Deep Learning for NLP Intro

Jay Urbain, PhD
jay.urbain@gmail.com
https://www.linkedin.com/in/jayurbain/

TOPICS

- What is Natural Language Processing
- Why is NLP hard?
- Core NLP problems: part-of-speech tagging, language modeling, etc.
- NLP Applications
- NLP Vocabulary

WHAT IS NATURAL LANGUAGE PROCESSING (NLP)?

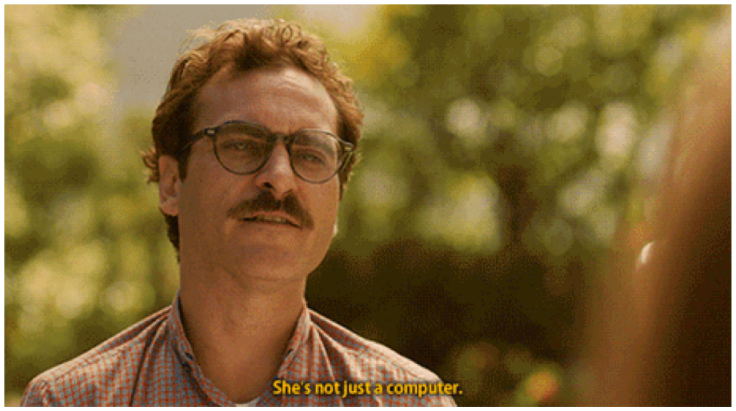
WHAT IS A NATURAL LANGUAGE PROCESSING (NLP)?

"Natural language processing (NLP) is an area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data. "

Wikipedia

GOAL: for computers to process or "*understand*" natural language in order to perform tasks that are useful such as question answering

- Add structure to unstructured text
- Full understanding for now is unsolved



Core NLP Tasks

- Tokenization
- Part-of-speech tagging
- Noun chunking
- Parsing
- Named entity classification
- Information extraction

Tokenization

```
import spacy

nlp = spacy.load("en_core_web_sm") # load model package "en_core_web_sm"
doc = nlp(u"Apple is looking at buying U.K. startup for $1 billion")
for token in doc:
    print(token.text)

doc[5]
```

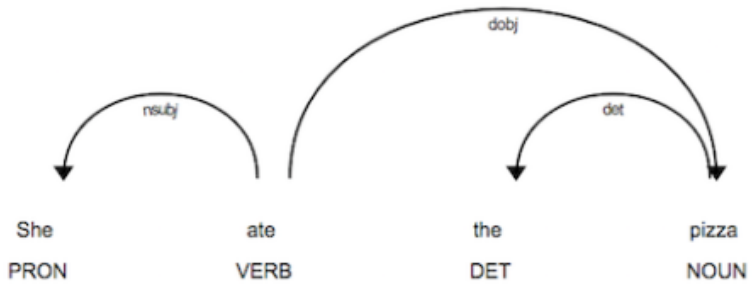
```
Apple
is
looking
at
buying
U.K.
startup
for
$
1
billion
```

Part-Of-Speech Tagging (POS)

```python
import spacy

# Load the small English model
nlp = spacy.load('en_core_web_sm')

# Process a text
doc = nlp("Chronic obstructive pulmonary disease (COPD) is a chronic

# Iterate over the tokens
for token in doc:
    # Print the text and the predicted part-of-speech tag
    print(token.text, token.pos_)
```

```
Chronic ADJ
obstructive ADJ
pulmonary ADJ
disease NOUN
( PUNCT
COPD NOUN
) PUNCT
is VERB
a DET
chronic ADJ
inflammatory ADJ
lung NOUN
disease NOUN
that DET
causes VERB
obstructed VERB
airflow ADV
from ADP
the DET
lungs NOUN
. PUNCT
```

Dependency Parsing



Label	Description	Example
nsubj	nominal subject	She
dobj	direct object	pizza
det	determiner (article)	the

- The pronoun "She" is a nominal subject attached to the verb "ate".
- The noun "pizza" is a direct object attached to the verb "ate". It is eaten by the subject, "she".
- The determiner "the", also known as an article, is attached to the noun "pizza".

Named Entity Identification

```
doc = nlp(u"Jay Urbain, is an aging caucasian male suffering from illusion

# Iterate over the predicted entities
for ent in doc.ents:
    # Print the entity text and its label
    print(ent.text, ent.label_)
```

```
Jay Urbain PERSON
LBP ORG
Jay PERSON
N. Tennyson PERSON
Disturbia PERSON
WI ORG
Kimberly Urbain PERSON
```

NLP APPLICATIONS

- Keyword search
- Text classification
- Named entity recognition
- Machine translation
- Question answering, chatbot dialog systems
- Much more...

CLINICAL & TRANSLATIONAL RESEARCH INFORMATICS

Patient Deidentification

https://cis.ctsi.mcw.edu/deid/

2019-04-04 15:13:09.055

Date offset:

10

Patient name:

Input text:

Jay Urbain, jay.urbain@gmail.com, born December 6, 2156 is an elderly caucasian male suffering from illusions of grandeur and LBP. He is married to Kimberly Urbain, who is much better looking. Patient father, Francis Urbain has a history of CAD and DM. Jay has been prescribed meloxicam, and venti americano. He lives at 9050 N. Tennyson Dr., Disturbia, WI with his wife and golden retriever Mel. You can reach him at 414-745-5102.

