By
Dr. Hussein Hazimeh

Lebanese University
Faculty of Information 1

Data Science Departement

2nd year – Data Analysis in R

March – 2022 – Chapter 7

Agenda

- Text analysis topics
  - Text pre-processing
    - Tf-idf
    - Text operation
      - Stemming
      - Stop word removal
      - Special character removal
  - Sentiment analysis
    - **Rule-based (lexicon-based)**
    - Supervised-based
  - Named entity recognition
    - **CRF**
    - Supervised-based
  - Topic analysis (detection)
    - **Rule-based**
    - Supervised-based

- What is information extraction
- Data extraction vs information extraction
- Main domains of information extraction
- Named Entity Recognition
- Topic detection

What is infromation extraction

Information Extraction (IE)

- Late 1970s within NLP field
- Find and extract automatically limited relevant parts of texts
- Move from unstructured/semi-structured data to structured data
  - Schemas
  - Relations (as a database)

What is IE

» Unstructured text

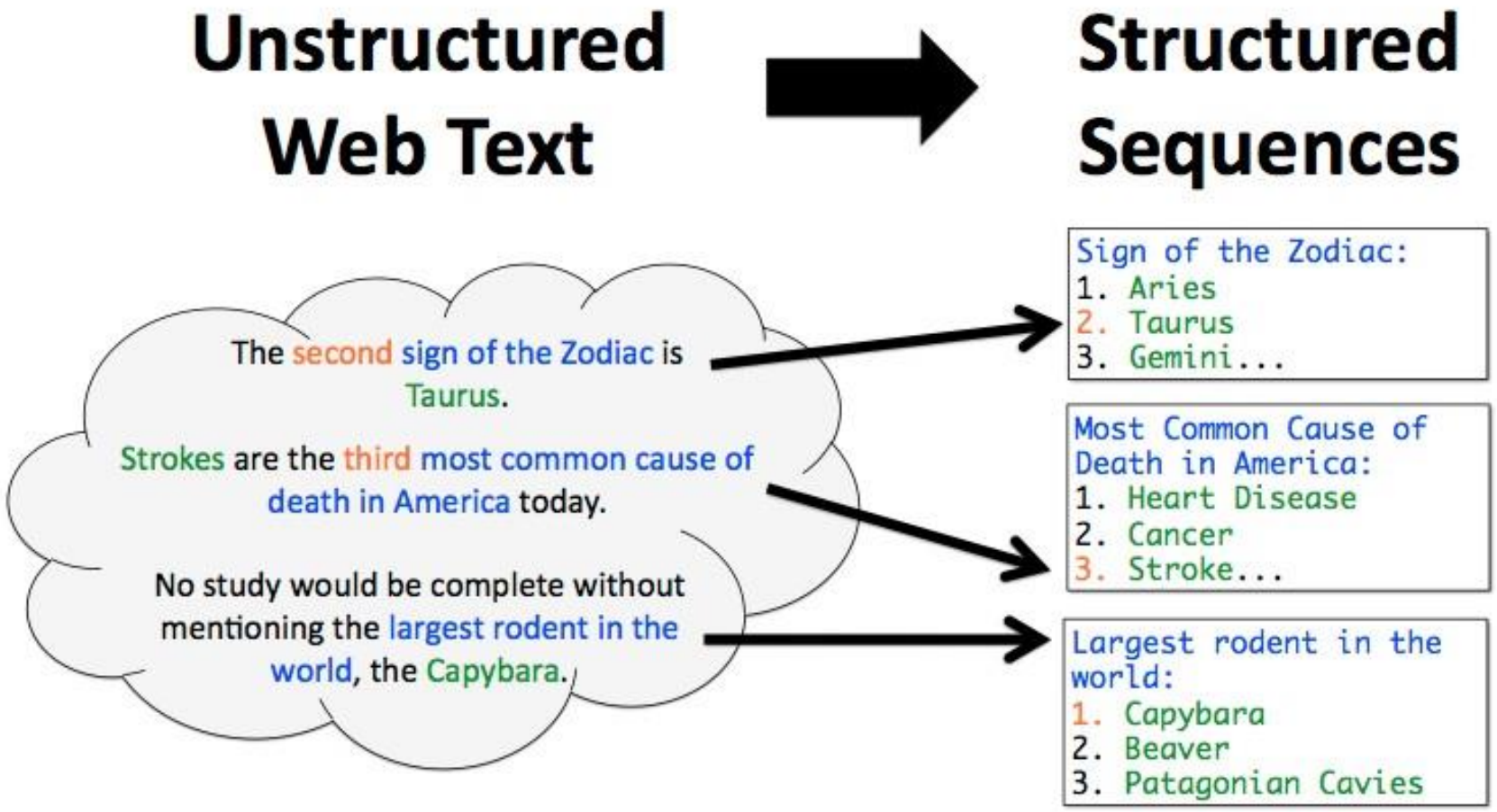
> Professor John Skvoretz, U. of South Carolina, Columbia, will present a seminar entitled "Embedded Commitment," on Thursday, May 4th from 4-5:30 in PH 223D.

