Analytics for Unstructured Data (MapReduce and Hadoop) The Hadoop Ecosystem In-database Analytics – SQL Essentials Advanced SQL and MADlib for In-database Analytics	Operationalizing an Analytics Project Creating the Final Deliverables Data Visualization Techniques + Final Lab – Application of the Data Analytics Lifecycle to a Big Data Analytics Challenge

Advanced analytics—theory and methods

Lesson: Text analysis

Text analysis

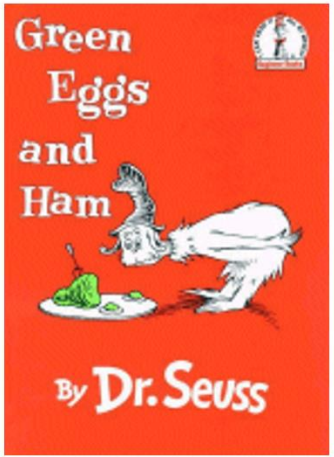
During this lesson, the following topics are covered:

- Challenges with text analysis
- Key tasks in text analysis
- Definition of terms used in text analysis
  - Term frequency, inverse document frequency
- Representation and features of documents and corpus
- Use of regular expressions in parsing text
- Metrics used to measure the quality of search results
  - Relevance with TF-IDF

Text analysis, cont.

Encompasses the processing and representation of text for analysis and learning tasks

- Analyzing tweets
  - Twitter currently generates 500 MM tweets per day
  - It generates tweets in close to 40 languages
  - Distinct number of words can be close to 100,000
- High-dimensionality
  - Every distinct term is a dimension
  - Green Eggs and Ham: A 50-D problem!
- Data is unstructured

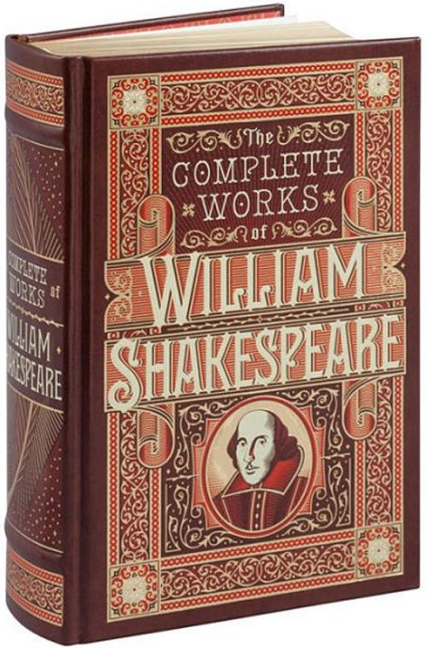


Text analysis—problem-solving tasks

- Parsing
  - Impose a structure on the unstructured or semi-structured text
  - Useful for later analysis
- Search and retrieval
  - Which documents have this word or phrase?
  - Which documents are about this topic or this entity?
- Text mining
  - "Understand" the content
  - Clustering and classification
- Tasks are not an ordered list:
  - Does not represent process
  - Perform tasks based on the problem you want to address

Example—term frequency

- Term frequency—Count of words or terms in each document or across documents
- Calculate the term frequency of the words in the complete works of Shakespeare
  - Total words—0.88 MM, Distinct words—18,000
- Each distinct work is stored as a document. Each document contains words.
- Within the works, there is a possibility that some of documents are dramatis personae—a document with character details.
- Find the short documents that are dramatis personae and filter them from the dataset.



Representing corpus—collection of documents and features

- A corpus is a collection of text documents.
  - Group of news articles
  - A set of emails or tweets
- Corpus metrics
  - Volume
  - Corpus-wide term frequencies
  - Inverse document frequency (IDF)
- Challenge: a corpus is often dynamic
  - Indexes and metrics must be updated continuously

Text classification—parsing and tokenizing

- Tokenization is the task of separating words from the body of text.
- Tokens are typically words, but can be fragments of words
- Not as straightforward as it seems
  - Corpus can help you by removing punctuations, numbers, and stop words so that they are not tokenized.
  - If space is used for tokenizing, “day.” could be one form of word.
  - If punctuation is used for tokenizing, you get "day" and ".". But, It also converts "we’ll" to "we" and "ll".
  - You must find the right token.
- Each language might need a new method of tokenization.
- Case folding is another method used for tokenizing.

