Analyzing Text Data

Susan B. Davidson and Zachary G. Ives

University of Pennsylvania

CIS 545 – Big Data Analytics

Portions of this lecture have been contributed to the OpenDS4All project, piloted by Penn, IBM, and the Linux Foundation

https://tinyurl.com/cis545-lecture-01-31-22

Most Data is "Unstructured Text"

- Web pages, Wikipedia article narratives, …
- Email, SMS, Twitter, Facebook, …
    - Captions in / comments on videos
- Newswire, blogs, …
- Product descriptions and reviews, …

Often in many languages!

https://tinyurl.com/cis545-lecture-01-31-22

Text Is Much More Accessible Today

- Giant training sets
- Deep learning
- Word embedding models
- More scalable processing platforms

… Have all led to better text analysis, understanding, etc.

https://tinyurl.com/cis545-lecture-01-31-22

Roadmap – Overview of Some Techniques

- Words give insights, e.g., sentiment analysis
- Information extraction from text
- Tools for natural language processing
- Entities and relationships
- Challenges in NLP

https://tinyurl.com/cis545-lecture-01-31-22

Sentiment Analysis

Susan B. Davidson and Zachary G. Ives

University of Pennsylvania

CIS 545 – Big Data Analytics

Penn Engineering

ODPi OpenDS4All

Portions of this lecture have been contributed to the OpenDS4All project, piloted by Penn, IBM, and the Linux Foundation

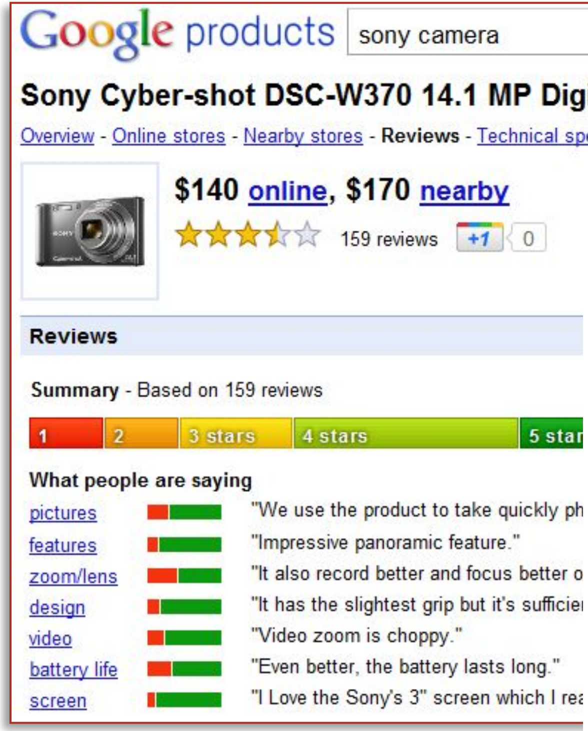
https://tinyurl.com/cis545-lecture-01-31-22

A Common "Big Data" Task

- Take products, services, companies, movies, celebrities, etc.
- Try to predict how well received they are
  - Tweets, reviews, etc.
  - Or if one needs to do extra PR / advertising / etc.
- Doesn't necessarily require we fully understand the text – we can find positive or negative *sentiment*

https://tinyurl.com/cis545-lecture-01-31-22

Sentiment Analysis



Google products | sony camera

Sony Cyber-shot DSC-W370 14.1 MP Digital Camera (Silver)

Overview - Online stores - Nearby stores - **Reviews** - Technical specifications - Similar items - Access

