Natural Language Processing

Natural Language Processing

Information Extraction

Instructor: Moushmi Dasgupta

Connect up with me on LinkedIn
www.linkedin.com/in/moushmi1234

Some of the material is from Georgia Institute of Technology, Atlanta, GA, USA.

Session Plans

Date	Day	Topic	Time
2nd Nov 2022	Wednesday	Text Classification	2:00PM – 4:00PM
4th Nov 2022	Friday	Text Classification	2:00PM – 4:00PM
5th Nov 2022	Saturday	IE	9:30AM - 1:30PM
9th Nov 2022	Wednesday	IE	2:00PM – 4:00PM
11th Nov 2022	Friday	Chatbot	2:00PM – 4:00PM
13th Nov 2022	Saturday	Chatbot	9:30AM - 1:30PM

AGENDA

Information Extraction

Module 3

4.1 IE Applications

4.2 IE Tasks

4.3 The General Pipeline for IE

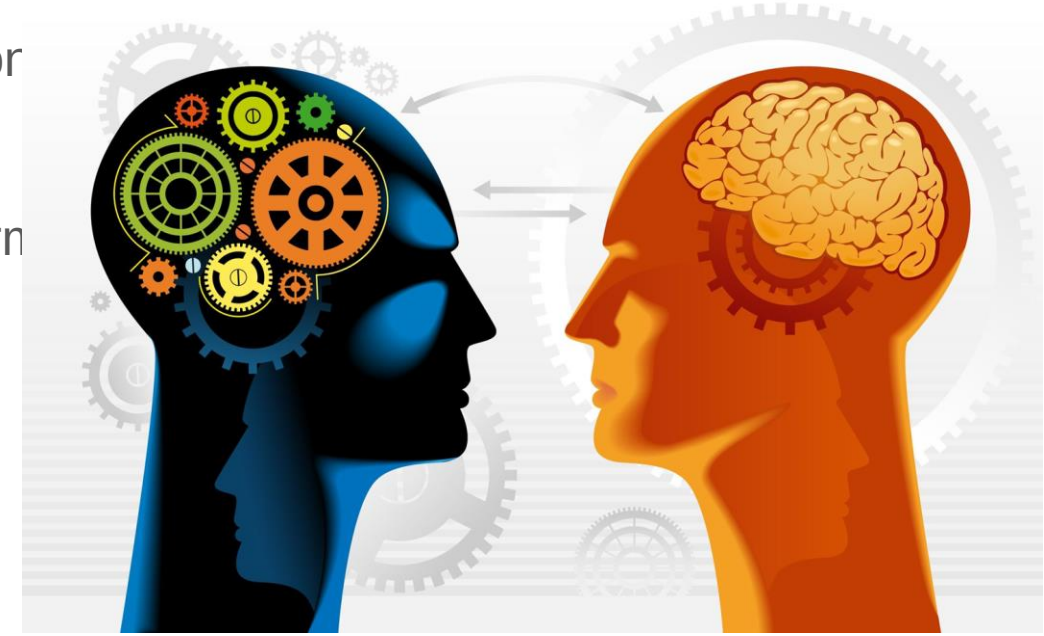
4.4 Case Study on IE

Pre-requisite:

1. Python programming
2. An understanding of Machine Learning
3. Fundamentals of NLP Understanding and hands-on with Python programming on and IDE
4. An Understanding of NLP Pipeline
5. Understanding of Text Classifications
6. Invest in attending classroom sessions (Weekly 1 or 2 classes of 3+ hours duration)
7. Invest in yourself with1 hour of self study everyday

