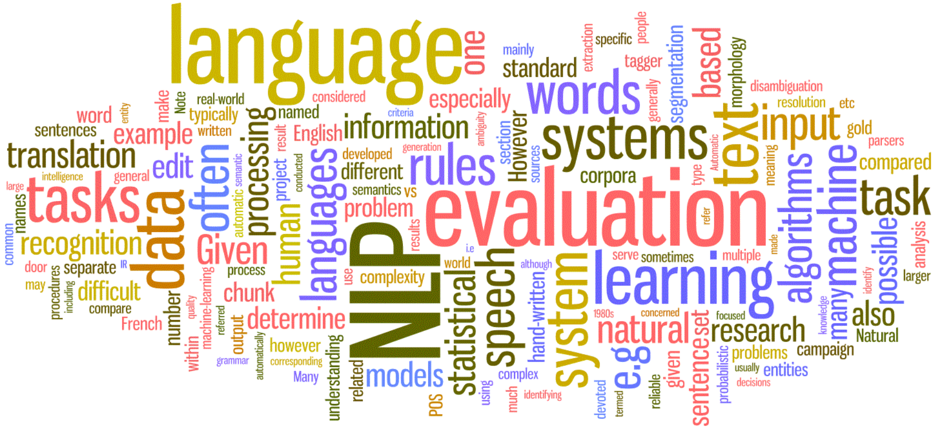
Information Extraction
Named Entities Recognition
Relation Extraction
De-Identification



Dr. Liao

3/31/2020

Overview

- Information Extraction (IE)
  - Definition, Architectures & Examples
  - Automatic Content Extraction (ACE)
- Tasks and Subtasks
  - Named Entity Recognition (NER)
  - Relation Extraction
  - Event Extraction
- Applied NLP Teches with Hands-On Programming Practice
  - Intro to SpaCy
  - NER in SpaCy & NLTK
  - De-Identification in Python
  - Web scraping for the entire webpage in BeautifulSoup
- Homework

Information Extraction

- Information extraction (IE) systems
  - Find and understand limited relevant parts of texts
  - Gather information from many pieces of text
  - Produce a structured representation of relevant information:
    - *relations* (in the database sense), a.k.a.,
    - a *knowledge base*
  - Goals:
    1. Organize information so that it is useful to people
    2. Put information in a semantically precise form that allows further inferences to be made by computer algorithms

IE Examples

- The process of converting unstructured text into structured information

New York Times Co. named Russell T. Lewis, 45, president and general manager of its flagship New York Times newspaper, responsible for all business-side activities. He was executive vice president and deputy general manager. He succeeds Lance R. Primis, who in September was named president and chief operating officer of the parent.

Person	Company	Post	State
Russell T. Lewis	New York Times newspaper	president and general manager	start
Russell T. Lewis	New York Times newspaper	executive vice president	end
Lance R. Primis	New York Times Co.	president and CEO	start

Yoav Artzi: Natural language processing

IE Examples (Cont.)

- Biography Info Extraction from the Webpage

Biography for
Elvis Presley More at IMD

Date of Birth
8 January 1935, Tupelo, Mississippi, USA

Date of Death
16 August 1977, Memphis, Tennessee, USA (

Birth Name
Elvis Aron Presley

Nickname
The Pelvis
The King
The King Of Rock 'n

Height
6' (1.83 m)

Mini Biography

Elvis Aaron Presley

Name	Birthplace	Birthdate
Elvis Presley	Tupelo, MI	1935-01-08
...	...	

DISCOVER ELVIS

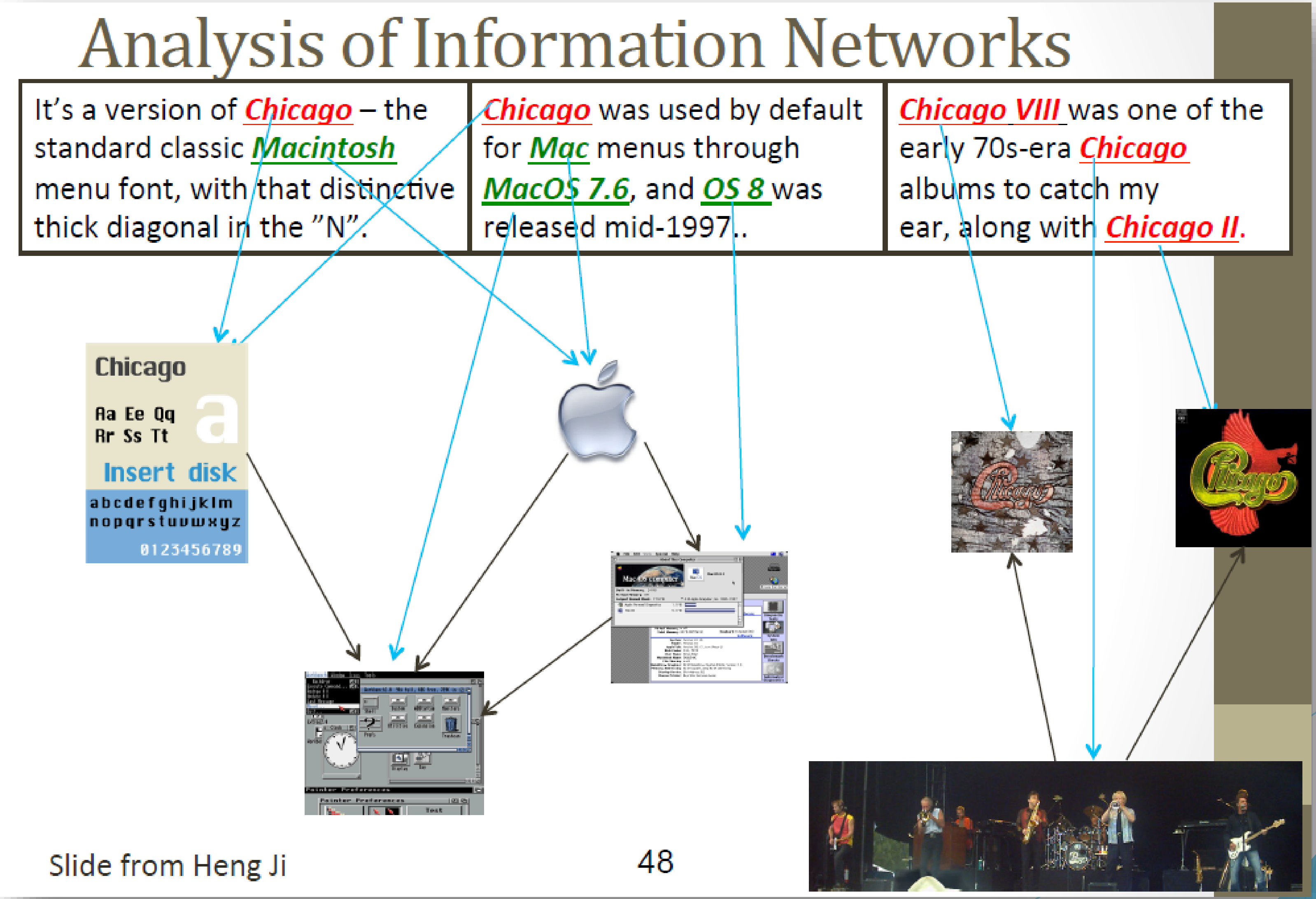
Biography
Overview / 1935-1957 / 1958-1965 / 1966-1969 / 1970-1977