Data Format Pretty Print

Submit

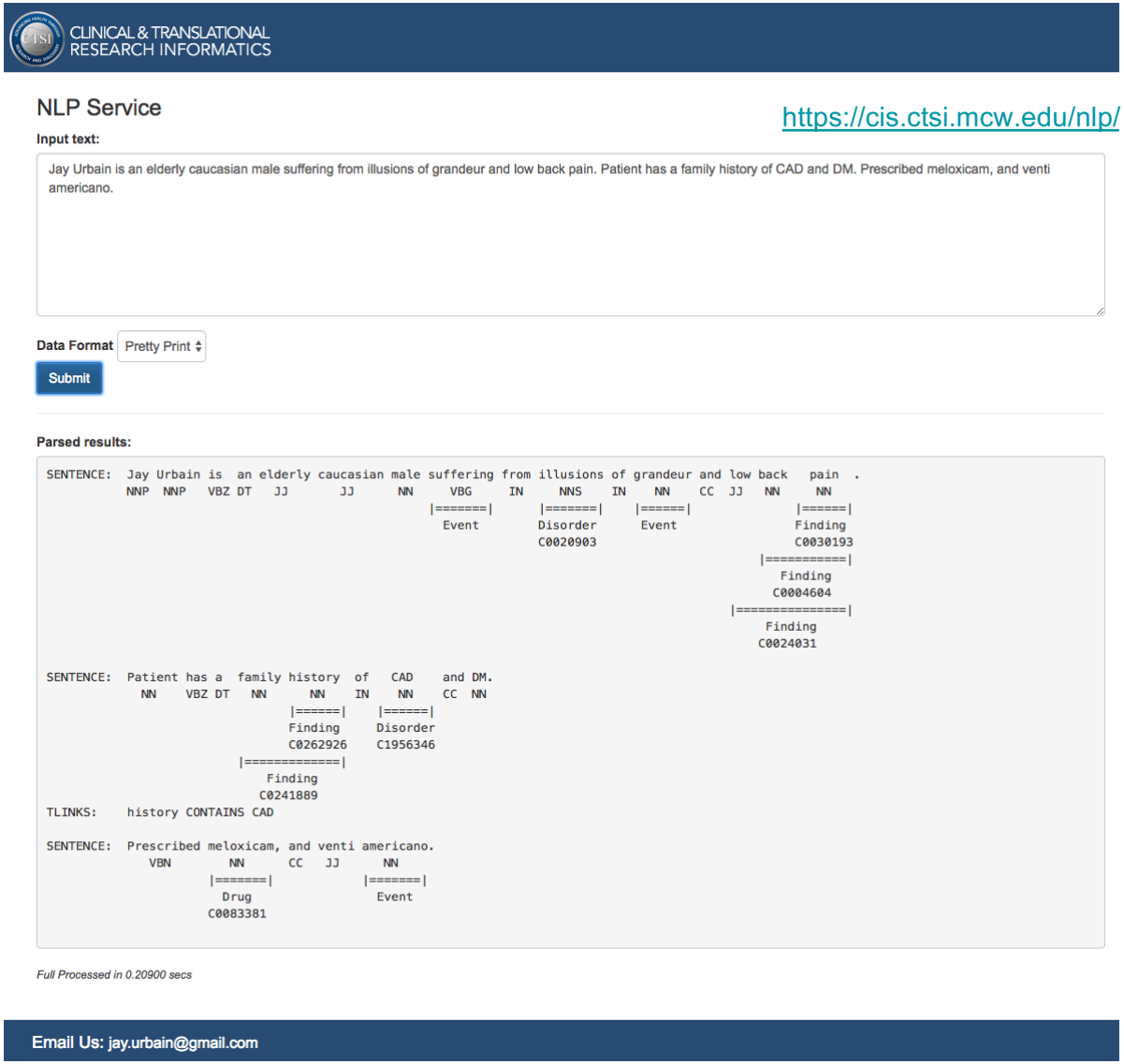
Parsed results:

[PERSON] [PERSON], [xxx@xxx.xxx] , born [12_16_2156] is an elderly caucasian male suffering from illusions of grandeur and LBP. He is married to [PERSON] [PERSON], who is much better looking. Patient father, [PERSON] [PERSON] has a history of CAD and DM. [PERSON] has been prescribed meloxicam, and venti americano. He lives at [xxxxx x. xxxxx] Dr., Disturbia, WI with his wife and golden retriever [PERSON]. You can reach him at [xxx_xxx_xxxx] .