Extract and represent text

Document representation:

- A structure for analysis

"Bag of words"

- Common representation
- A vector with one dimension for every unique term in space

Term frequency (tf) =

number of times a term occurs

- Good for basic search, classification

Reduce dimensionality

- Term space—not ALL terms
  - No stop words: "the", "a"
  - Often no pronouns
- Stemming
  - “Phone" = “Phones"

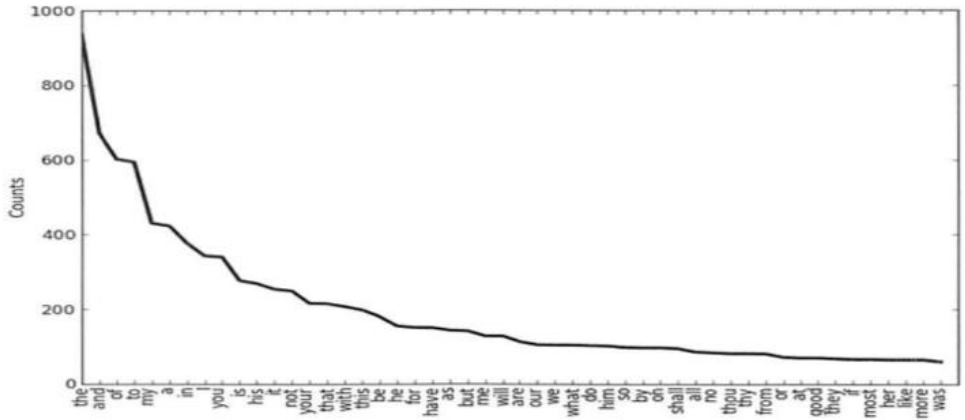
”We know what we are, but not what we may be."

Convert this sentence to a vector in the term space:

We	3
what	2
how	0
are	1
now	0
may	1
doubt	0
thing	0

Extract and represent text, cont.

- Using single words as identifiers with the bag-of-words representation, the term frequency of each word can be calculated.
- As an example, here is a plot for terms with highest counts from Shakespeare's Hamlet.



Computing relevance—term frequency

- Along with the Bag-of-words method, you can also use this function to calculate weights.
- Assign each term in a document a weight for that term.
- The weight of a term t in a document d is a function of the number of times t is displayed in d.
  - The weight can be set to the number of occurrences of t in d:

    $$tf(t, d) = count(t, d)$$

Inverse document frequency (IDF)

idf(t) = log [N/(df(t)+1)]

- N: Number of documents in the corpus
- df(t): Number of documents in the corpus that contain a term t
- Measures term uniqueness in corpus
  - ”love" vs. ”truth“
- Indicates the importance of the term
  - Search—relevance
  - Classification—discriminatory power

NLP Approaches

Rule based Approach	Statistical Approach
Relies on hand-constructed rules that are to be acquired from language specialists	developers do not need language specialists expertise
requires only small amount of training data	requires large amount of annotated training data (very large corpora)
development could be very time consuming , require expert in language	automated
some changes may be hard to accommodate	Less quality - does not explicitly deal with syntax
not easy to obtain high coverage of the linguistic knowledge	
useful for limited domain	

More possibilities with text analysis

- Topic tagging—segment the text documents into the specific categories
  - Segment books into the right categories. There are 300,000 titles published in 2013 in the US only.
  - In the current example, the plays can be categorized into comedies, tragedies, and histories.
- **Sentiment analysis** refers to tasks that use statistics and natural language processing to mine opinions to identify and extract subjective information from texts.
  - This analysis is performed on Twitter to determine overall opinion on a particular trending topic.
  - Companies and brands often utilize sentiment analysis to monitor brand reputation across social media platforms.

What is Natural Language Processing (NLP)

- Getting computers to perform useful tasks involving human languages whether for
  - *Enabling human-machine communication*
  - *Improving human-human communication*
  - *Doing things with spoken or textual material*

- Examples:
  - Spoken Conversational Agents
  - Machine Translation
  - Question Answering

  ……..