Overview

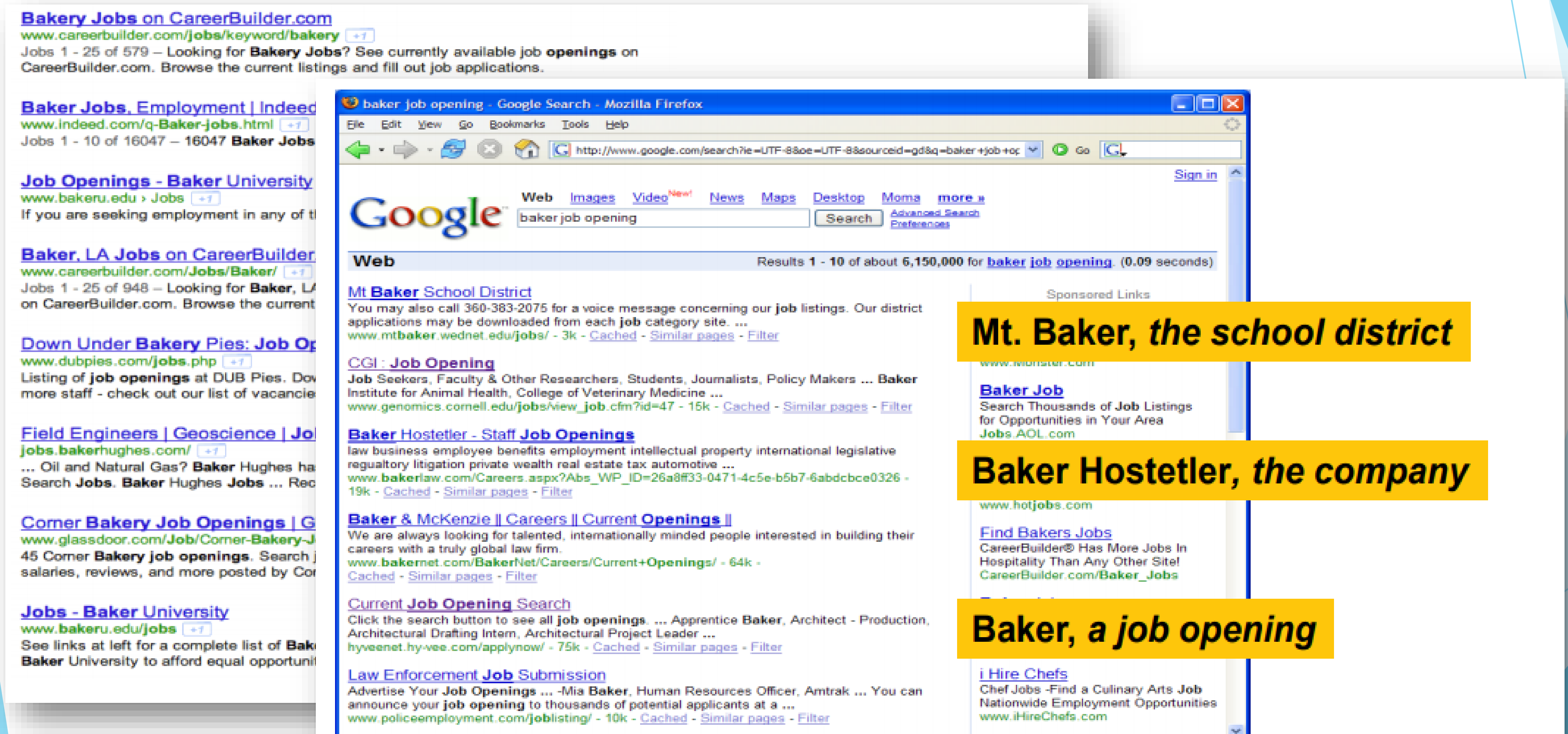
Elvis Aaron Presley, in the humblest of circumstances, was born to Vernon and Gladys Presley in a two-room house in Tupelo, Mississippi on January 8, 1935. His twin brother, Jessie Garon, was stillborn, leaving Elvis to grow up as an only child. He and his parents moved to Memphis, Tennessee in 1948, and Elvis graduated from Humes High School there in 1953.