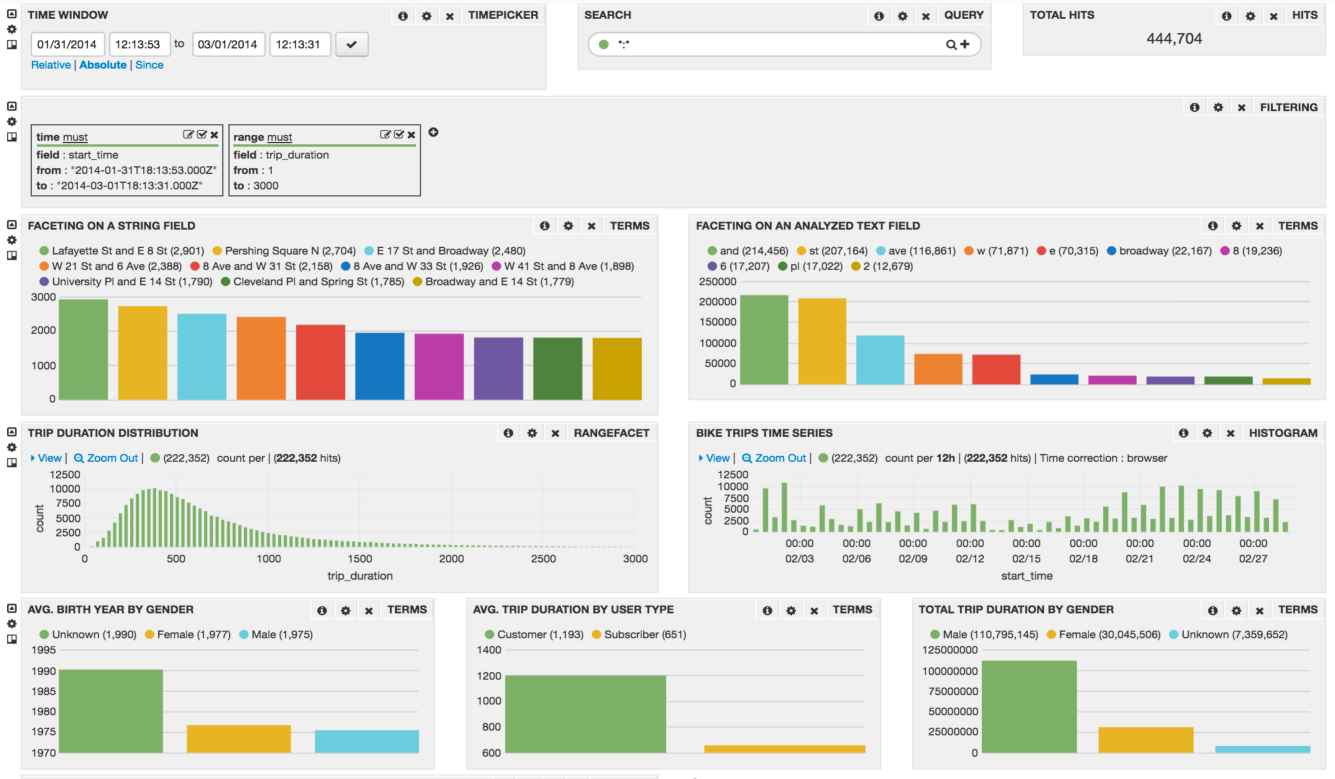
Processed in 0.24000 secs

Email Us: jay.urbain@gmail.com

Clinical Named Entity Identification



Search driven discovery: aggregating statistics from named entities



Prologue – Medical Record Machine “Translation”

- work with Bradley Crotty, MD

COLONOSCOPY Procedure Note. Date of Procedure: [1_21_2016]. Primary Physician: [PERSON] [PERSON] [PERSON], MD. Attending Physician: [PERSON] [PERSON], MD. Fellow:None. Indications: Colon cancer screening for family history of colon cancer Colon cancer screening for family history of polyps.. Previous COLONOSCOPY: Yes. Date: [8_16_2007]. Medications Administered: Agents given by the anesthesia service during MAC. Procedure Details: The patient was placed in the left lateral position and monitored continuously with ECG tracing, pulse oximetry monitoring and direct observations. Medications were administered incrementally over the course of the procedure to achieve an adequate level of moderate sedation. After anorectal examination was performed, the Olympus CF H180 was inserted into the rectum and advanced under direct vision to the terminal ileum. The procedure was considered not difficult. During withdrawal examination, the final quality of the prep was good.. Bowel Prep Scale Right Colon: Grade 2 (minor amount of residual staining, small fragments of stool, and/or opaque liquid, but mucosa of colon segment is seen well) . Bowel Prep Scale for Transverse Colon: Grade 3 (entire mucosa of colon segment seen well, with no residual staining, small fragments of stool, or opaque liquid) . Bowel Prep Scale for Left Colon: Grade 2 (minor amount of residual staining, small fragments of stool, and/or opaque liquid, but mucosa of colon segment is seen well) Additional rinsing and suctioning (600 ml) were necessary to obtain adequate views. A careful inspection was made as the colonoscope was withdrawn, which did include a retroflexed evaluation of the rectum. Findings and interventions are described below. Appropriate photo documentation was obtained. Overall [PATIENT] L [PATIENT] did tolerate the procedure well, without undue discomfort, hypotension or desaturation. At the completion of the procedure she was transferred from the endoscopy suite to be recovered in the FSC observation area per protocol and was discharged when criteria was met. After adequate recovery from sedation, she was discharged to home, with appropriate plans for follow up in place. Scope in time: 0741 Cecum reached time: 0747. Scope out time: 0759. Findings:. Findings for the Anorectal: No mass. Findings for the Terminal Ileum:

Translation

Date of procedure:

- 1_21_2016

Indications for colon cancer screening:

- Family history of polyps

Complications:

- The procedure was considered not difficult

Findings for the transverse colon:

- 1 polyp 2mm
- Tubular adenoma
- Hyperplastic polyp

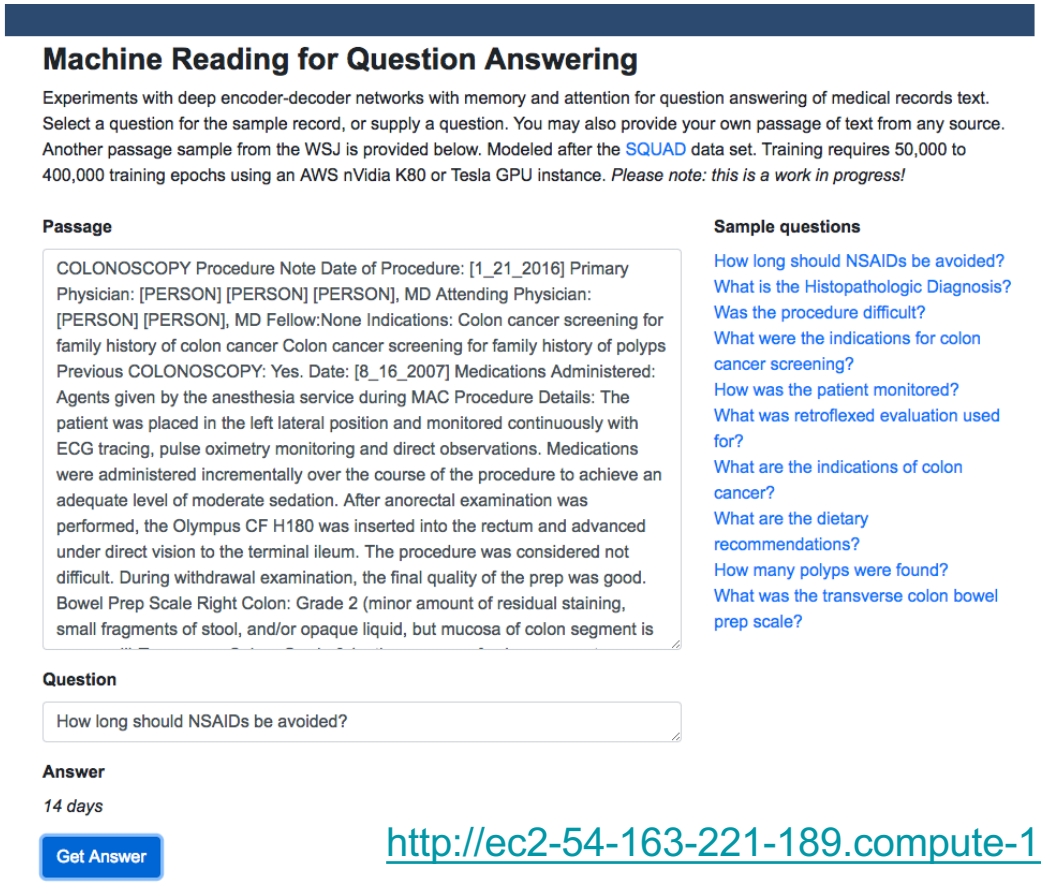
NSAIDS should be avoided:

- 14 days

Recommendations:

- High fiber diet

Question Answering



http://ec2-54-163-221-189.compute-1.amazonaws.com:8080/

Open Domain Question Answering

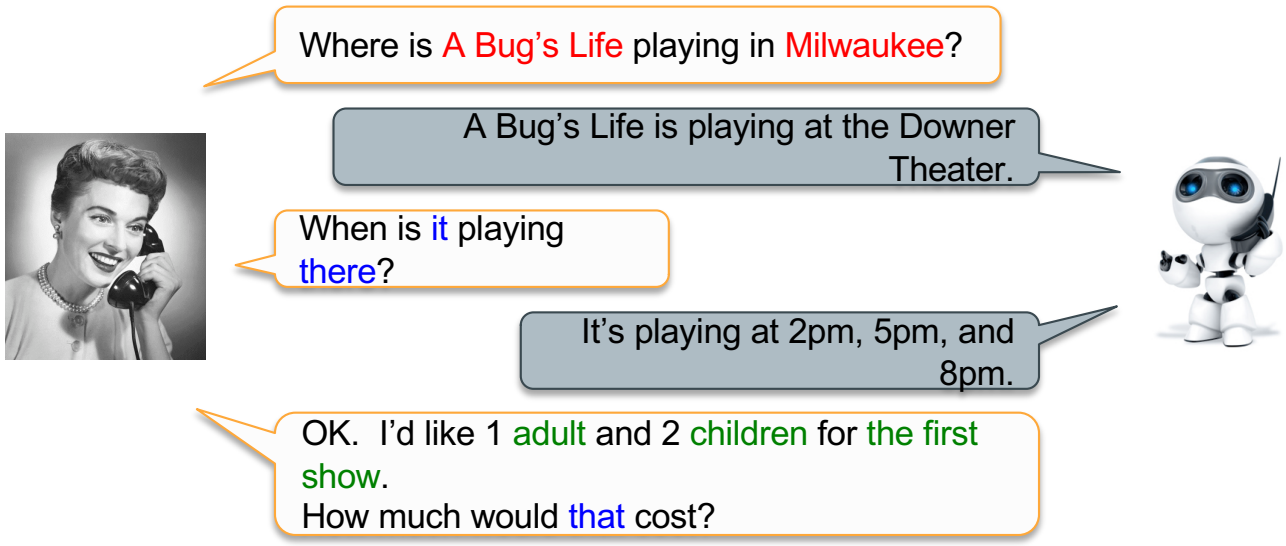
Enter Question? Submit question

Question: What causes diabetes?