Natural Language Processing

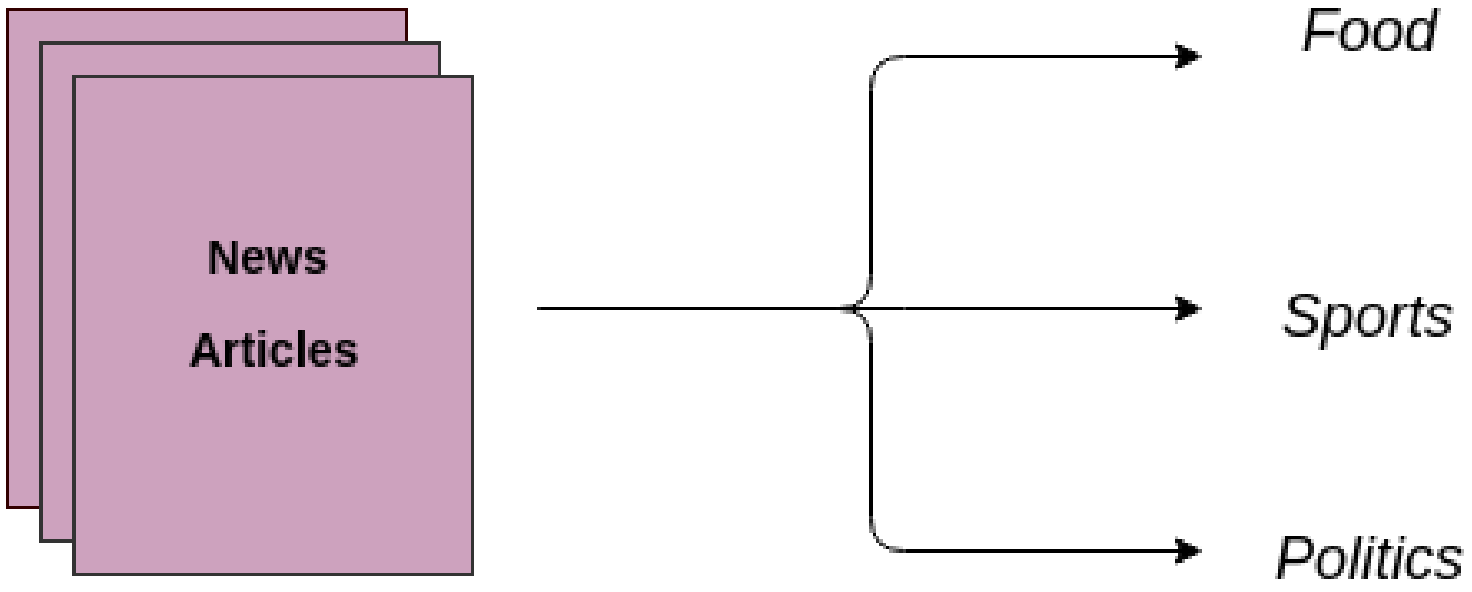
1. Natural Language Processing is a subfield of artificial intelligence concerned with methods of communication between computers and natural languages such as english, hindi, etc.
2. Objective of Natural Language processing is to perform useful tasks involving human languages like
    - Sentiment Analysis
    - Machine Translation
    - Part of Speech Tags
    - Human-Machine communication(chatbots)