Slide from Suchanek

IE Examples (Cont.) – Coreference Resolution (disambiguation to Wikipedia)

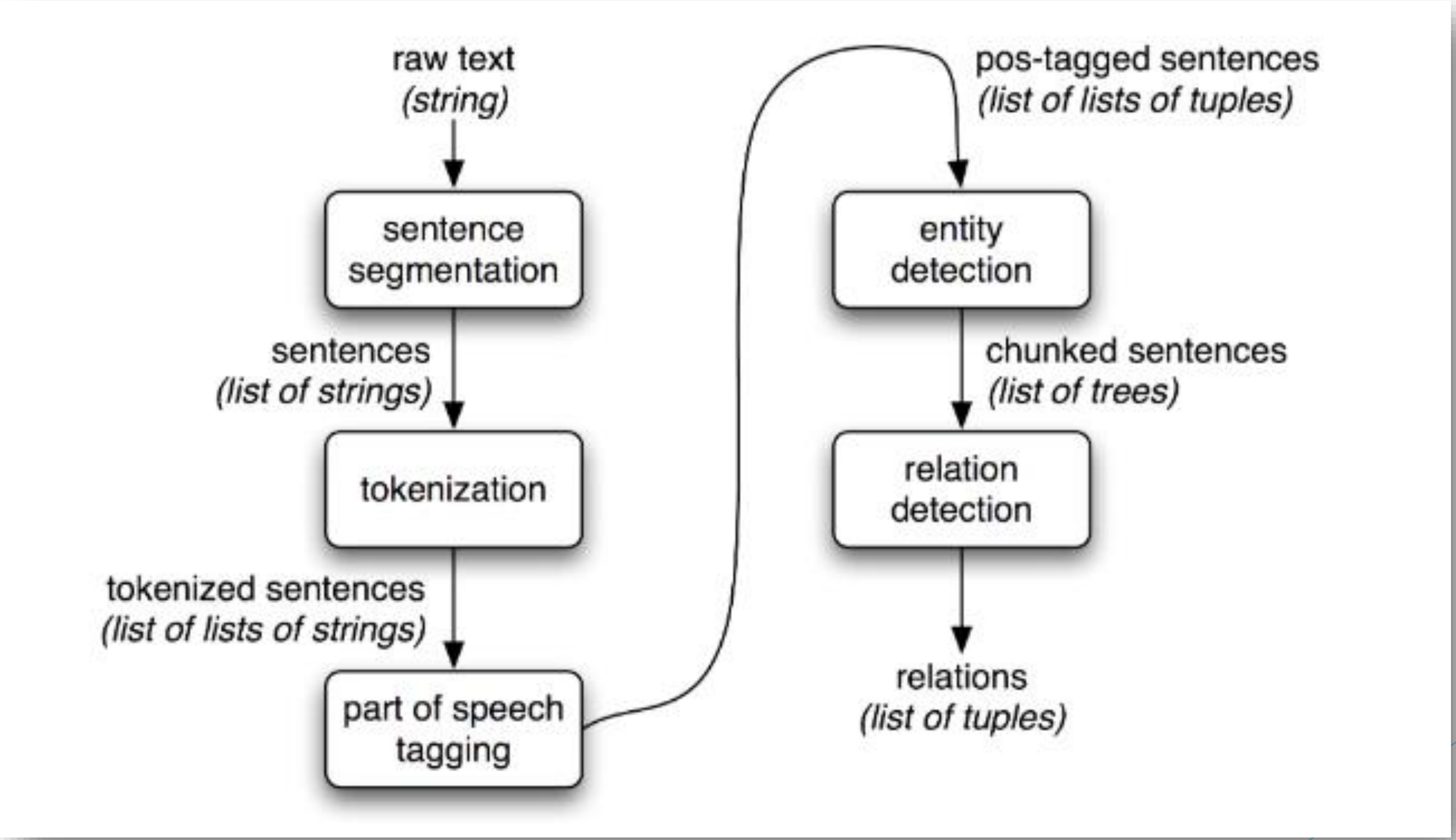


IE Examples (Cont.)



Slide from Cohen/Mccallum

A Simple IE Architecture



NLTK book: [Information Extraction]

IE Major Tasks and Subtasks

- Named Entity Recognition (NER)
  - Go to Lecture 9 - Named Entity Recognition_Rev.pptx
- Relation Extraction
- Events Extraction

Examples of Extraction of **Named Entities**, **Relations**, and **Events**

Dr. Liao, a professor from George Mason University in Fairfax VA, is teaching a virtual online class now.

- **Named Entities:**
  - Dr. Liao (PER), George Mason University (ORG), Fairfax VA (LOC)
- Relations
  - Person – org:
    - Dr. Liao from George Mason University
  - Org – Location:
    - George Mason University in Fairfax VA
- Event:
  - Dr. Liao is teaching a virtual online class (now)

Why Relation Extraction?

- Create new structed knowledge bases, useful for any apps
- Augment current knowledge bases
  - Adding words to WordNet thesaurus, facts to FreeBase or DBPedia
- Support Question Answering (QA)
  - Q: Where is George Mason University?
  - A: Fairfax, VA
  - Q: What is Dr. Liao teaching now?
  - A: A virtual online class.