Doc Source	Span	Doc score	Span score
Native American disease and epidemics	heart disease	153.33	19048.22
Gestational diabetes	insulin resistance	140.95	13148.75
Diabetes management	blood sugar levels	162.54	7199.86
A2 milk	consuming milk with A1 casein	156.44	6424.36
Diabetes mellitus	high blood sugar	166.46	6396.39

WHY IS NLP HARD?
… and why Deep Learning may help

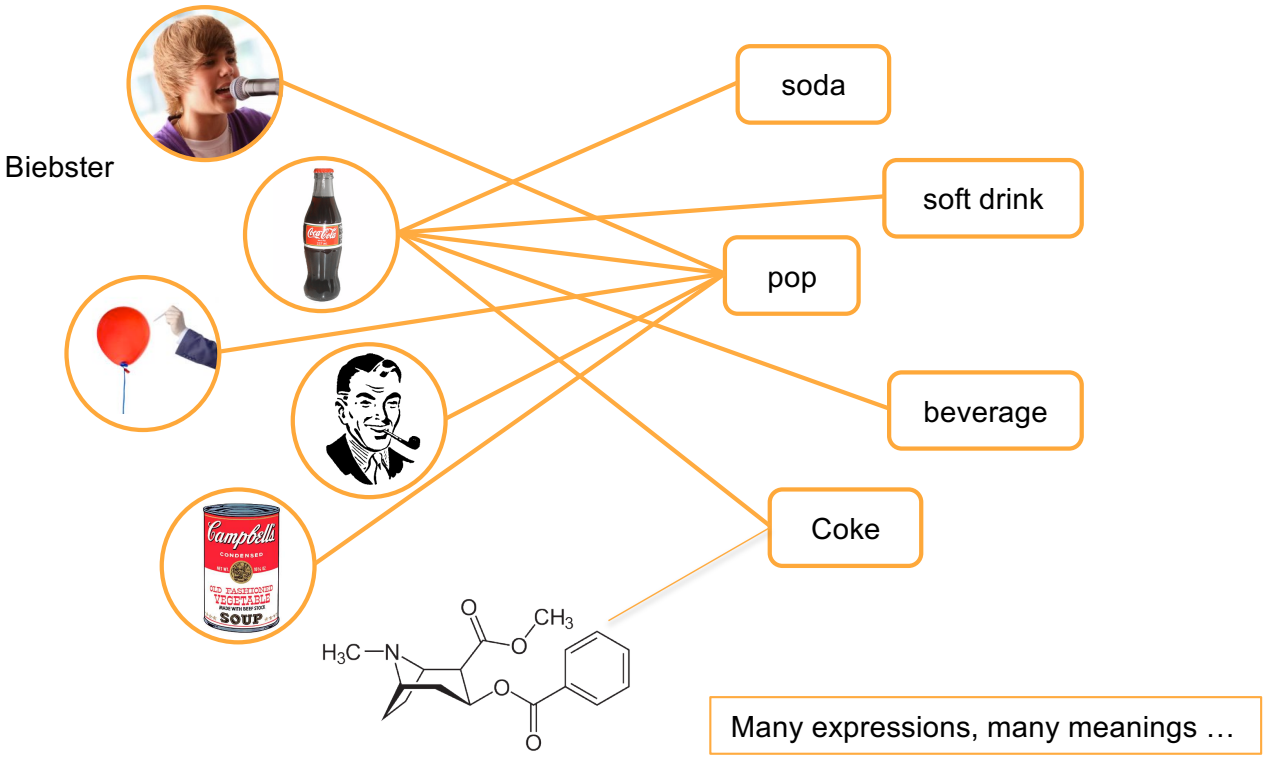
Language: the ultimate UI



But we need domain knowledge, discourse knowledge, world knowledge

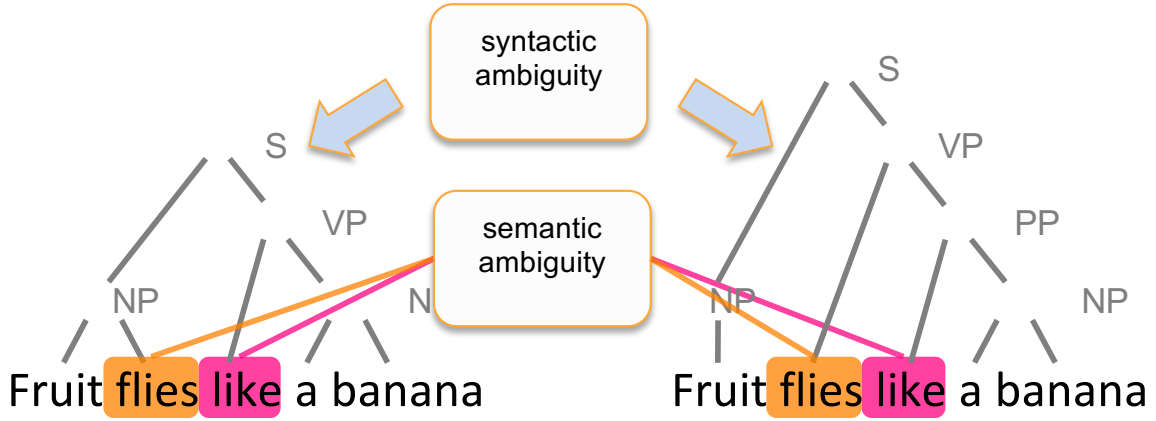
Meanings and expressions

Big obstacle: relation between meanings and expressions is not one-to-one



Many expressions, many meanings …

Syntactic & semantic ambiguity

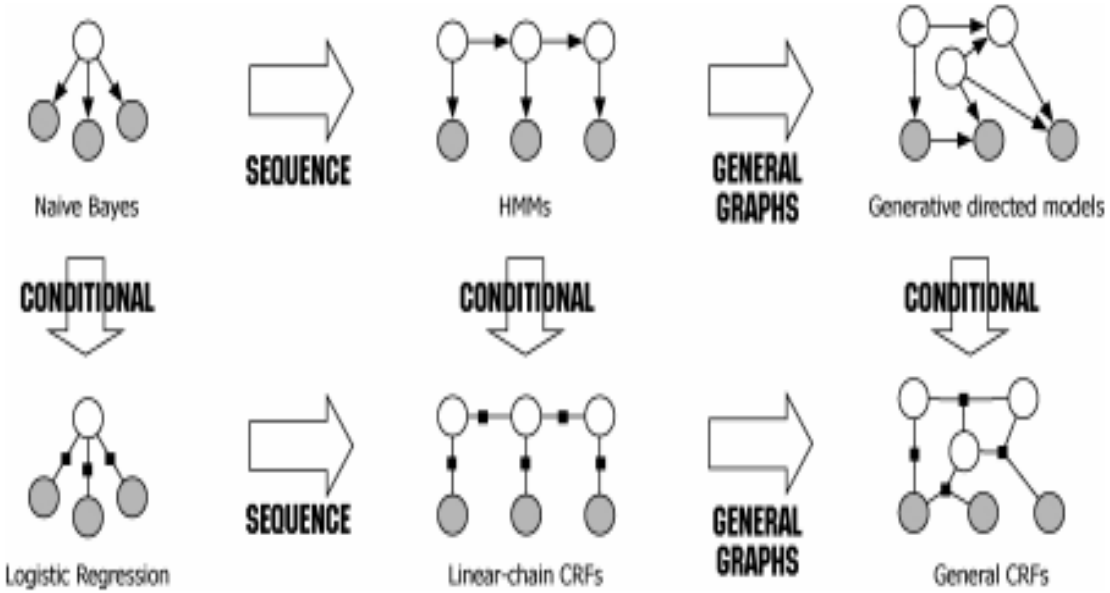


["like" is a remarkable word: it can be used as a noun, verb, adverb, adjective, preposition, particle, conjunction, or interjection.]

photos from worth1000.com

Traditional NLP use “shallow” statistical machine learning models: HMM, MEMM, CRF

- http://homepages.inf.ed.ac.uk/csutton/publications/crftut-fnt.pdf



Traditional Models: HMM, MEMM, CRF

- http://homepages.inf.ed.ac.uk/csutton/publications/crftut-fnt.pdf
- Linear chain CRF for sequence class, Named entity CRF features

Definition 2.2. Let Y, X be random vectors, $\theta = \{\theta_k\} \in \Re^K$ be a parameter vector, and $\mathcal{F} = \{f_k(y, y', \mathbf{x}_t)\}_{k=1}^K$ be a set of real-valued feature functions. Then a *linear-chain conditional random field* is a distribution $p(\mathbf{y}|\mathbf{x})$ that takes the form:

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{t=1}^{T} \exp\left\{ \sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right\}, \qquad (2.18)$$

where $Z(\mathbf{x})$ is an input-dependent normalization function

$$Z(\mathbf{x}) = \sum_{\mathbf{y}} \prod_{t=1}^{T} \exp\left\{ \sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right\}. \qquad (2.19)$$