What is IE

» Semi-structured text

Name: Dr. Jeffrey D. Hermes
Affiliation: Department of AutoImmune Diseases
Research & Biophysical Chemistry Merck Research Laboratories
Title: "MHC Class II: A Target for Specific Immunomodulation of the Immune Response"
Host/e-mail: Robert Murphy, murph@a.crf.cmu.edu
Date: Wednesday, May 3, 1995
Time: 3:30 p.m.
Place: Mellon Institute Conference Room
Sponsor: MERCK RESEARCH LABORATORIES

What is IE


Unstructured
Web Text
Structured
Sequences
The second sign of the Zodiac is Taurus.
Strokes are the third most common cause of death in America today.
No study would be complete without mentioning the largest rodent in the world, the Capybara.
Sign of the Zodiac:
1. Aries
2. Taurus
3. Gemini...
Most Common Cause of Death in America:
1. Heart Disease
2. Cancer
3. Stroke...
Largest rodent in the world:
1. Capybara
2. Beaver
3. Patagonian Cavies

Main goals IE

Goals of IE

- Fill a predefined “template” from raw text
- Extract *who did what to whom and when?*
  - *Event extraction*
- Organize information so that is useful to people

IE tasks and sub tasks

- **Named Entity Recognition (NER)**
  - Detection ↠ Mr. Smith eats bitterballen [Mr. Smith] : ENTITY
  - Classification ↠ Mr. Smith eats bitterballen [Mr. Smith] : PERSON

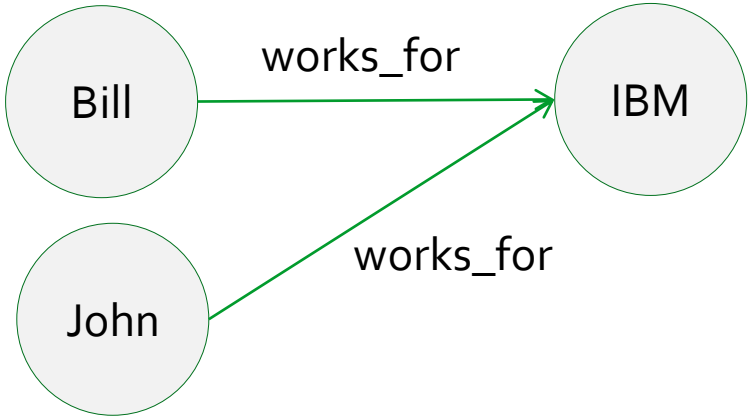
- Event extraction
  - The thief broke the door with a hammer
    - CAUSE_HARM ↠

Verb:	break
Agent:	the thief
Patient:	the door
Instrument:	a hammer

- Coreference resolution
  - [Mr. Smith] eats bitterballen. Besides to this, [he] only drinks Belgium beer.