But *which relations should we extract?*

Automatic Content Extraction (ACE)

- ACE is a research program for developing advanced IE technologies
  - 1999-2008 by NIST, succeeding MUC and preceding Text Analysis Conference
- Given a text in natural language, the ACE challenge is to detect:
  - **Entities**
    - persons, organizations, locations, facilities, weapons, vehicles, and geo-political entities
  - **Relations** between entities
    - Relation types include: role, part, located, near, and social
    - E.g., person A is the manager of company B
  - **Events**
    - interaction, movement, transfer, creation and destruction

Content reference from Wikipedia

Automatic Content Extraction (ACE)

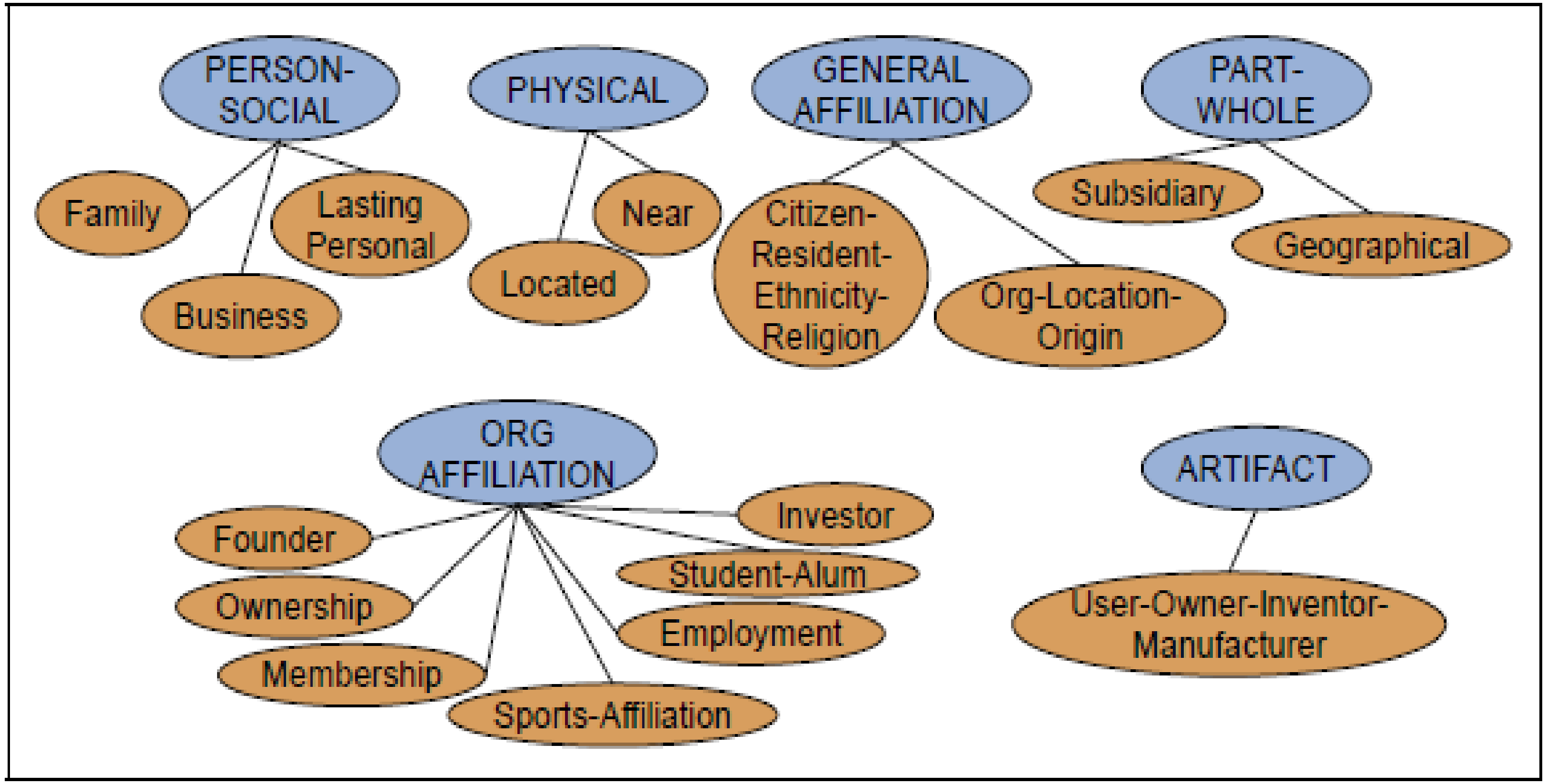


Figure 17.9 The 17 relations used in the ACE relation extraction task.

Image Courtesy of Stanford Professor Dr. Dan Jurafsky

Automatic Content Extraction (ACE)

- Physical-Located PER-GPE

  He was in Tennessee

- Part-Whole-Subsidiary ORG-ORG

  XYZ, the parent company of ABC

- Person-Social-Family PER-PER

  John's wife Yoko

- Org-AFF-Founder PER-ORG

  Steve Jobs, co-founder of Apple...

Content Courtesy of Stanford Professor Dr. Dan Jurafsky

Relation Extraction

- Approaches
  - Pattern Matching
  - Supervised Learning
  - Semi-Supervised Learning
  - Unsupervised Learning
  - Distantly Supervised Learning

Relation Extraction (Cont.)