W=v	$w_t = v$	$\forall v \in \mathcal{V}$
T=j	part-of-speech tag for w_t is j (as determined by an automatic tagger)	$\forall$POS tags j
P=I−j	w_t is part of a phrase with syntactic type j (as determined by an automatic chunker)	
Capitalized	w_t matches `[A-Z][a-z]+`	
Allcaps	w_t matches `[A-Z][A-Z]+`	
EndsInDot	w_t matches `[^\.]+.*\.`	
	w_t contains a dash	
	w_t matches `[A-Z]+[a-z]+[A-Z]+[a-z]`	
Acro	w_t matches `[A-Z][A-Z\\.]*\\.[A-Z\\.]*`	
Stopword	w_t appears in a hand-built list of stop words	
CountryCapital	w_t appears in list of capitals of countries	
⋮	many other lexicons and regular expressions	
$q_k(\mathbf{x}, t+\delta)$ for all k and $\delta \in [-1, 1]$		

Custom solutions for each domain, and for each application.
Difficulty capturing long-range dependencies, context, and semantics

NLP: With deep Learning, reality is getting closer to vision

NLP VOCABULARY

VOCABULARY

- *token* - A unit of text, typically a word. Could also be:
  - Group of words like "New York"
  - Sub-word like "mega" in "megabyte"
  - Letter like "m"
- *tokenizing* - Creating tokens, representing tokens as a distinct number
- *document* - Sequence of *tokens*
  - Could be whole book or a tweet
  - Classify each *document* as having a positive or a negative sentiment
- *corpus* - A set of *documents*
  - We'll use a *corpus* containing movie reviews since each *document* is paired with a rating indicating *sentiment*

VOCABULARY

- *n-grams* - adjoining sequence of *n* tokens (words)
  - Sometimes referred to as shingles
- *sentiment analysis* - evaluate mood of n-grams
  - Determine the attitude / emotion of text
- *tf-idf* = term frequency * inverse document frequency
  - Numerical statistic intended to reflect significance of a word to a document in a corpus
  - Popular term-weighting schemes
  - TF increases each time a word appears in a document
  - IDF decreases each time a word appears in a corpus