Natural language processing

- Natural language processing (NLP) is the capacity of a computer to "understand" natural language text at a level that allows for meaningful interaction between the computer and a person working in a particular application domain.
- Most companies use NLP extensively...
- **Usage of NLP**
  - Social media monitoring
  - Text analytics
  - Formulate responses using natural language
  - Sentiment classification
  - Chatbots

Components of NLP

- **Natural Language Understanding**

- **Natural Language Generation**

20

Tough nut to crack—NLP

- NLP must overcome ambiguity and volume of data for it to understand the sentence.
- Types of ambiguity
  - Lexical ambiguity – words having multiple meanings
  - Syntactic ambiguity – sentence having multiple parse trees
  - Semantic ambiguity – sentence having multiple meanings
  - Anaphoric ambiguity – phrase or word that is previously mentioned but having a different meaning
- A complete guide to NLP is now available in the Advanced Methods in Data Science and Big Data Analytics course

Challenges—text analysis

- Finding the right structure for your unstructured data
- Very high dimensionality
- Thinking about your problem the right way

Why NL Understanding is hard?

- Natural language is extremely rich in form and structure, and **very ambiguous**.
  - How to represent meaning,
  - Which structures map to which meaning structures.

23

Why NL Understanding is hard?

- One input can mean many different things. Ambiguity can be at different levels.
  - Lexical (word level) ambiguity -- different meanings of words
  - Syntactic ambiguity -- different ways to parse the sentence
  - Interpreting partial information -- how to interpret pronouns
  - Contextual information -- context of the sentence may affect the meaning of that sentence.
    - Many input can mean the same thing.
    - Interaction among components of the input is not clear.

24

Phonology

25

Levels of Analysis for language Processing

- **Phonology** – concerns how words are related to the sounds that realize them.

26

Morphology

27

Levels of Analysis for language Processing

- **Morphology** – concerns how words are constructed from more basic meaning units called morphemes. A morpheme is the primitive unit of meaning in a language.

Levels of Analysis for language Processing

- We can usefully divide morphemes into two classes
  - Stems: The core meaning bearing units
  - Affixes: Bits and pieces that adhere to stems to change their meanings and grammatical functions

Regulars… follow the rules	**Irregulars**… don't follow the rules
Walk, walks, walking, walked, walked Table, tables kick, kicks, kicked, kicking	Eat, eats, eating, ate, eaten Catch, catches, catching, caught, caught Cut, cuts, cutting, cut, cut Goose, geese

29

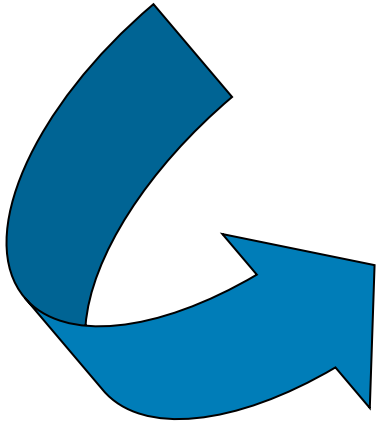
Morphological Parsing

- These regularities enable us to create software to parse words into their component parts

Example : morphological analysis

Surface form

I want to print Ali's .init file



stems

I (pronoun)
want (verb)
to (prep)
to(infinitive)
print (verb)
Ali (noun)
's (possessive)
.init (adj)
file (noun)
file (verb)

Syntax

32

Levels of Analysis for language Processing

- **Syntax** – concerns how can be put together to form correct sentences and determines what structural role each word plays in the sentence and what phrases are subparts of other phrases.
- Syntax is the study of regularities and laws of word order and phrase structure

33

Syntax

- Syntax: Structural relationship between words.
- The main issues here are structural ambiguities, as in:

  - Word Order
    - John hit Bill
    - Bill was hit by John
    - Bill hit John
    - Bill, John hit
    - Who John hit was Bill

34

Syntax

- In English, we cannot determine the meaning of the sentence from the meaning of the words.
  - Mary gave Peter a book. Peter gave Mary a book.
- The basic word order in English is: Subject-Verb-Object
- This holds for declarative sentences,
  - The children should eat spinach
- but the order changes to express a particular "mood":
  - Interrogative (question): Should the children eat spinach?
  - Imperative (command, request): Eat spinach!