- <u>Pattern Matching</u>
  - Patterns
    - "[PER] was born in [LOC]"
    - "[PER] was graduated from [ORG]"
  - Matching Techniques
    - Exact matching
    - Flexible matching

Which techniques we have learnt can be used?

Relation Extraction (Cont.)

- Supervised Learning
  - Classifier
    - Naïve Bayes, SVM, etc.
  - Features
    - Types of two named entities
    - Bag of words
    - POS of words in between
  - Pros
    - Doesn’t require iteratively expanding patterns

Relation Extraction

- Approaches
  - Pattern Matching (rule based)
  - Supervised Learning
  - Semi-Supervised Learning
    - Bootstrapping
  - Unsupervised Learning
    - Uses very large amounts of unlabeled data
    - Not sensitive to genre issues in training corpus
  - Distantly Supervised Learning
    - Combo

Relation Extraction
- Example 1: Disease Outbreaks

May 19 1995, Atlanta -- The Centers for Disease Control and Prevention, which is in the front line of the world's response to the deadly Ebola epidemic in Zaire , is finding itself hard pressed to cope with the crisis...

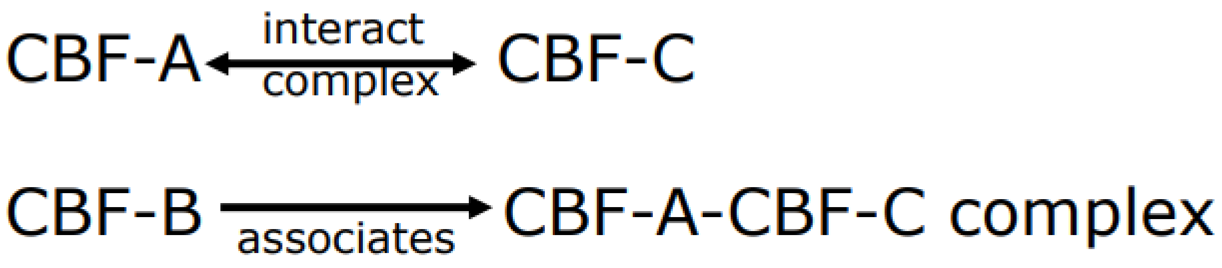
Information Extraction System

Date	*Disease Name*	*Location*
Jan. 1995	Malaria	Ethiopia
July 1995	Mad Cow Disease	U.K.
Feb. 1995	Pneumonia	U.S.

Slide from Manning

Relation Extraction
- Example 2: Protein Interactions

"We show that CBF-A and CBF-C interact with each other to form a CBF-A-CBF-C complex and that CBF-B does not interact with CBF-A or CBF-C individually but that it associates with the CBF-A-CBF-C complex."



Slide from Manning

Rough Accuracy of Information Extraction

Information type	Accuracy
Entities	90-98%
Attributes	80%
Relations	60-70%
Events	50-60%

- Errors cascade (error in entity tag → error in relation extraction)
- These are very rough, actually optimistic, numbers
  - Hold for well-established tasks, but lower for many specific/novel IE tasks

Slide from Manning

NLP Hands-On Programming in Class

- Code Examples & Tutorials for
  - **NER and De-Identification** with NLTK, SpaCy, and Python
  - **Web scraping for the entire webpage** in BeautifulSoup
  - Dr. Liao wrote them **particularly** for
    - this course learning
    - **Assignments, labs, and final project** examples
- All programming tutorials & code example demos
  - Using online **Jupyter Lab** in class
- References:
  - NLTK book: Information Extraction
  - SpaCy NER documentations

Intro to SpaCy - Industrial-Strength NLP

- SpaCy Website: https://spacy.io/
- Install SpaCy
- Download / Import / Load SpaCy English Model
- Import SpaCy displacy for rendering NER

```
import spacy
# Use the command to install the SpaCy:
# > pip install -U spacy
```

```
## Use the command to download the SpaCy English model:
# > python -m spacy download en_core_web_sm
```

```
# Import SpaCy English model
import en_core_web_sm

# Load English tokenizer, tagger, parser, NER and word vectors
nlp = en_core_web_sm.load()
```

```
from spacy import displacy
```

The code snippets from Dr. Liao's code examples

SpaCy Annotations for NER

- [SpaCy Annotation for NER]

TYPE	DESCRIPTION
PERSON	People, including fictional
NORP	Nationalities or religious or political groups
FACILITY	Buildings, airports, highways, bridges, etc
ORG	Companies, agencies, institutions, etc
GPE	Countries, cities, states
LOC	Non-GPE locations, mountain ranges, bodies of water
PRODUCT	Objects, vehicles, foods, etc (Not services)
EVENT	Named hurricanes, battles, wars, sports events, etc
WORK_OF_ART	Titles of books, songs, etc
LAW	Named documents made into laws
LANGUAGE	Any named language
DATE	Absolute or relative dates or periods.
TIME	Times smaller than a day
PERCENT	Percentage, including "%".
MONEY	Monetary values, including unit
QUANTITY	Measurements, as of weight or distance
ORDINAL	"first", "second", etc
CARDINAL	Numerals that do not fall under another type

IOB Scheme

TAG	ID	DESCRIPTION
"I"	1	Token is inside an entity.
"O"	2	Token is outside an entity.
"B"	3	Token begins an entity.
""	0	No entity tag is set (missing value).

BILUO Scheme

TAG	DESCRIPTION
B EGIN	The first token of a multi-token entity.
I N	An inner token of a multi-token entity.
L AST	The final token of a multi-token entity.
U NIT	A single-token entity.
O UT	A non-entity token.