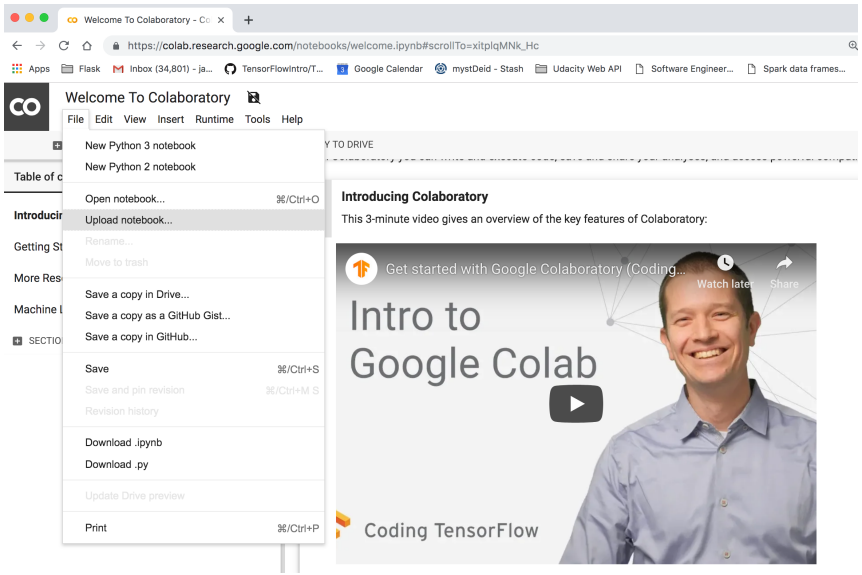
Notebook -
Introduction to NLP using Spacy

https://github.com/jayurbain/DeepNLPIntro2019/blob/master/notebooks/spaCy/spaCy-Intro.ipynb

Running the NLP workshop labs

Option 1 (recommended) Google Colab:

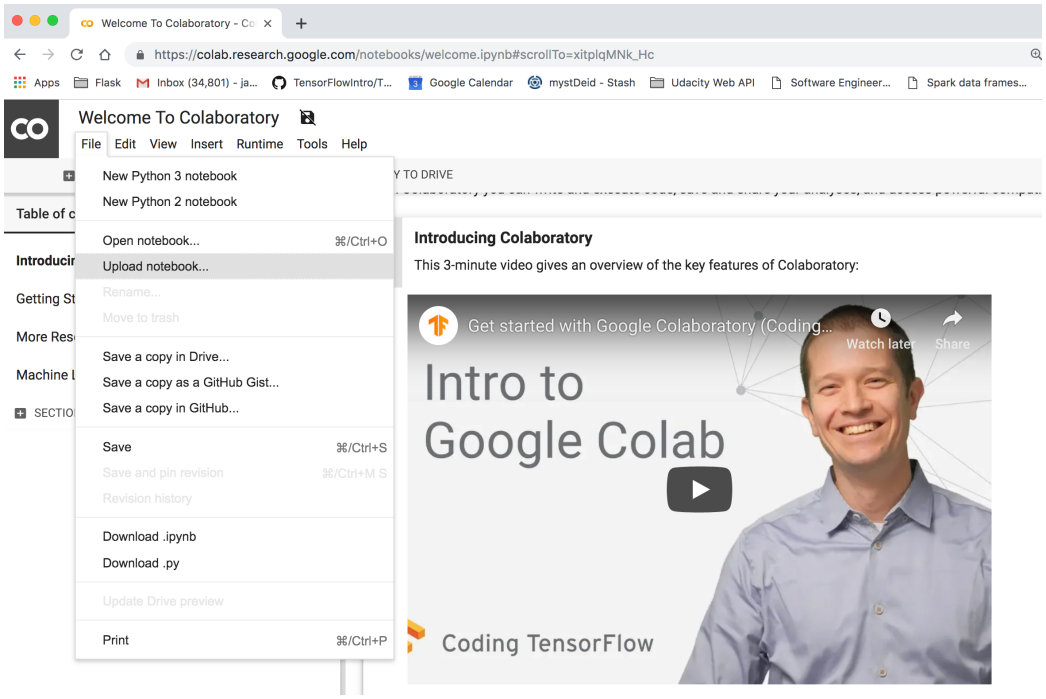
- Navigate to Google Colab:
- https://colab.research.google.com/notebooks/welcome.ipynb



Running the NLP workshop labs

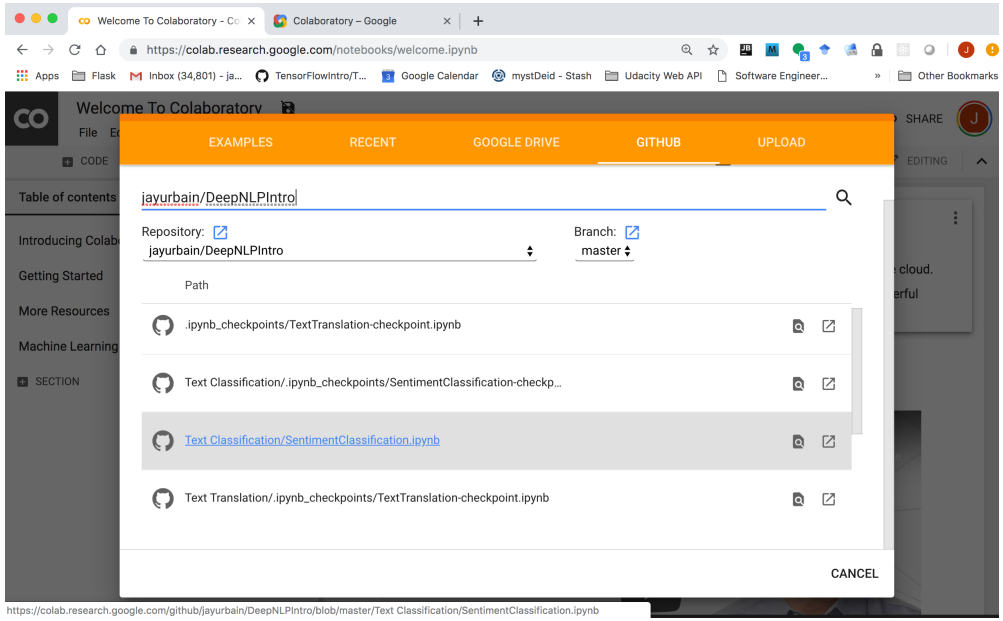
Google Colab continued:

- Select File -> Upload notebook...