35

Parsing (Syntactic Analysis)

- Assigning a syntactic and logical form to an input sentence
  - uses knowledge about word and word meanings (lexicon)
  - uses a set of rules defining legal structures (grammar)

  - Ex”
- Ahmad ate the apple.

```
(S  (NP (NAME Ahmad))
    (VP  (V ate)
         (NP  (ART the)
              (N apple))))
```


36

Rewrite Rules

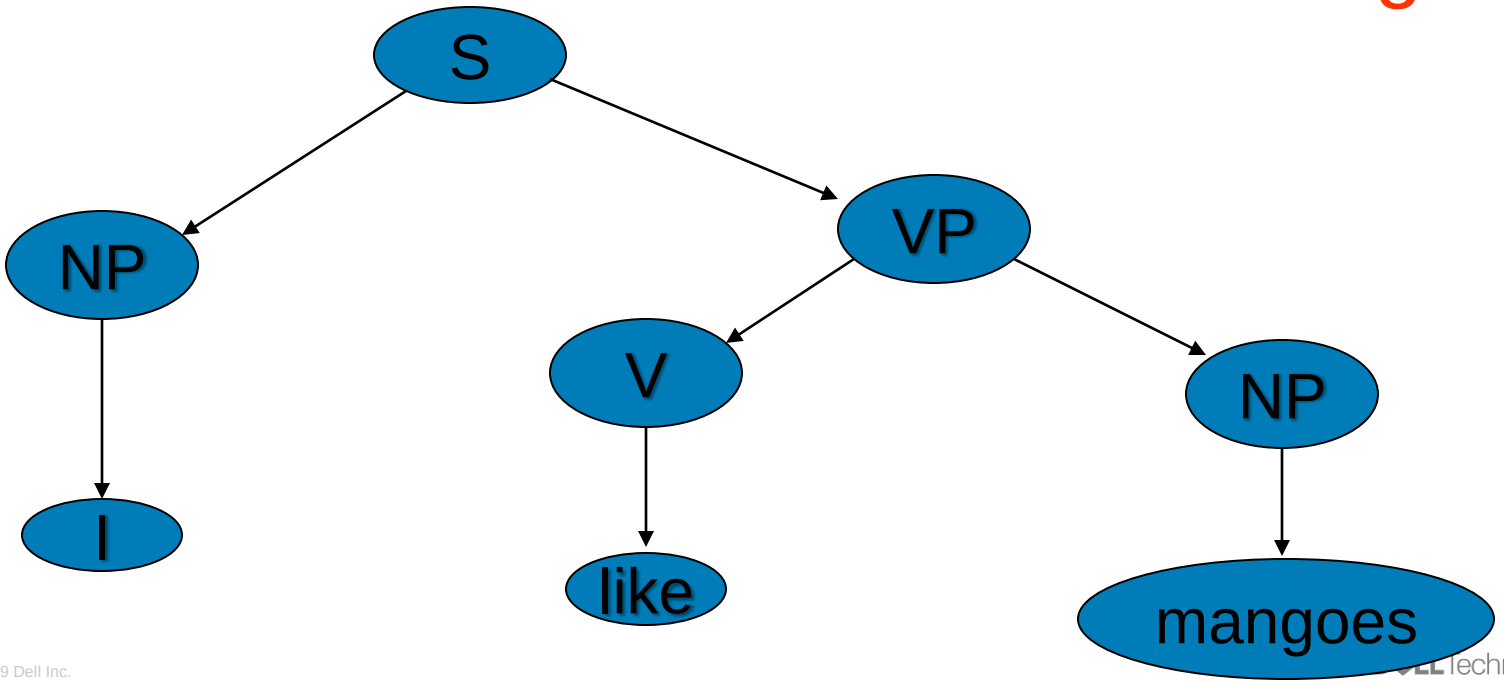
- Symbols that cannot be decomposed are called terminal symbols.
- Symbols that can be decomposed are called nonterminals.
- An intuitive way to represent a sentence structure is as a tree, in which each nonterminal represents the application of the rewrite tree. T
- The following example present a tree representation of the sentence
  - I like Mango