IE tasks and sub tasks

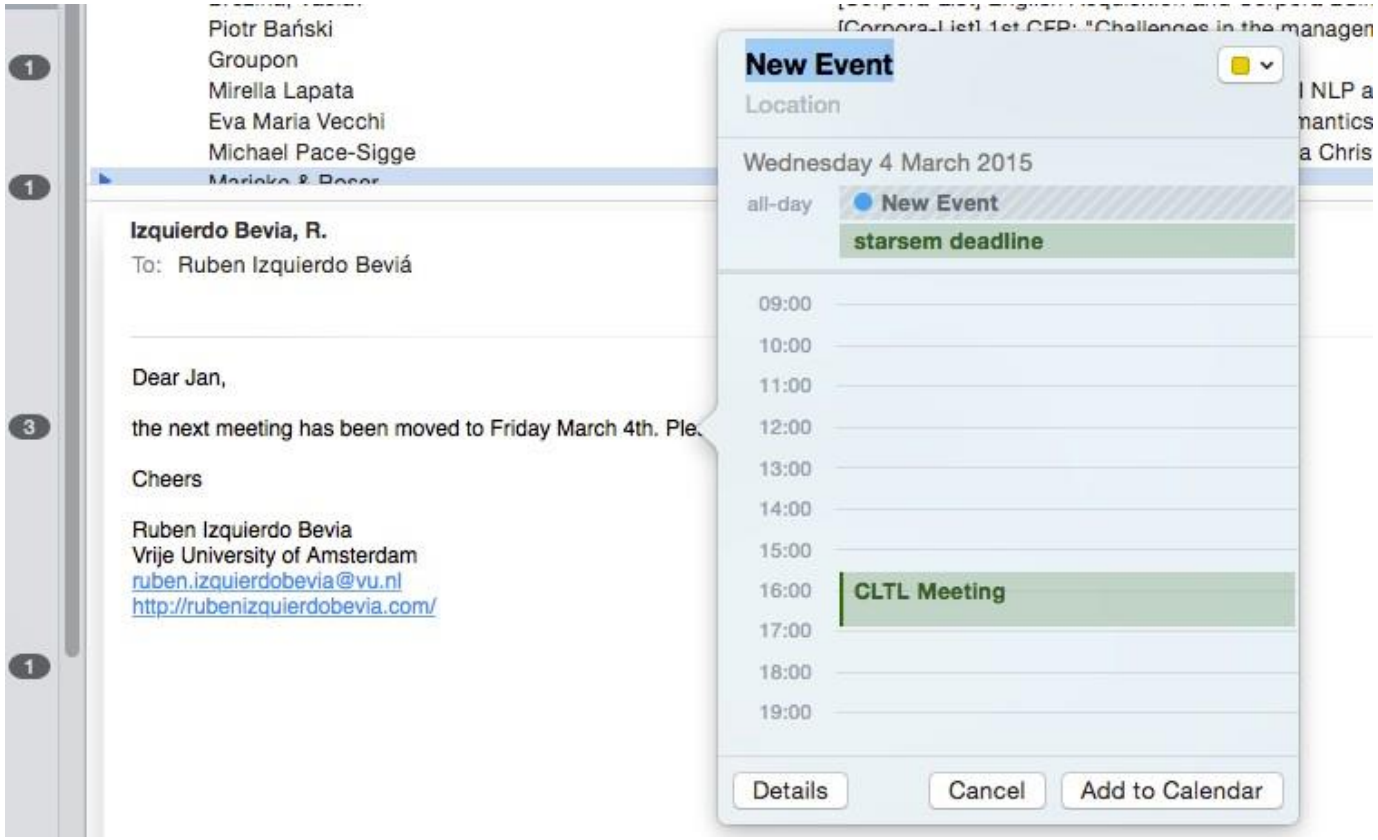
- Relationship extraction
  - Bill works for IBM PERSON works for ORGANISATION