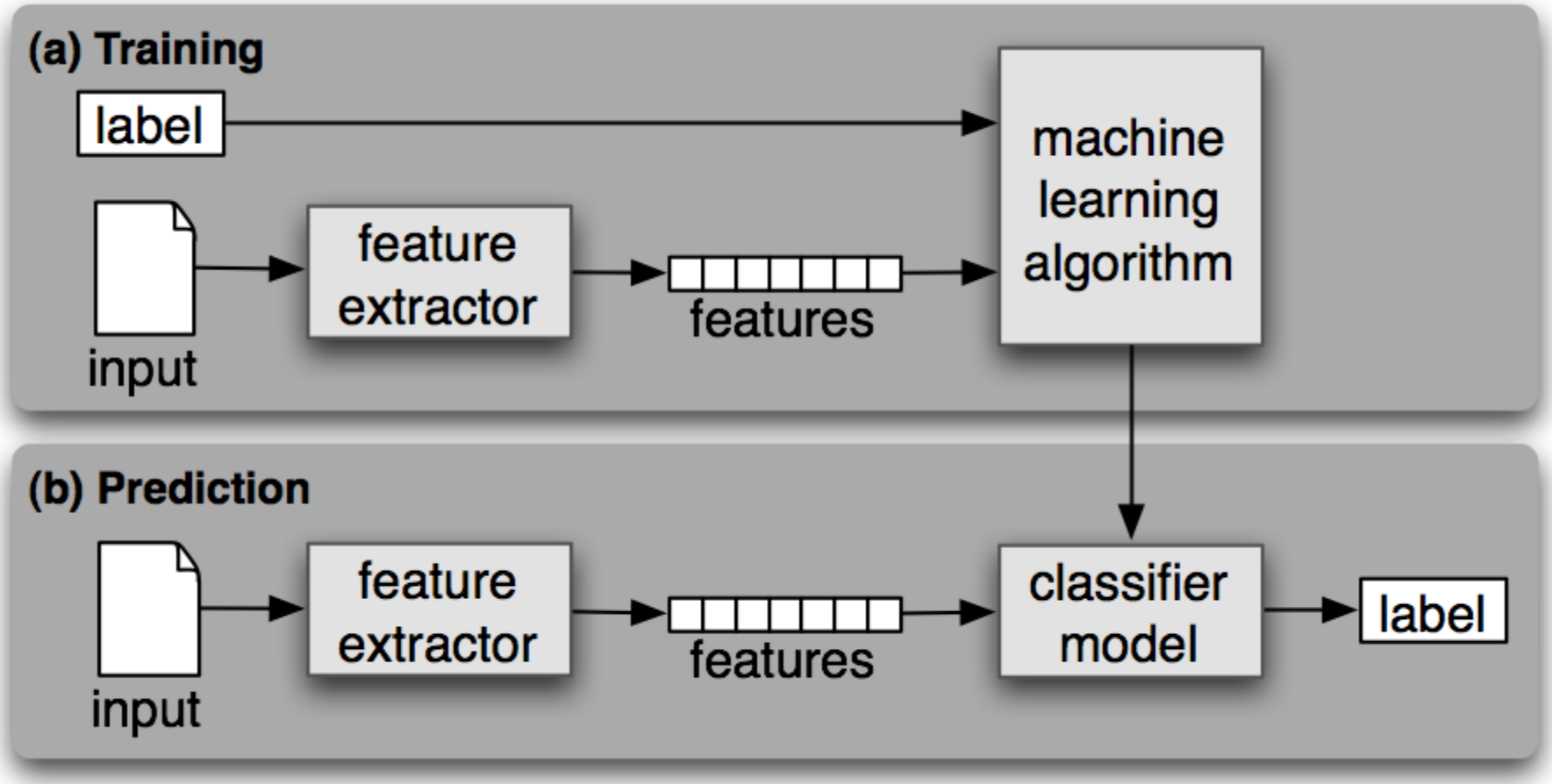
NLP Pipeline Steps

Some of the material is from Georgia Institute of Technology, Atlanta, GA, USA.