$140 online, $170 nearby

★★★⯪☆ 159 reviews +1 0

Reviews

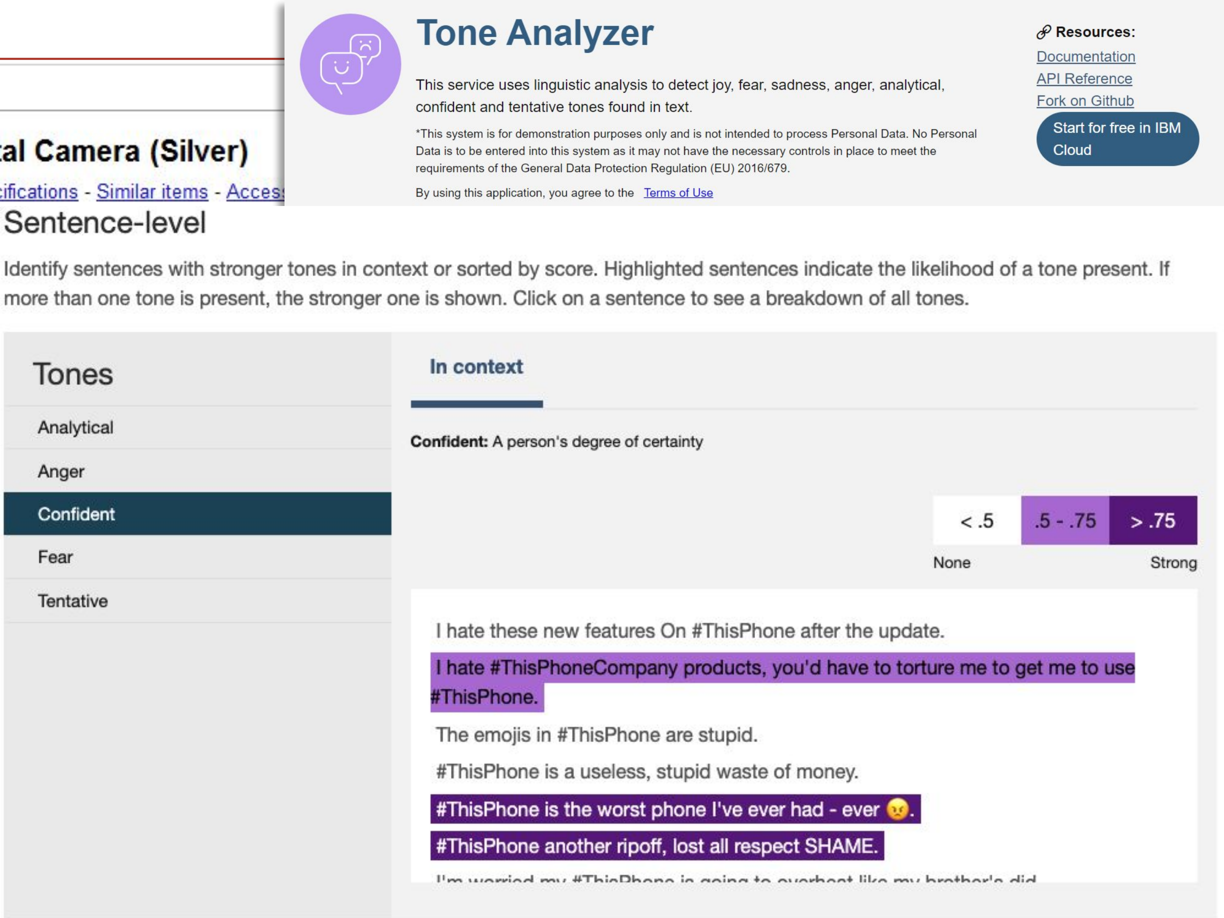
Summary - Based on 159 reviews

1 | 2 | 3 stars | 4 stars | 5 star

What people are saying

pictures	🟥🟩	"We use the product to take quickly ph
features	🟥🟩	"Impressive panoramic feature."
zoom/lens	🟥🟩	"It also record better and focus better o
design	🟥🟩	"It has the slightest grip but it's sufficie
video	🟥🟩	"Video zoom is choppy."
battery life	🟥🟩	"Even better, the battery lasts long."
screen	🟥🟩	"I Love the Sony's 3" screen which I rea

https://tinyurl.com/cis545-lectu



Tone Analyzer

This service uses linguistic analysis to detect joy, fear, sadness, anger, analytical, confident and tentative tones found in text.