- Terminology extraction (tf-idf)
  - Finding relevant terms of multi words from a given corpus
- Some concrete examples
  - Extracting earnings, profits, board members, headquarters from company reports
  - Searching on the WWW for e-mails for advertising (spamming)
  - Learn drug-gene product interactions from biomedical

IE tasks and sub tasks

- Apple mail



Methods for IE

- Rule based methods
  - Rule based
  - Regular expressions
- Learning based approaches
  - Traditional classifiers
    - Bayes, MME, SVM …
  - Sequence label models
    - HMM, CMM, CRF
- Unsupervised approaches
- Hybrid approaches

Regular expressions

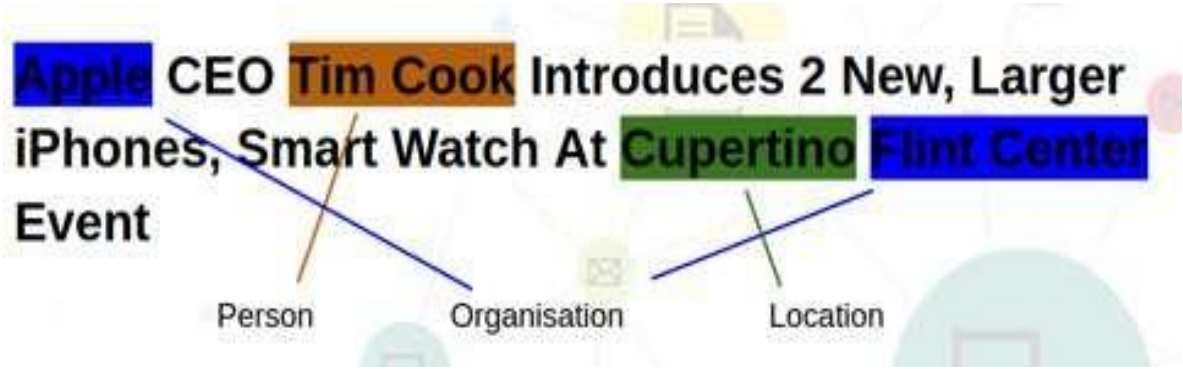
Character	Description
a	The character a
.	Any single character
[abc]	Any character in the brackets (OR) ‘a’ or ‘b’ or ‘c’
[^abc]	Any character not in the brackets. Any symbol that is not ‘a ‘ or ‘b’ or ‘c’
*	Quantifier. Matches the preceding element ZERO or more times
+	Quantifier. Matches the preceding element ONE or more times
?	Matches the previous element zero or one time
|	Choice (OR) Matches one of the expressions (before of after the |)

Named Entity Recognition

What is NER

- Sub-domain under NLP (Natural Language Processing)
- A part of IE (Information Extraction)
- Automatic identification and counting of occurrences of named entities in a collection of information.
- Associating the named entities to their appropriate types