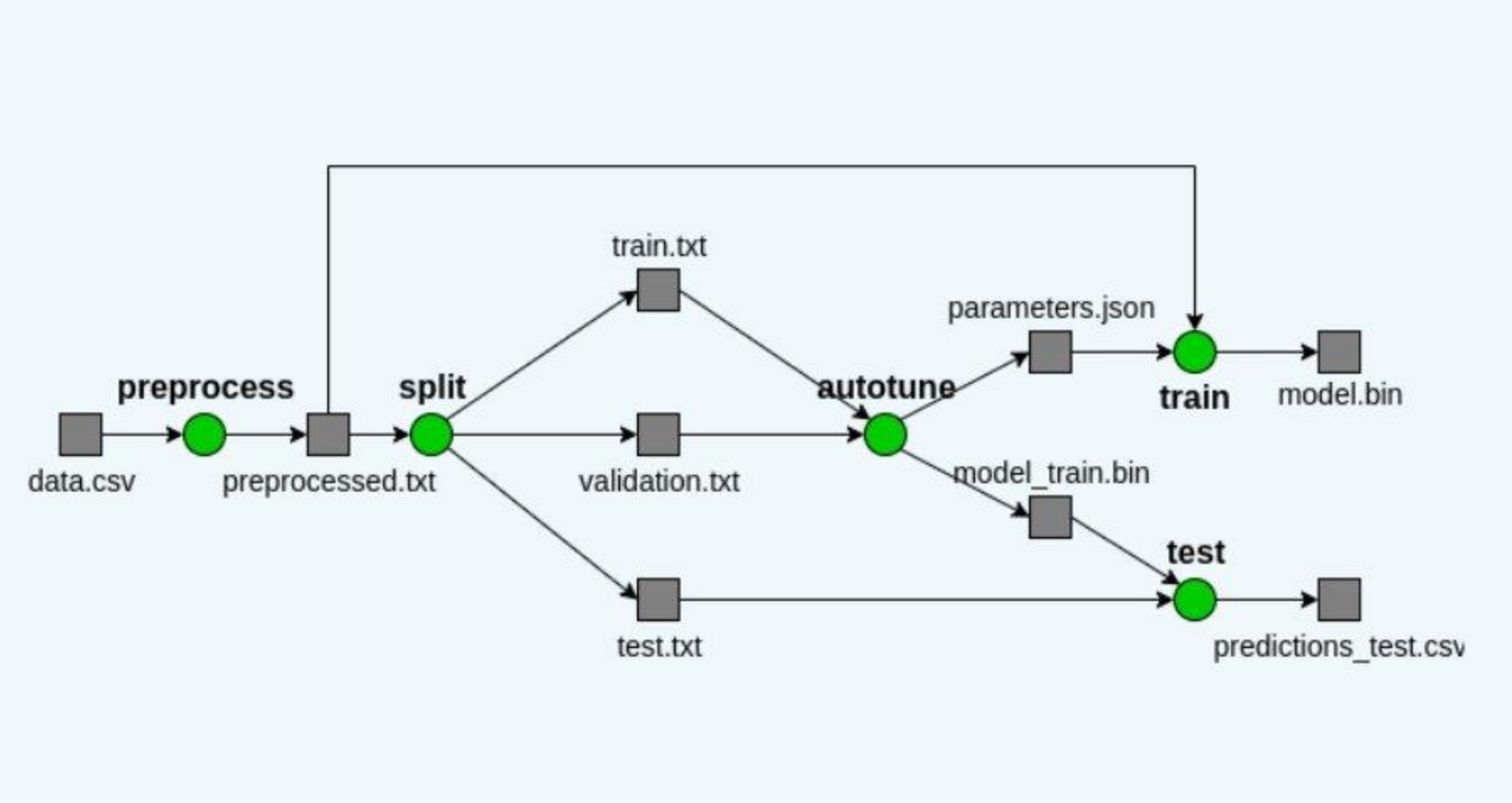
Text Classification



A classifier is called **supervised** if it is built based on training corpora containing the correct label for each input. The framework used by supervised classification is shown in



3.1 A Pipeline for Building Text Classification Systems



A production level pipeline for building Txt classification systems

3.2 Using Existing Text Classification APIs

Open-Source Libraries for Text Classification

1.Scikit-learn
2.Natural Language Toolkit (NLTK)
3. SpaCy

(Other Open source Libraries of interest)
1.TensorFlow
2.PyTorch
3.Keras

Paid API's

Google Cloud NLP
IBM Watson
Lexaytics
Amazon Comprehend
Aylien

Scikit Learn – Text Analytics

https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html

Practice labs as demoed in class

NLTK – Natural Language Toolkit

https://www.nltk.org/

Perform all labs as demoed in the Class

SpaCy

https://spacy.io/

Perform all labs as demoed in the Class

Named Entity Recognition:

A named entity is a "real-world object" that's assigned a name.
Example, a person, a country, a product or a book title.

We also get named entity recognition as part of spacy package.

It is inbuilt in the english language model and we can also train our own entities if needed.

Dependency Parser

A dependency parser analyzes the grammatical structure of a sentence.

It establishing relationships between "head" words and words which modify those heads.

Spacy can be used to create these dependency parsers which can be used in a variety of tasks.

Word Similarity:

Spacy has word vector model as well.
We can use the same to find similar words.