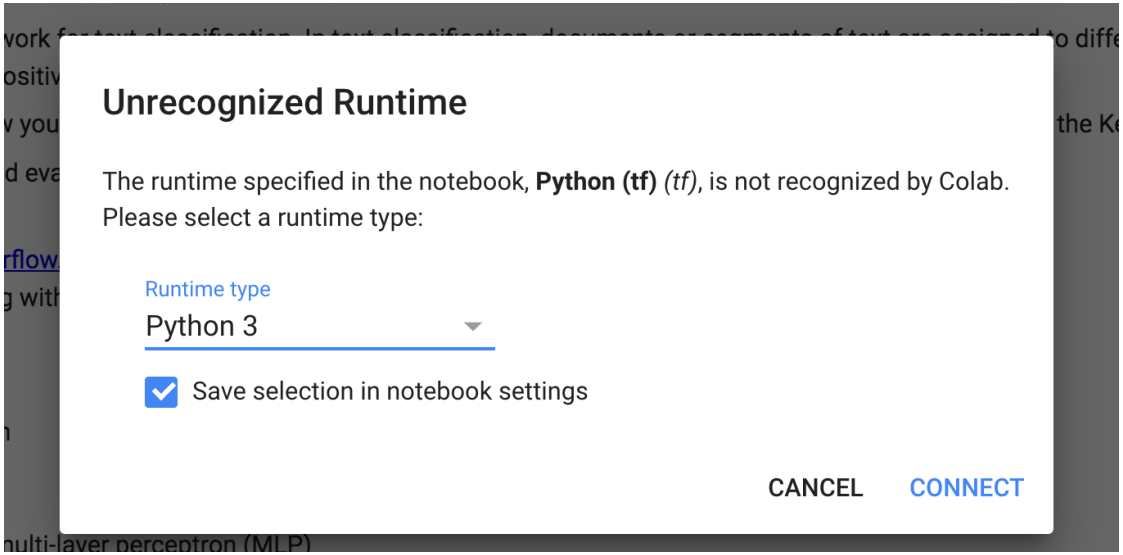
Running the NLP workshop labs

- Select GITHUB.
- Enter jayurbain/DeepNLPIntro2019
- Select the appropriate notebook for lab: ***notebooks/spaCy/spaCy-Intro.ipynb***
-



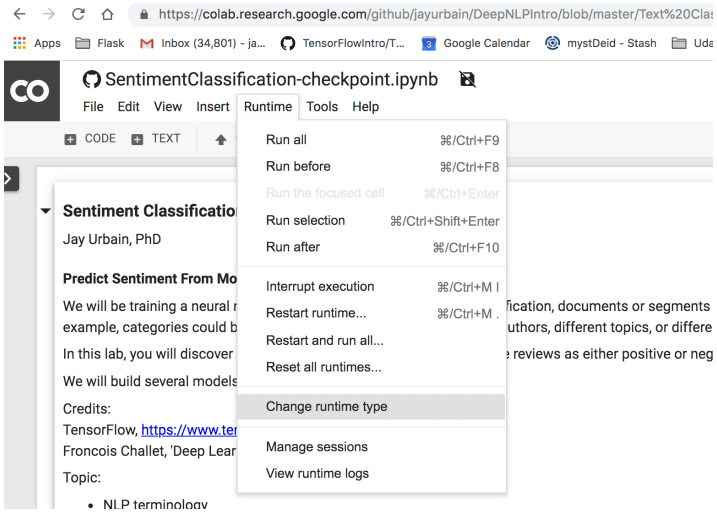
Running the NLP workshop labs

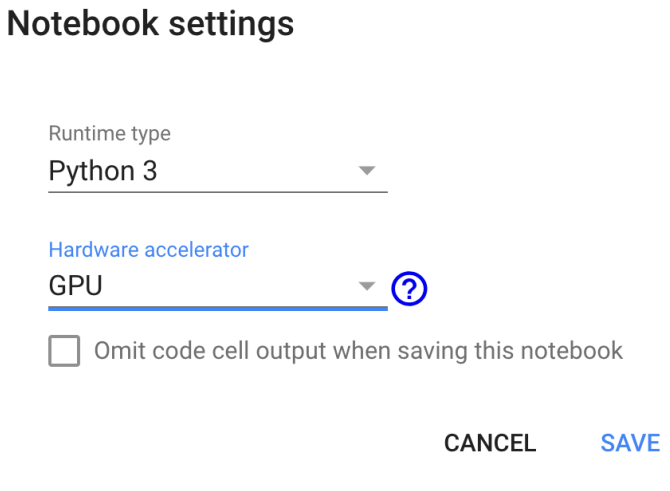
- Select the Python 3 runtime environment.
- Select Connect.



Running the NLP workshop labs

- Optional: Select Runtime -> Change runtime type -> Hardware accelerator -> GPU, Save!





Running the NLP workshop labs

Option 2: Local repository

1) git clone or download zip file

https://github.com/jayurbain/DeepNLPIntro2019

2) Create an environment.

conda

$ conda create –n deepnlp2019

$ source activate deepnlp2019

OR virtualenv

$ python -m venv ~/envs/ deepnlp2019

$ source ~/envs/deepnlp2019/bin/activate

3) cd DeepNLPIntro2019

4) pip install –r requirements.txt