*This system is for demonstration purposes only and is not intended to process Personal Data. No Personal Data is to be entered into this system as it may not have the necessary controls in place to meet the requirements of the General Data Protection Regulation (EU) 2016/679.

By using this application, you agree to the Terms of Use

Resources:

Documentation

API Reference

Fork on Github

Start for free in IBM Cloud

Sentence-level

Identify sentences with stronger tones in context or sorted by score. Highlighted sentences indicate the likelihood of a tone present. If more than one tone is present, the stronger one is shown. Click on a sentence to see a breakdown of all tones.

Tones

- Analytical
- Anger
- Confident
- Fear
- Tentative

In context

Confident: A person's degree of certainty

< .5 | .5 - .75 | > .75

None — Strong

I hate these new features On #ThisPhone after the update.

I hate #ThisPhoneCompany products, you'd have to torture me to get me to use #ThisPhone.

The emojis in #ThisPhone are stupid.

#ThisPhone is a useless, stupid waste of money.

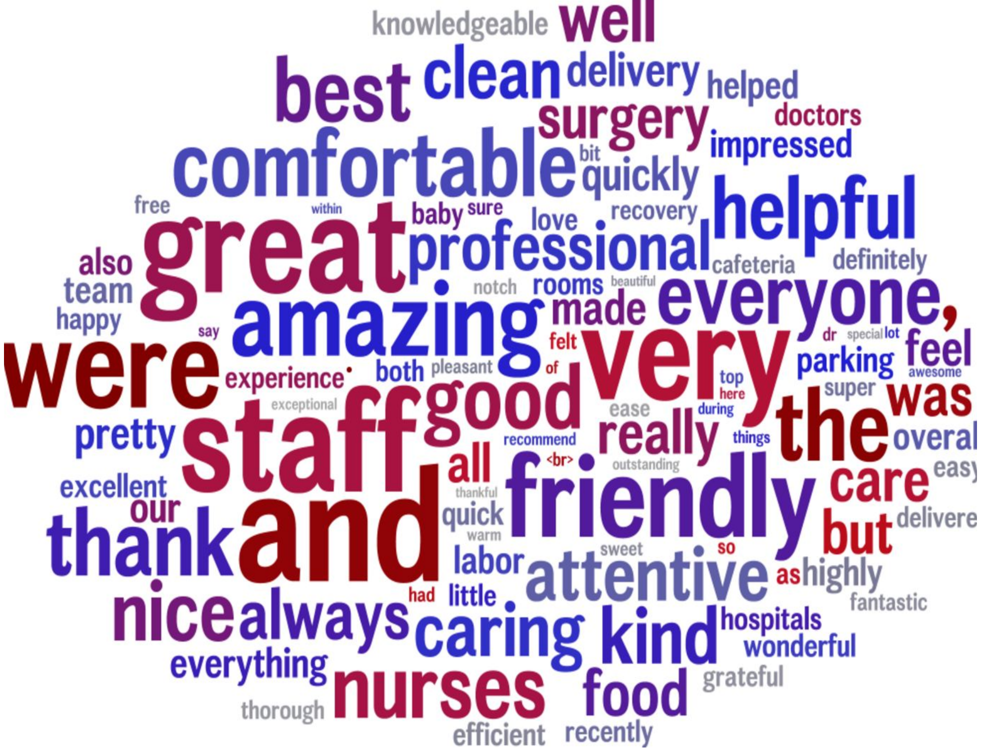
#ThisPhone is the worst phone I've ever had - ever 😠.

#ThisPhone another ripoff, lost all respect SHAME.