3.3 Use Case for Sentiment Analysis on Amazon Customer Reviews Data

Practical Case Study under Discussion

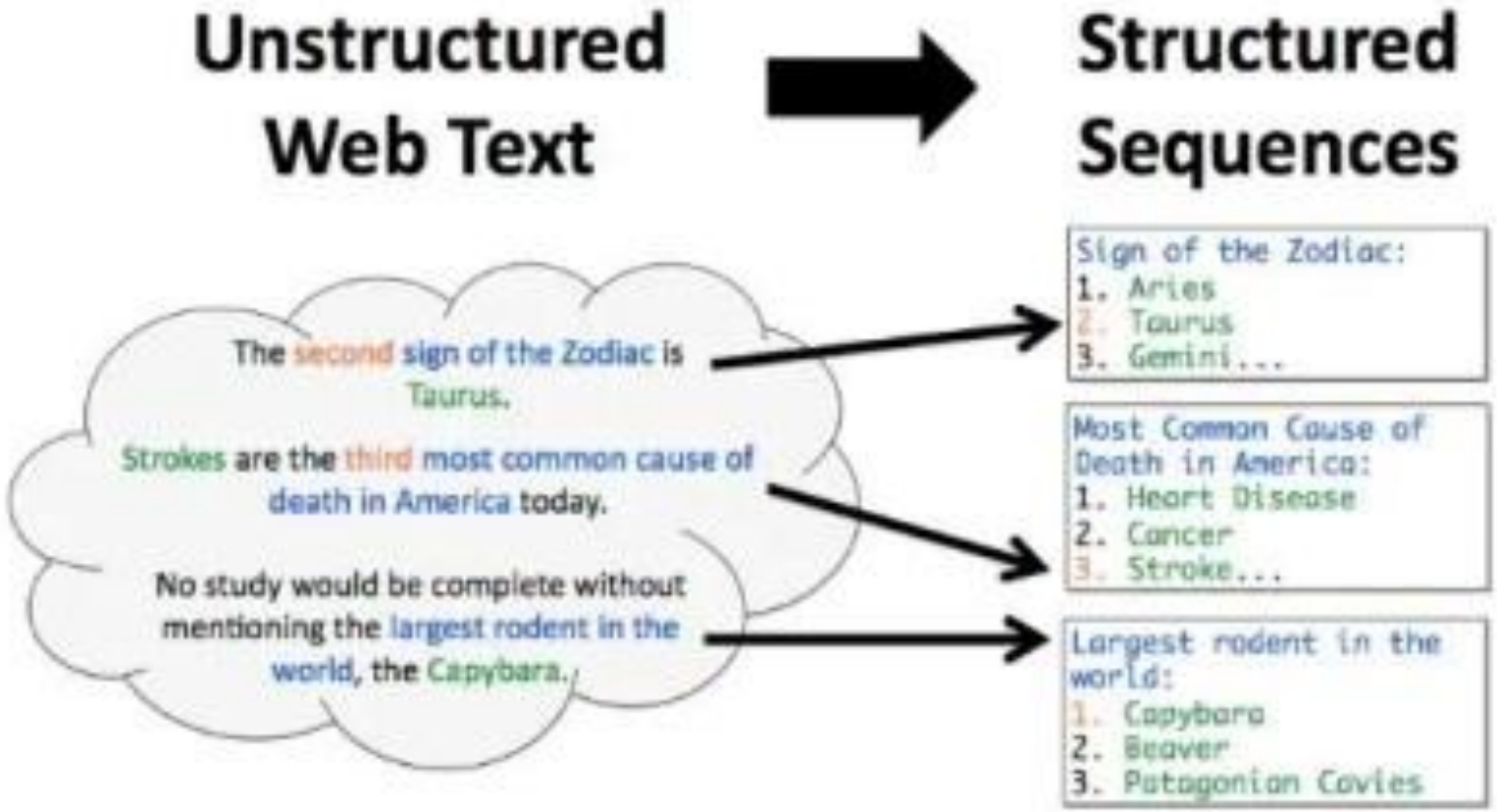
Check and execute Scripts using NLP Tools and API