Wikipedia scheme

Models trained on Wikipedia corpus (Nothman et al., 2013) use a less fine-grained NER annotation scheme and recognise the following entities:

TYPE	DESCRIPTION
PER	Named person or family.
LOC	Name of politically or geographically defined location (cities, provinces, countries, international regions, bodies of water, mountains).
ORG	Named corporate, governmental, or other organizational entity.
MISC	Miscellaneous entities, e.g. events, nationalities, products or works of art.

The tables from SpaCy website

SpaCy Annotation for Token Entity

Annote the token-level entity

using the BILUO tagging scheme to describe the entity boundaries

```
pprint([(X, X.ent_iob_, X.ent_type_) for X in mytext])
```

```
[(A, 'O', ''),
 (U.S., 'B', 'NORP'),
 (Marine, 'I', 'NORP'),
 (who, 'O', ''),
 (is, 'O', ''),
 (assigned, 'O', ''),
 (to, 'O', ''),
 (Fort, 'B', 'GPE'),
 (Belvoir, 'I', 'GPE'),
 (in, 'O', ''),
 (Fairfax, 'B', 'GPE'),
 (County, 'I', 'GPE'),
 (and, 'O', ''),
 (lives, 'O', ''),
 (at, 'O', ''),
 (Marine, 'B', 'ORG'),
 (Corps, 'I', 'ORG'),
 (Base, 'I', 'ORG'),
 (Quantico, 'I', 'ORG'),
```

The code snippets & running results from Dr. Liao's code examples

POS Extraction & Lemmatization In SpaCy

Extract part-of-speech and lemmatize the text

```
[(x.orth_,x.pos_, x.lemma_) for x in [y
                                      for y
                                      in mytext
                                      if not y.is_stop and y.pos_ != 'PUNCT']]
```

```
[('U.S.', 'PROPN', 'U.S.'),
 ('Marine', 'PROPN', 'Marine'),
 ('assigned', 'VERB', 'assign'),
 ('Fort', 'PROPN', 'Fort'),
 ('Belvoir', 'PROPN', 'Belvoir'),
 ('Fairfax', 'PROPN', 'Fairfax'),
 ('County', 'PROPN', 'County'),
 ('lives', 'VERB', 'live'),
 ('Marine', 'PROPN', 'Marine'),
 ('Corps', 'PROPN', 'Corps'),
 ('Base', 'PROPN', 'Base'),
 ('Quantico', 'PROPN', 'Quantico'),
 ('Prince', 'PROPN', 'Prince'),
 ('William', 'PROPN', 'William'),
 ('County', 'PROPN', 'County'),
 ('state', 'NOUN', 'state'),
 ('diagnosed', 'VERB', 'diagnose'),
 ('case', 'NOUN', 'case'),
 ('Virginia', 'PROPN', 'Virginia'),
 ('Governor', 'PROPN', 'Governor'),
 ('Ralph', 'PROPN', 'Ralph'),
 ('Northam', 'PROPN', 'Northam'),
 ('asking', 'VERB', 'ask'),
 ('volunteers', 'NOUN', 'volunteer'),
 ('staff', 'VERB', 'staff'),
 ('Virginia', 'PROPN', 'Virginia'),
 ('Medical', 'PROPN', 'Medical'),
 ('Reserve', 'PROPN', 'Reserve'),
 ('Corps', 'PROPN', 'Corps'),
 ('state', 'NOUN', 'state'),
 ('reaches', 'VERB', 'reach'),
 ('600', 'NUM', '600'),
 ('coronavirus', 'NOUN', 'coronavirus'),
 ('cases', 'NOUN', 'case')]
```

The code snippets & running results from Dr. Liao's code examples

Get Named Entities in SpaCy

Get the named entities

[42]:
```
pprint([(X.text, X.label_) for X in mytext.ents])
```

```
[('U.S. Marine', 'NORP'),
 ('Fort Belvoir', 'GPE'),
 ('Fairfax County', 'GPE'),
 ('Marine Corps Base Quantico', 'ORG'),
 ('Prince William County', 'GPE'),
 ('first', 'ORDINAL'),
 ('Virginia', 'GPE'),
 ('Ralph Northam', 'PERSON'),
 ('the Virginia Medical Reserve Corps', 'ORG'),
 ('more than 600', 'CARDINAL')]
```

Count all the named entities

[45]:
```
labels = [x.label_ for x in mytext.ents]
Counter(labels)
```

[45]:
```
Counter({'NORP': 1,
         'GPE': 4,
         'ORG': 2,
         'ORDINAL': 1,
         'PERSON': 1,
         'CARDINAL': 1})
```

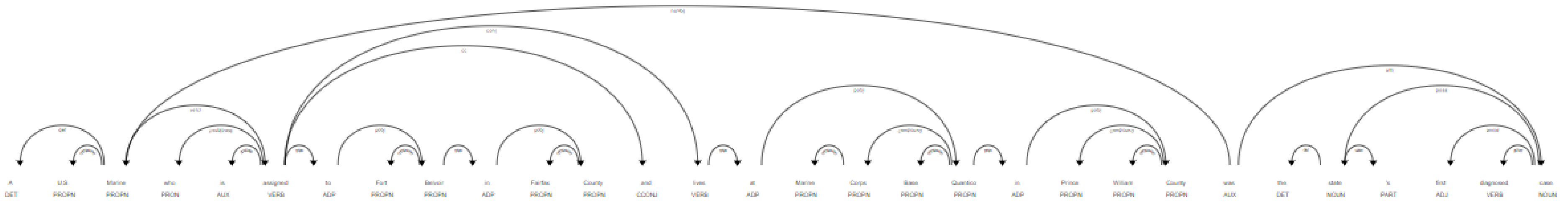
The code snippets & running results from Dr. Liao's code examples