I'm worried my #ThisPhone is going to overheat like my brother's did

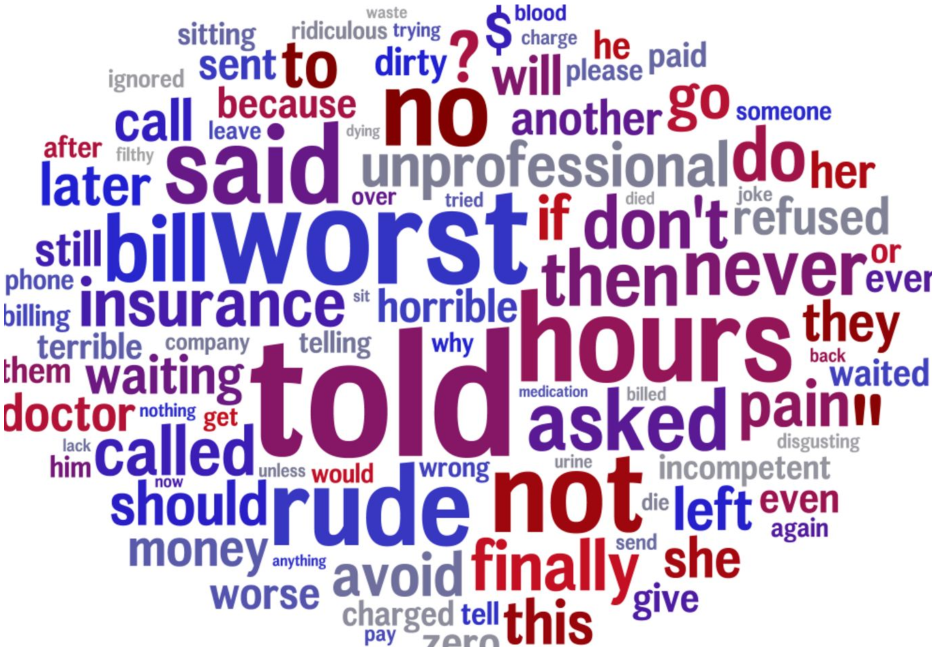
Yelp – Positive Hospital Reviews

(from Lyle Ungar)



https://tinyurl.com/cis545-lecture-01-31-22

Yelp – Negative Hospital Reviews



https://tinyurl.com/cis545-lecture-01-31-22

How Does It Work?

An example web service:
https://www.paralleldots.com/sentiment-analysis

Typically:

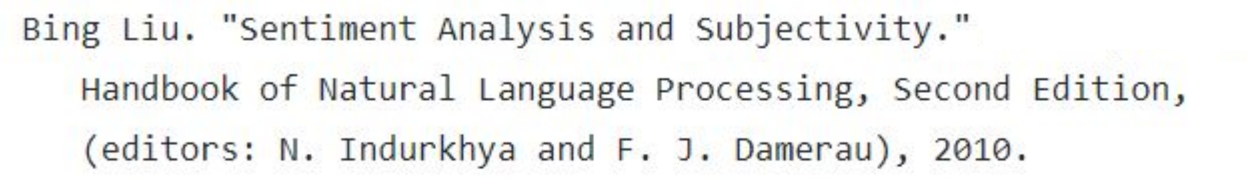
- Parse into sentences
- Via machine learning and/or rules
  - Determine if sentence is *subjective* or *objective*
  - Classify *positive*, *negative*, or *neutral*
  - Potentially identify the *subject* and the *opinion holder*
- Combine into an overall sentiment

We'll only *use* sentiment analysis tools in this course – and revisit when we get to neural nets – but CIS 530 covers how to *build* them!

https://tinyurl.com/cis545-lecture-01-31-22

Examples of Positive & Negative Words

https://github.com/jeffreybreen/twitter-sentiment-analysis-tutorial-201107/blob/master/data/opinion-lexicon-English/