Information Extraction

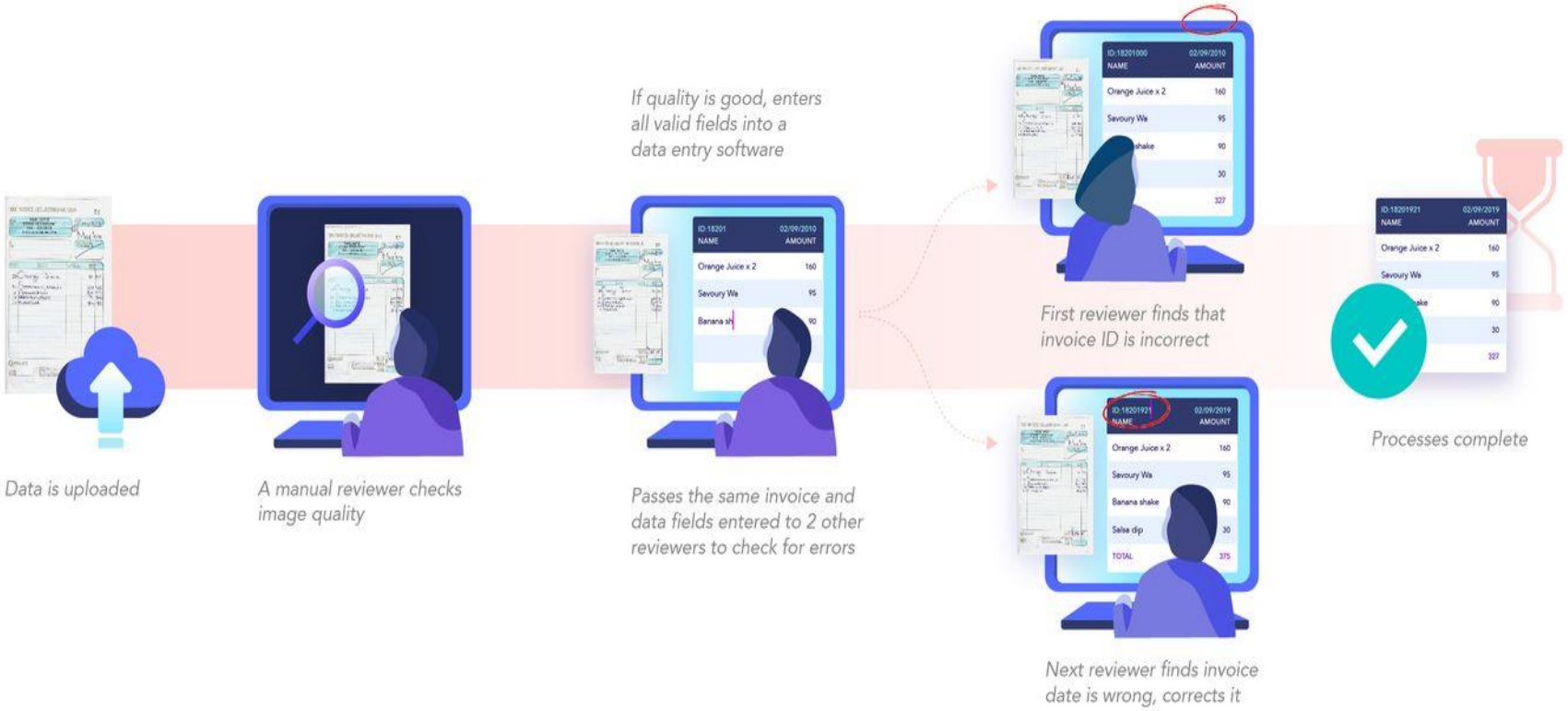
Information Extraction is the process of parsing through unstructured data and extracting essential information into more editable and structured data formats.

Some of the material is from Georgia Institute of Technology, Atlanta, GA, USA.

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

Information Extraction



Get Familiar with these concepts

NAME	DESCRIPTION
Tokenization	Segmenting text into words, punctuations marks etc.
Part-of-speech (POS) Tagging	Assigning word types to tokens, like verb or noun.
Dependency Parsing	Assigning syntactic dependency labels, describing the relations between individual tokens, like subject or object.
Lemmatization	Assigning the base forms of words. For example, the lemma of “was” is “be”, and the lemma of “texts” is “text”.
Sentence Boundary Detection (SBD)	Finding and segmenting individual sentences.

Get Familiar with these concepts

NAME	DESCRIPTION
Named Entity Recognition (NER)	Labelling named “real-world” objects, like persons, companies or locations.
Entity Linking (EL)	Disambiguating textual entities to unique identifiers in a knowledge base.
Similarity	Comparing words, text spans and documents and how similar they are to each other.
Text Classification	Assigning categories or labels to a whole document, or parts of a document.
Rule-based Matching	Finding sequences of tokens based on their texts and linguistic annotations, similar to regular expressions.
Training	Updating and improving a statistical model’s predictions.
Serialization	Saving objects to files or byte strings.