Visualize Entities and Dependencies in SpaCy

displacy.render(mytext, style='ent’, , jupyter=True)

[A U.S. **GPE** Marine who is assigned to Fort Belvoir **GPE** in Fairfax County **GPE** and lives at Marine Corps Base Quantico **ORG** in Prince William County **GPE** was the state's first **ORDINAL** diagnosed case., Virginia **GPE** Governor Ralph Northam **PERSON** is asking for volunteers to staff the Virginia Medical Reserve Corps **ORG** as the state reaches more than 600 **CARDINAL** coronavirus cases.]

displacy.render(mytext), style='dep’, jupyter = True, options = {'distance': 100})



The code snippets & running results from Dr. Liao’s code examples nltk.ne_chunk() recognizes named entities using a classifier: PERSON, ORGANIZATION, and GPE

```
ne_tree = nltk.chunk.ne_chunk(pos_tag(word_tokenize(mytext)))
print(ne_tree)
```

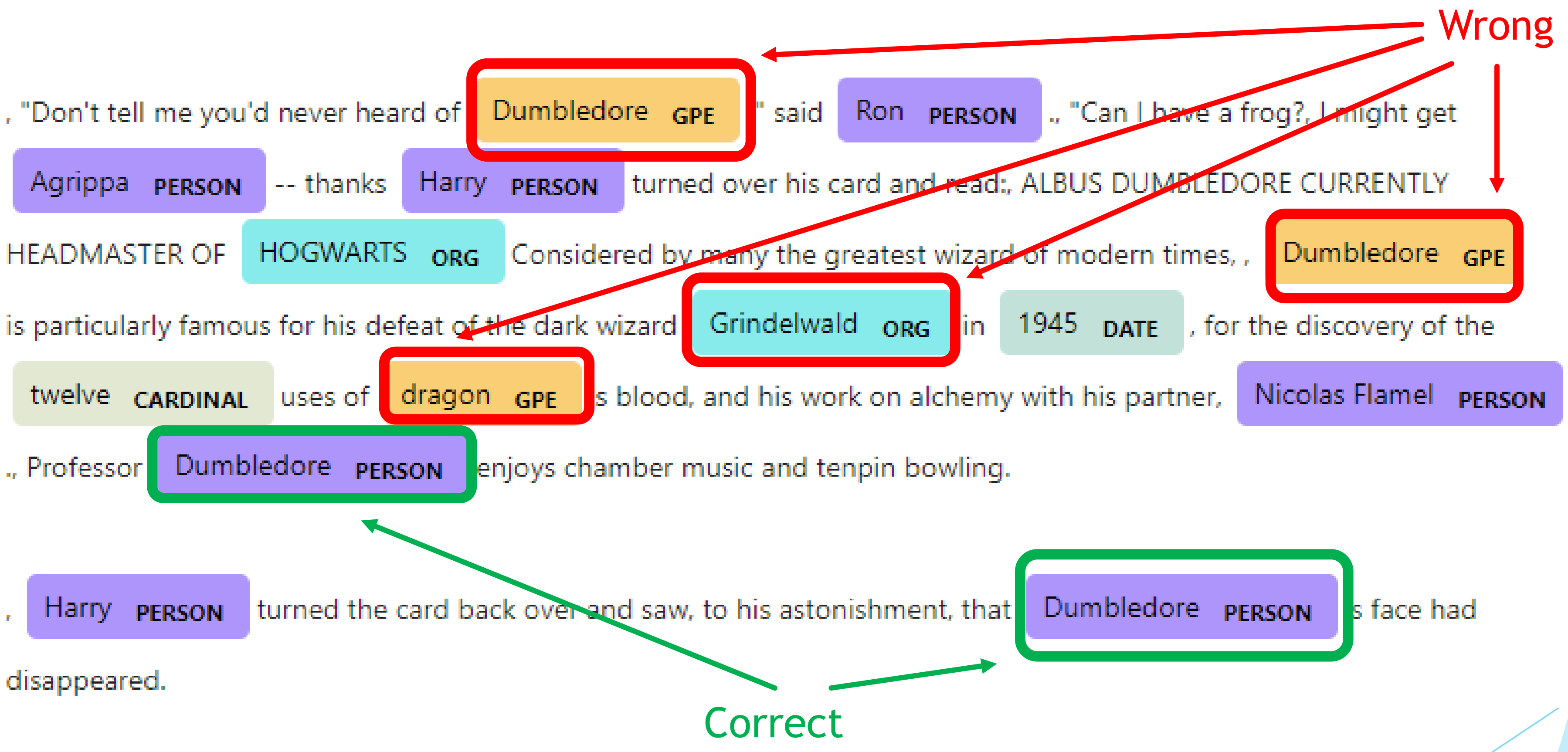
```
(S
  A/DT
  (GPE U.S./NNP)
  Marine/NNP
  who/WP
  is/VBZ
  assigned/VBN
  to/TO
  (ORGANIZATION Fort/NNP Belvoir/NNP)
  in/IN
  (GPE Fairfax/NNP County/NNP)
  and/CC
  lives/NNS
  at/IN
  (FACILITY Marine/NNP Corps/NNP Base/NNP Quantico/NNP)
  in/IN
  (GPE Prince/NNP)
  (PERSON William/NNP County/NNP)
  was/VBD
  the/DT
  state/NN
  's/POS
  first/JJ
  diagnosed/VBN
  case./NN
  ,/,
  (GPE Virginia/NNP)
  Governor/NNP
  (PERSON Ralph/NNP Northam/NNP)
  is/VBZ
  asking/VBG
  for/IN
  volunteers/NNS
  to/TO
  staff/NN
  the/DT
  (ORGANIZATION Virginia/NNP Medical/NNP Reserve/NNP Corps/NNP)
  as/IN
  the/DT
  state/NN
  reaches/VBZ
  more/JJR
  than/IN
  600/CD
```

NLTK - NER

- nltk.ne_chunk()
- Wrong Recognition
  - **Prince** is recognized as GPE.
  - **William County** is recognized as a person.