Bing Liu. "Sentiment Analysis and Subjectivity."
Handbook of Natural Language Processing, Second Edition,
(editors: N. Indurkhya and F. J. Damerau), 2010.

abound
abounds
abundance
abundant
accessable
accessible
acclaim
acclaimed
acclamation
accolade
accolades

+

2-faces
abnormal
abolish
abominable
abominably
abominate
abomination
abort
aborted
aborts
abraded

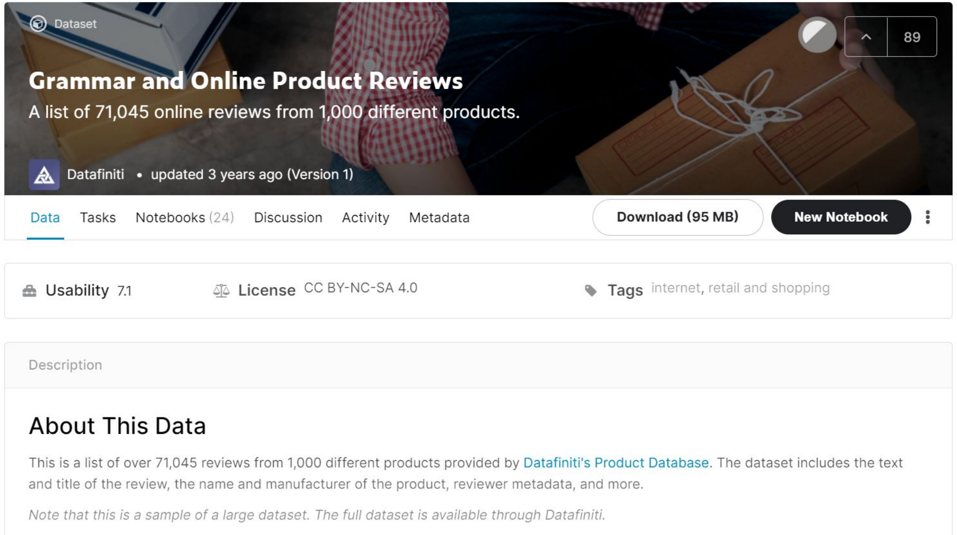
-

https://tinyurl.com/cis545-lecture-01-31-22

Using Sentiment Analysis

http://tinyurl.com/cis545-notebook-02

Suppose we want to understand how well reviewed products are…



Dataset

89

Grammar and Online Product Reviews

A list of 71,045 online reviews from 1,000 different products.

Datafiniti • updated 3 years ago (Version 1)

Usability 7.1

License CC BY-NC-SA 4.0

Tags internet, retail and shopping

Description

About This Data

This is a list of over 71,045 reviews from 1,000 different products provided by Datafiniti's Product Database. The dataset includes the text and title of the review, the name and manufacturer of the product, reviewer metadata, and more.

Note that this is a sample of a large dataset. The full dataset is available through Datafiniti.

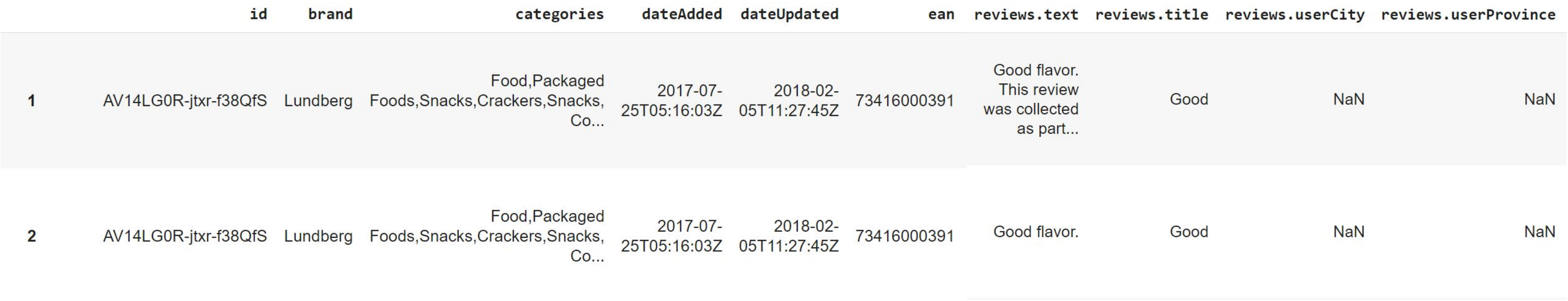
https://www.kaggle.com/datafiniti/grammar-and-e-product-reviews