37

Syntax Processing Stage

Structure Detection

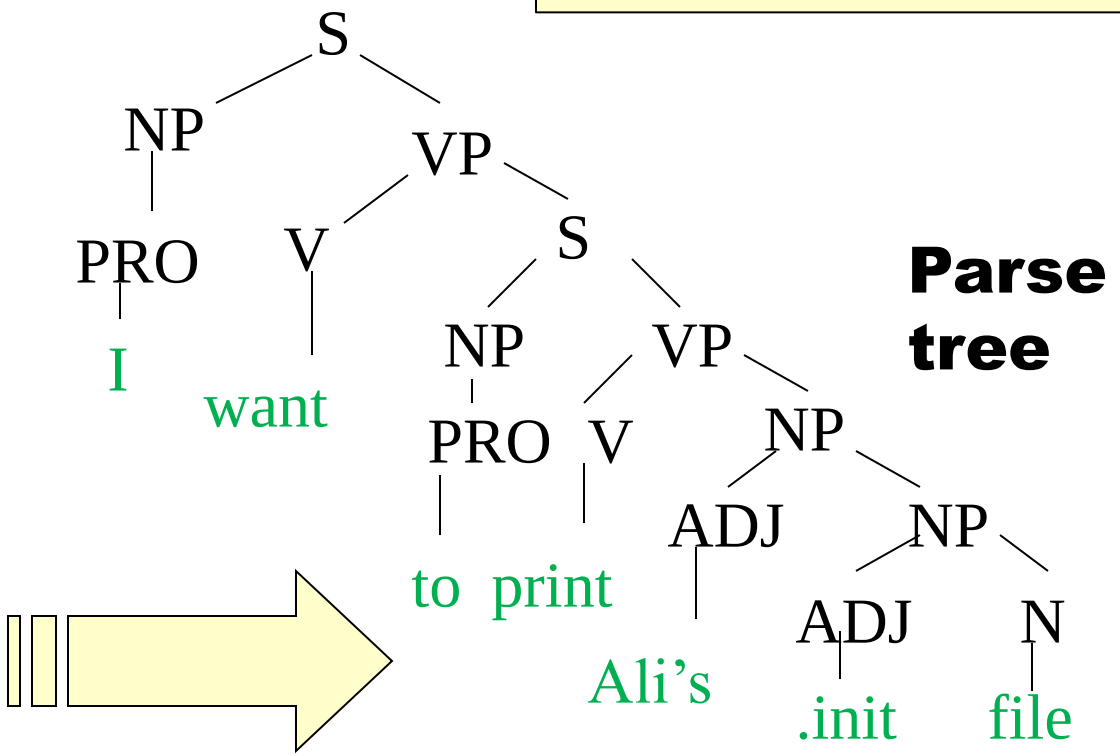
I like Mango



I want to print Ali's .init file

stems

I (pronoun)
want (verb)
to (prep)
to(infinitive)
print (verb)
Ali (noun)
's (possessive)
.init (adj)
file (noun)



Parse tree

Semantics

40

Levels of Analysis for language Processing

- **Semantics** – concerns what words mean and how these meaning combine in sentences to form sentence meaning. The study of context-independent meaning.

41

Semantics

- Semantics: the study of the meaning of language. Can be decomposed into:

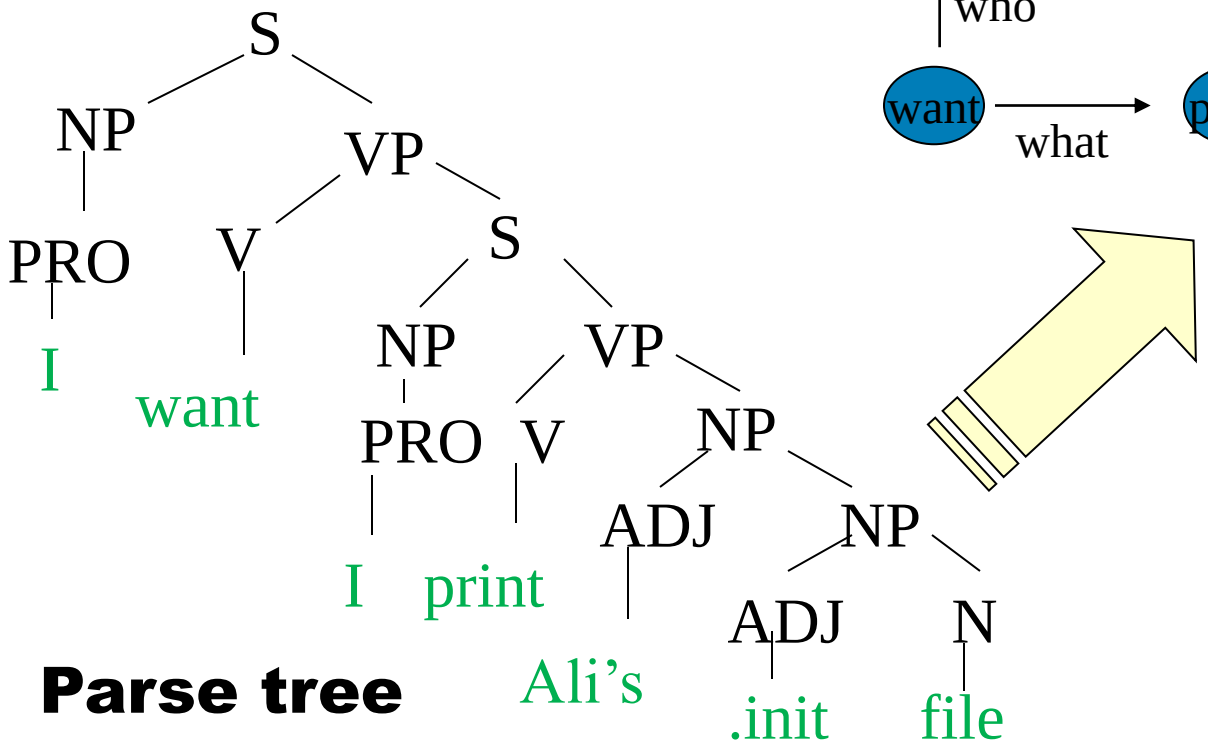
Lexical semantics: the study of meaning of **individual** words