Prince William County should be recognized as **GPE**.

Incorrect NER with SpaCy in RED



The running results generated from Dr. Liao's code examples

De-Identification

- Redact Names

[A U.S. GPE Marine who is assigned to Fort Belvoir GPE in Fairfax County GPE and lives at Marine Corps Base Quantico ORG in Prince William County GPE was the state's first ORDINAL diagnosed case., Virginia GPE Governor Ralph Northam PERSON is asking for volunteers to staff the Virginia Medical Reserve Corps ORG as the state reaches more than 600 CARDINAL coronavirus cases.]

```
A U.S. Marine who is assigned to Fort Belvoir in Fairfax County and lives at
Marine Corps Base Quantico in Prince William County was the state's first dia
gnosed case., Virginia Governor [REDACTED] [REDACTED] is asking for volunteer
s to staff the Virginia Medical Reserve Corps as the state reaches more than
600 coronavirus cases.
```

```
displacy.render(nlp(redacted), jupyter=True, style='ent')
```

A U.S. Marine NORP who is assigned to Fort Belvoir GPE in Fairfax County GPE and lives at Marine Corps Base Quantico ORG in Prince William County GPE was the state's first ORDINAL diagnosed case., Virginia GPE Governor [REDACTED] [REDACTED] is asking for volunteers to staff the Virginia Medical Reserve Corps ORG as the state reaches more than 600 CARDINAL coronavirus cases.

The running results generated from Dr. Liao's code examples

Web Scraping for the Entire Webpage

Scrape the text from the webpage using BeautifulSoup

```
from bs4 import BeautifulSoup
import requests
import re
```

```
def _scrape_webtext(url):
    """
    Scrape the text from the webpage using BeautifulSoup
    """
    res = requests.get(url)
    html = res.text
    soup = BeautifulSoup(html, 'html5lib')
    for script in soup(["script", "style", 'aside']):
        script.extract()
    return " ".join(re.split(r'[\n\t]+', soup.get_text()))
```

Extract named entity from the webpage

```
news = _scrape_webtext('https://www.nbcwashington.com/news/local/latest-updates-how-many-coronavirus-covid-diagnosed-confirmed-cases-test-death
webtext = nlp(news)
print(news[:1500])
```

```
The Latest: 3,405 Coronavirus Cases Diagnosed in DC, Maryland, Virginia – NBC4 Washington Skip to content coronavirus The Latest: 3,405 Coronav
irus Cases Diagnosed in DC, Maryland, Virginia Here are the latest numbers on COVID-19 diagnoses and related deaths in D.C., Maryland and Virgi
nia By Sophia Barnes and NBC Washington Staff • Published March 3, 2020• Updated 3 hours ago NBC Universal, Inc. D.C. Mayor Muriel Bowser has i
ssued a stay-at-home order. News4's Mark Segraves reports city leaders are concerned that some who have died from coronavirus were scared to se
ek medical treatment due to their immigration status. As of Tuesday, 3,405 cases of coronavirus had been announced. D.C. had 495 cases, Marylan
d had 1,660 and Virginia had 1,250. The virus has infected a broad range of people, from an 8-week-old baby boy to elderly nursing home residen
ts. For most people, the coronavirus causes only mild or moderate symptoms including fever. shortness of breath and cough. Recovery might take
about two weeks. Severe illness including pneumonia can occur, especially in the elderly and people with existing health problems, and recovery
could take six weeks in such cases.  Local Maryland 28 mins ago 3 More Deaths in COVID-19 Outbreak at Maryland Nursing Home coronavirus in west
virginia 2 hours ago West Virginia Announces 17 New Coronavirus Cases; Hospital Closes Due To Financial Hardships At least fifty-four people in
D.C., Maryland and Virginia have died from COVID-19, health officia
```

The code snippets & running results from Dr. Liao's code examples

More Code Examples

- Please see Dr. Liao’s code examples and tutorials for both NLTK and SpaCy in class
  - NER
  - De-Identification
  - Web scraping for the entire webpage

Student Presentations

- Team 4
  - NAACL-HLT 2019
  - GraphIE: A Graph-Based Framework for Information Extraction
  - Yujie Qian, Enrico Santus, Zhijing Jin, Jiang Guo, Regina Barzilay
  - MIT
- Team 6
  - EMNLP 2018
  - Improving Neural Abstractive Document Summarization with Structural Regularization
  - Wei Li, Xinyan Xiao, Yajuan Lyu, Yuanzhuo Wang
  - Chinese Academy of Sciences & Baidu

Homework

- Previous Homework
  - Programming Assignment 3
    - WSD
    - Due on 3/31
  - Term Project Checkpoint 2
    - Due on 4/5 Midnight
- Today's Homework
  - Optional Lab 3 –
    NER & De-Identification
    - Due on 4/14
  - Student Presentation on Class 10
    - Team 3 & Team 6
    - PPT Slides Due on 4/6 Midnight