https://tinyurl.com/cis545-lecture-01-31-22

Food-Product Review Sentiment

https://tinyurl.com/cis545-notebook-02

```
reviews_df = pd.read_csv( 'https://penn-cis545-files.s3.amazonaws.com/Gramm
arandProductReviews.csv' )
food_df = reviews_df[reviews_df[ 'categories' ].apply( lambda x: 'Food,' in x)
]
```



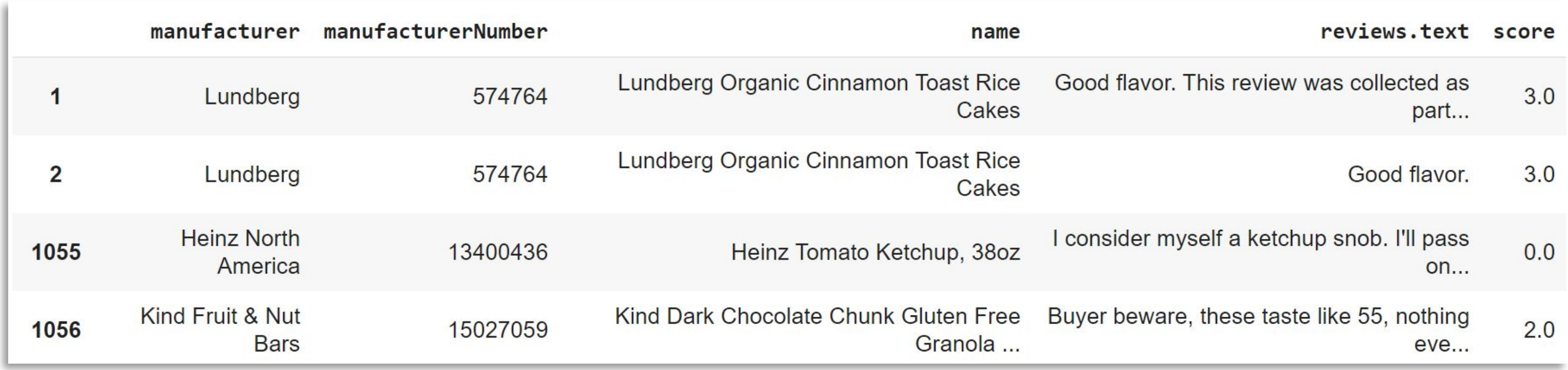
	id	brand	categories	dateAdded	dateUpdated	ean	reviews.text	reviews.title	reviews.userCity	reviews.userProvince
1	AV14LG0R-jtxr-f38QfS	Lundberg	Food,Packaged Foods,Snacks,Crackers,Snacks, Co...	2017-07-25T05:16:03Z	2018-02-05T11:27:45Z	73416000391	Good flavor. This review was collected as part...	Good	NaN	NaN
2	AV14LG0R-jtxr-f38QfS	Lundberg	Food,Packaged Foods,Snacks,Crackers,Snacks, Co...	2017-07-25T05:16:03Z	2018-02-05T11:27:45Z	73416000391	Good flavor.	Good	NaN	NaN

https://tinyurl.com/cis545-lecture-01-31-22

Let’s Use AFINN Toolkit

```
afinn = Afinn(language='en')

reviews_text_df['score'] = reviews_text_df['reviews.text'].apply(
        afinn.score)
```