4.1 IE Applications

Identify specific pieces of information (data) in a unstructured or semi-structured textual document.
Transform unstructured information in a corpus of documents or web pages into a structured database.

Applied to different types of text:

- Newspaper articles
- Web pages
- Scientific articles
- Newsgroup messages
- Classified ads
- Medical notes

4.1 IE Applications

Job postings
Job resumes
Seminar announcements
Company information from the web
Apartment rental ads
Molecular biology information from MEDLINE

Sample Job Posting

Subject: US-TN-SOFTWARE PROGRAMMER
Date: 17 Nov 1996 17:37:29 GMT
Organization: Reference.Com Posting Service
Message-ID: <56nigp$mrs@bilbo.reference.com>

SOFTWARE PROGRAMMER

Position available for Software Programmer experienced in generating software for PC-Based Voice Mail systems. Experienced in C Programming. Must be familiar with communicating with and controlling voice cards; preferable Dialogic, however, experience with others such as Rhetorix and Natural Microsystems is okay. Prefer 5 years or more experience with PC Based Voice Mail, but will consider as little as 2 years. Need to find a Senior level person who can come on board and pick up code with very little training. Present Operating System is DOS. May go to OS-2 or UNIX in future.

Please reply to:
Kim Anderson
AdNET
(901) 458-2888 fax
kimander@memphisonline.com

Extracted Job Template

computer_science_job
id: 56nigp$mrs@bilbo.reference.com
title: SOFTWARE PROGRAMMER
salary:
company:
recruiter:
state: TN
city:
country: US
language: C
platform: PC \ DOS \ OS-2 \ UNIX
application:
area: Voice Mail
req_years_experience: 2
desired_years_experience: 5
req_degree:
desired_degree:
post_date: 17 Nov 1996

Amazon Book Description

```
....
</td></tr>
</table>
<b class="sans">The Age of Spiritual Machines : When Computers Exceed Human Intelligence</b><br>
<font face=verdana,arial,helvetica size=-1>
by <a href="/exec/obidos/search-handle-url/index=books&field-author=
        Kurzweil%2C%20Ray/002-6235079-4593641">
Ray Kurzweil</a><br>
</font>
<br>
<a href="http://images.amazon.com/images/P/0140282025.01.LZZZZZZZ.jpg">
<img src="http://images.amazon.com/images/P/0140282025.01.MZZZZZZZ.gif" width=90
   height=140 align=left border=0></a>
<font face=verdana,arial,helvetica size=-1>
<span class="small">
<span class="small">
<b>List Price:</b> <span class=listprice>$14.95</span><br>
<b>Our Price: <font color=#990000>$11.96</font></b><br>
<b>You Save:</b> <font color=#990000><b>$2.99 </b>
(20%)</font><br>
</span>
<p> <br>…
```

Extracted Book Template

Title: The Age of Spiritual Machines :
When Computers Exceed Human Intelligence
Author: Ray Kurzweil
List-Price: $14.95
Price: $11.96
:
:

Web Extraction

Many web pages are generated automatically from an underlying database.
Therefore, the HTML structure of pages is fairly specific and regular (*semi-structured*).
However, output is intended for human consumption, not machine interpretation.
An IE system for such generated pages allows the web site to be viewed as a structured database.
An extractor for a semi-structured web site is sometimes referred to as a *wrapper*.
Process of extracting from such pages is sometimes referred to as *screen scraping*.

Template Types

Slots in template typically filled by a substring from the document.

Some slots may have a fixed set of pre-specified possible fillers that may not occur in the text itself.

- Terrorist act: threatened, attempted, accomplished.
- Job type: clerical, service, custodial, etc.
- Company type: SEC code

Some slots may allow multiple fillers.

- Programming language

Some domains may allow multiple extracted templates per document.

- Multiple apartment listings in one ad

Simple Extraction Patterns

Specify an item to extract for a slot using a regular expression pattern.

Price pattern: “\b\$\d+(\.\d{2})?\b”

May require preceding (pre-filler) pattern to identify proper context.

Amazon list price:

Pre-filler pattern: “List Price: ”
Filler pattern: “\$\d+(\.\d{2})?\b”

May require succeeding (post-filler) pattern to identify the end of the filler.