What basically a named entity



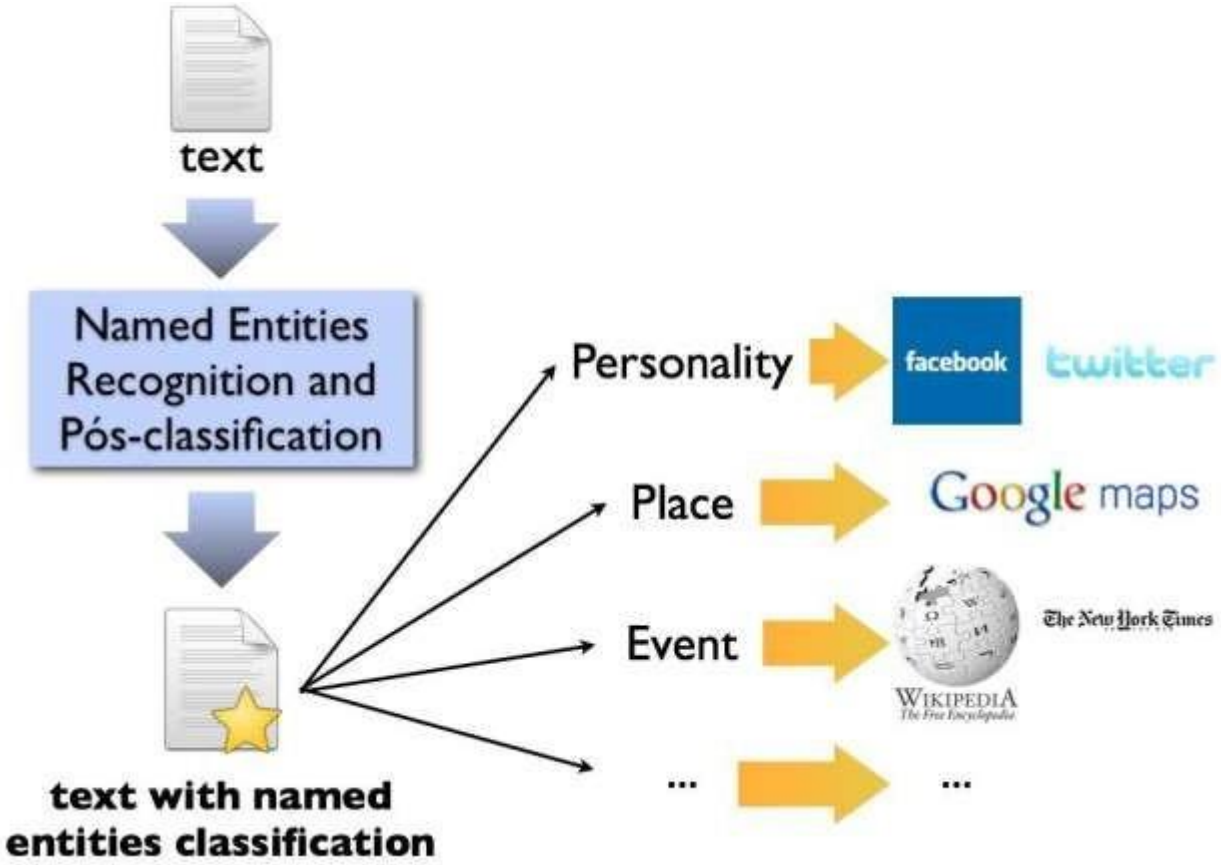
- Word or Phrase that identifies one item from a set of items that have similar attributes
- Semantic elements that carry a meaning

Named Entities with their labels are recognized as follows:

- **ENAMEX** : Person(Tim Cook) , Organization (Apple , Flint Center), Location(Cupertino)
- **TIMEX** : Date , Time
- **NUMEX** : Money , Percentage , Quantity

- Named Entities are either dependent on the Proper Names tagging or on the Part Of Speech (POS) tagging.

Types of named entities



➢ GENERIC NE:

Includes names of persons , organizations, etc.

For Example, any general requirement consisting of names of persons, organization , URLs, Location and so on.

➢DOMAIN SPECIFIC NE:

Consists of entities related to domains

For example,

In a medical domain, names of diseases , names of medicines form the entities whereas

In a manufacturing domain names of products , manufacturers , attributes of products form the named entities.