Global semantics: how the meaning of individual words are **combined** into meaning of sentences (or more).

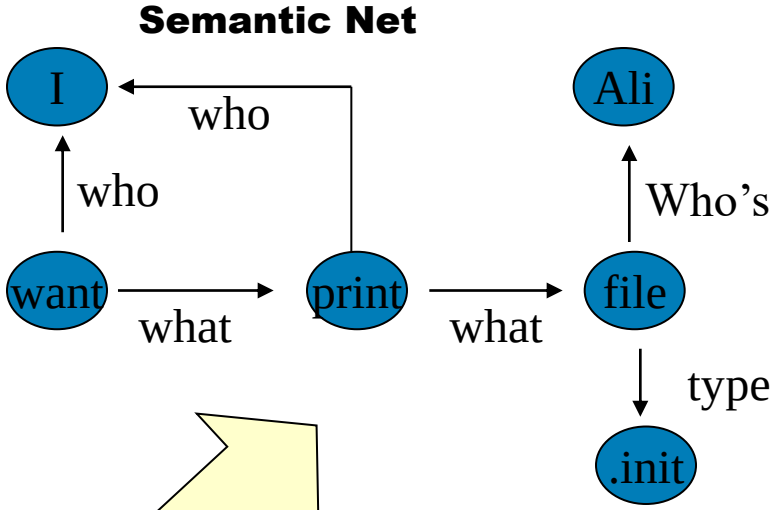
One approach to lexical semantics is to study how word meanings are related to each other. To study this, words can be organized into lexical hierarchies