Amazon list price:

Pre-filler pattern: “List Price: ”
Filler pattern: “.+”
Post-filler pattern: “”

Pre-Specified Filler Extraction

If a slot has a fixed set of pre-specified possible fillers, text categorization can be used to fill the slot.

- Job category
- Company type

Treat each of the possible values of the slot as a category, and classify the entire document to determine the correct filler.

4.2 IE Tasks

It can be a Human-Performed IE Task or an Automated IE Task (which can be solved by an IE System that implements an IE algorithm).

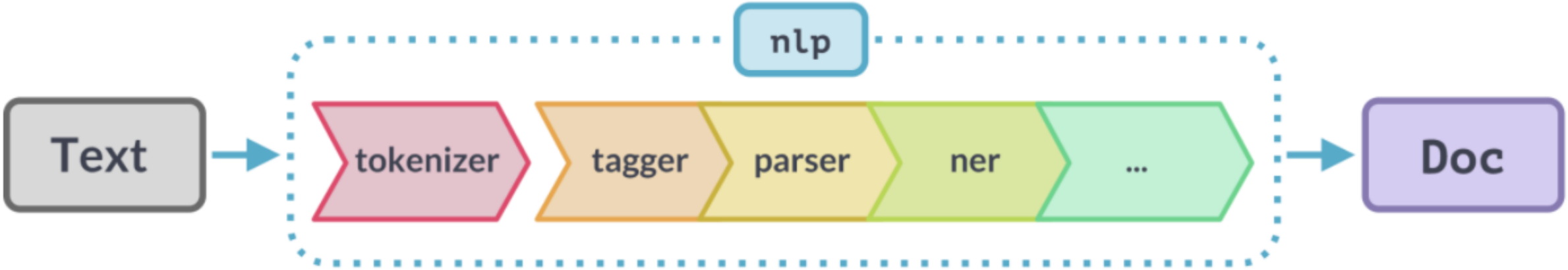
It can be Unstructured IE Task (such as from text or from images) or a Semi-Structured IE Task (such as from web-tables) to being an Structured IE Task (such as from databases)

It can be a Manual IE Task or automated Automated IE Task.

Example(s):

A Citation Extraction Task, that can support a Citation Search Service.
A Biomedical Information Extraction Task.
The extraction of Consumer Product Data to facilitate Searching (e.g. Google Products)
Employment Posting Extraction and Resume Extraction.

4.3 The General Pipeline for IE



IE Models

There are several state-of-the-art models we could rely on. Below are some of the frequently use open-source models:

1.Named Entity Recognition on CoNLL 2003 (English)

2.Key Information Extraction From Documents: Evaluation And Generator

3.Deep Reader: Information extraction from Document images
via relation extraction and Natural Language

Evaluation of the Model

We evaluate the training process is crucial before we use the models in production. This is usually done by creating a **testing dataset** and finding some key metrics:

- **Accuracy**: the ratio of correct predictions made against the size of the test data.
- **Precision**: the ratio of true positives and total predicted positives.
- **Recall** the ratio of true positives and total actual positives.
- **F1-Score**: harmonic mean of precision and recall.

Evaluating IE Accuracy

Always evaluate performance on independent, manually-annotated test data not used during system development.

Measure for each test document:

Total number of correct extractions in the solution template: *N*

Total number of slot/value pairs extracted by the system: *E*

Number of extracted slot/value pairs that are correct (i.e. in the solution template): *C*

Compute average value of metrics adapted from IR:

Recall = *C*/*N*

Precision = *C*/*E*

F-Measure = Harmonic mean of recall and precision

Demo Application

Difference between Information Retrieval and Information Extraction

Extraction means “pulling out” and Retrieval means “getting back.”

Information retrieval is about returning the information that is relevant for a specific query or field of interest of the user.

While information extraction is more about extracting general knowledge (or relations) from a set of documents or information.

Information extraction is the standard process of taking data and extracting structured information from it so that it can be used for various purposes, one of which may be in a search engine.

Source: Geeks for Geeks

4.4 Case Studies on IE

Case Study 1 – PwC AI automation Information Extraction

https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/publications/ai-automation-data-extraction.html

Case Study 2 – Research paper on Information Extraction

https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0254-8

Next topic: Chatbot