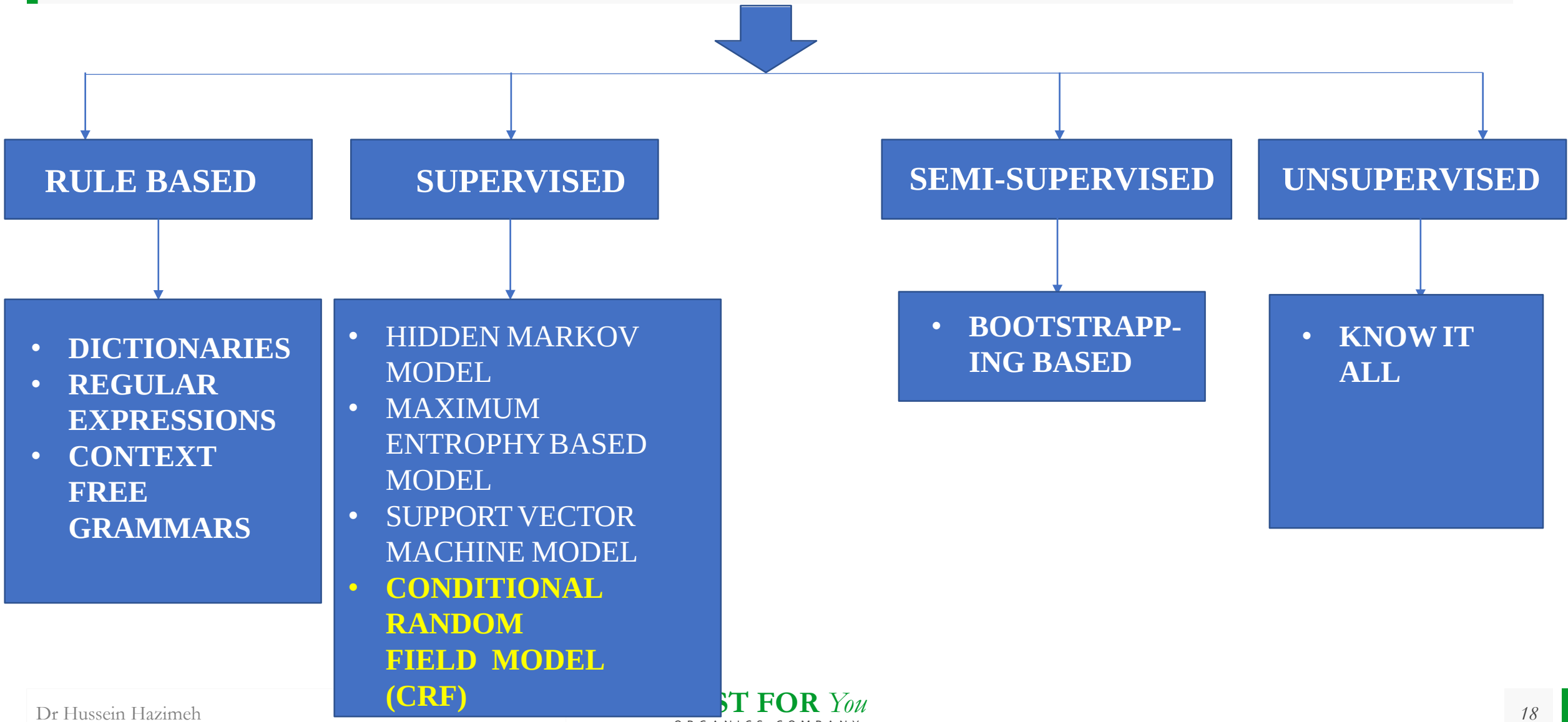
NER input and output

{"document":"Jim went to Stanford University, Tom went to the University of Washington. They both work for Microsoft."}

OUTPUT

[[["Jim", "PERSON"], ["Stanford", "ORGANIZATION"], ["University", "ORGANIZATION"], ["Tom", "PERSON"], ["University", "ORGANIZATION"], ["of", "ORGANIZATION"], ["Washington", "ORGANIZATION"]], [["Microsoft", "ORGANIZATION"]]]

Technique for NER



Topic Classification

Topic classification

- In machine learning and natural language processing, a topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents.
- Topic modeling is a frequently used text-mining tool for discovery of hidden semantic structures in a text body.

Topic classification

- By now, you're probably a bit more clued-up about machine learning; it goes something like this – humans lead by example and algorithms follow suit, right? Now, let's throw you off a little (don't worry, it will all become clear again)… It's also possible to build a topic classifier without machine learning!

Rule-based Topic classification

- Proposed solution:

LEXICONS