42

Example :Semantic analysis



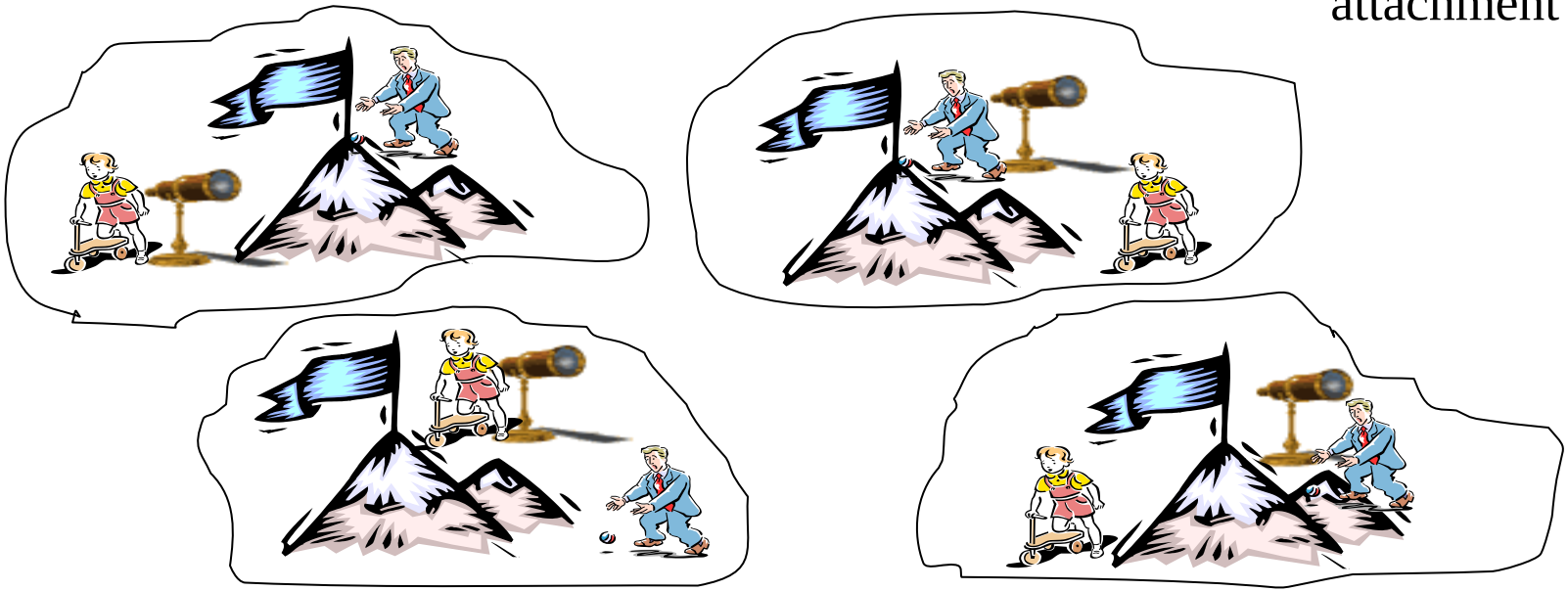
Parse tree



Example :Ambiguity

The boy saw the man on the mountain with a telescope

Prepositional phrase attachment



Pragmatics/Discourse

45

Levels of Analysis for language Processing(cont.)

- **Pragmatics** :–

  concerns how sentences are used in different situations and how use affects the interpretation of the sentence.

- **Discourse** :–

  concerns how the immediately preceding sentences affect the interpretation of the next sentence. For example, interpreting pronouns and interpreting the temporal aspects of the information.

- **World Knowledge** :–

  includes general knowledge about the world. What each language user must know about the other's beliefs and goals.

46

Pragmatics and Discourse

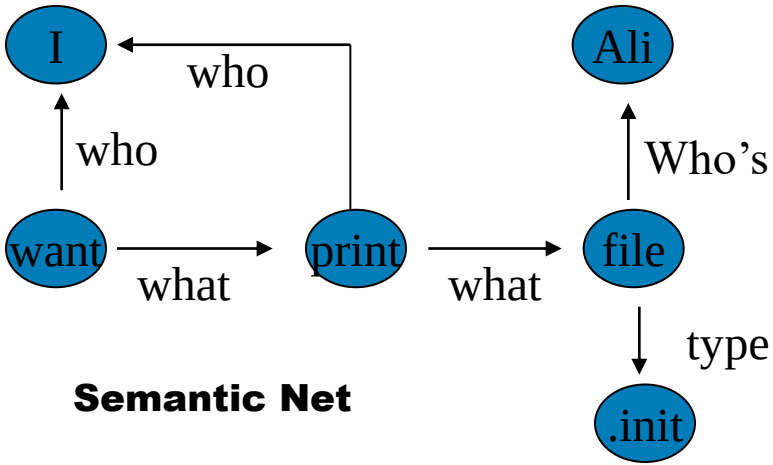
- Very hard problem
- Pragmatics: How a sentence is used; its purpose.
- E.g.: Rules of conversation:
  - Can you tell me what time it is
  - Could I have the salt
- Model user intention
  - *Tourist (in a hurry, checking out of the hotel, motioning to the service boy): Boy, go upstairs and see if my sandals are under the divan. Do not be late. I just have 15 minutes to catch the train.*
  - *Boy (running upstairs and coming back panting): yes sir, they are there.*

Example :Discourse

To whom the pronoun 'I' refers

To whom the proper noun 'Ali' refers

What are the files to be printed



Semantic Net

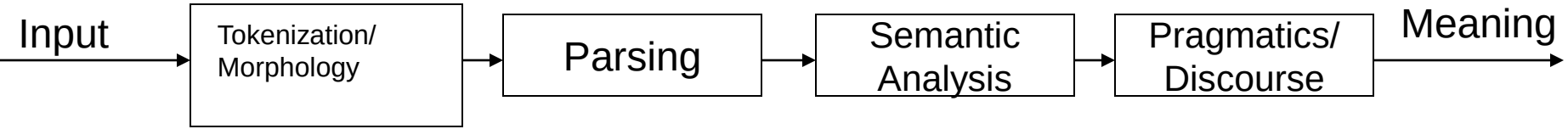
Example : Pragmatics

Execute the command

```
lpr /ali/stuff.init
```


48

Natural Language Understanding



- Key issues:
  - Knowledge
    - How acquire this knowledge of language?
      - Hand-coded? Automatically acquired?
  - Ambiguity
    - How determine appropriate interpretation?

Your Turn

Ambiguity

I made her duck.

- Mention three different interpretations for this sentence ?
- What are the reasons for the ambiguity?

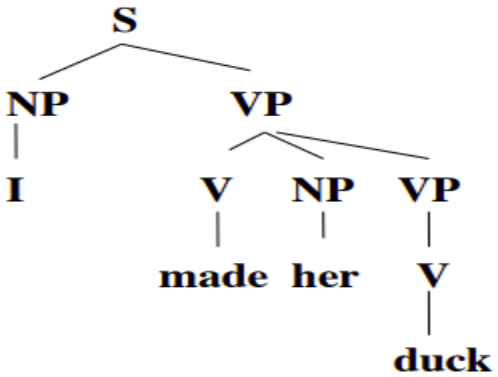
- I cooked waterfowl for her benefit (to eat)
- I cooked waterfowl belonging to her
- I created the duck she owns
- I caused her to quickly lower her head or body
- I waved my magic wand and turned her into undifferentiated waterfowl

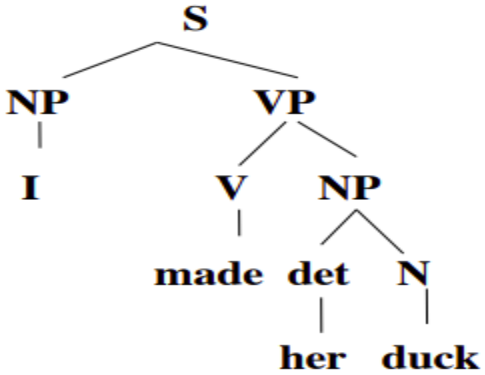
Ambiguity is Everywhere

- Lexical category: part of speech
  - **Duck** can be a Noun or Verb
    - V: Duck! I caused her to quickly lower her head or body.
    - N: I cooked waterfowl for her benefit
  - **Her** can be possessive (of her) or dative (for her)
    - Possessive: I cooked waterfowl belonging to her.
    - Dative: I cooked waterfowl for her benefit
- Lexical Semantics:
- ▸ Make can mean create or cook
- • create: I made the (plaster) duck statue she owns
- • cook: I cooked waterfowl for her benefit

Syntactic Ambiguity

- Structural ambiguity: one sentence can have many syntactic representations





I caused her to quickly lower her head or body

I created her waterfowl

Ambiguity is Everywhere

- Phonetics! –

I mate or duck
I'm eight or duck

- Eye maid; her duck
- Aye mate, her duck
- I maid her duck
- I'm aid her duck
- I mate her duck
- I'm ate her duck
- I'm ate or duck
- I mate or duck

Context to the rescue

- Q1: What did you cook for Mary last night?
- A1: I made her duck.

- Q2: Where did Mary get that great plaster duck?
- A2: I made her duck.

Check your knowledge

1. What are the two major challenges in the problem of text analysis?
2. Why is the corpus metrics dynamic. Provide an example and a scenario that explains the dynamism of the corpus metrics.
3. How does tf-idf enhance the relevance of a search result?
4. List and discuss a few methods that are deployed in text analysis to reduce the dimensions.

Text analysis—summary

During this lesson, the following topics were covered:

- Challenges with text analysis
- Key tasks in text analysis
- Definition of terms used in text analysis
  - Term frequency, inverse document frequency
- Representation and features of documents and corpus
- Metrics used to measure the quality of search results
  - Relevance with TF-IDF