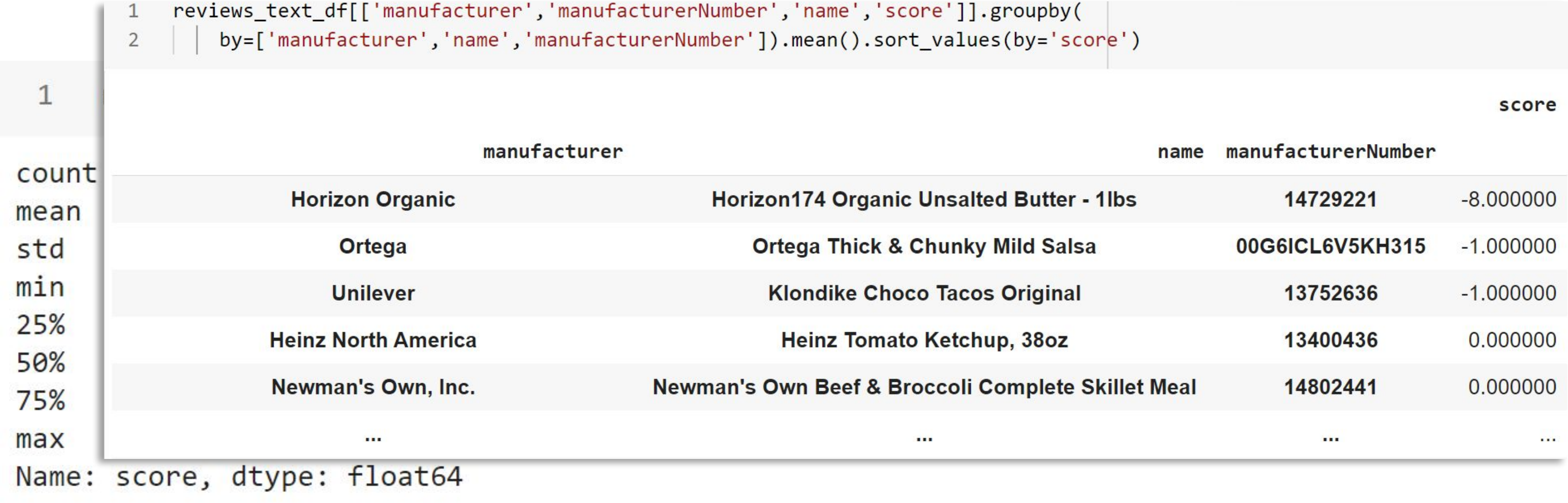
	manufacturer	manufacturerNumber	name	reviews.text	score
1	Lundberg	574764	Lundberg Organic Cinnamon Toast Rice Cakes	Good flavor. This review was collected as part...	3.0
2	Lundberg	574764	Lundberg Organic Cinnamon Toast Rice Cakes	Good flavor.	3.0
1055	Heinz North America	13400436	Heinz Tomato Ketchup, 38oz	I consider myself a ketchup snob. I'll pass on...	0.0
1056	Kind Fruit & Nut Bars	15027059	Kind Dark Chocolate Chunk Gluten Free Granola ...	Buyer beware, these taste like 55, nothing eve...	2.0

Nielsen, ESWC Workshop on “Making Sense of Microposts”, 2011

https://github.com/fnielsen/afinn

https://tinyurl.com/cis545-lecture-01-31-22

By Quality



```
1
count
mean
std
min
25%
50%
75%
max
Name: score, dtype: float64
```

```
reviews_text_df[['manufacturer','manufacturerNumber','name','score']].groupby(
    by=['manufacturer','name','manufacturerNumber']).mean().sort_values(by='score')
```

			score
manufacturer	name	manufacturerNumber	
Horizon Organic	Horizon174 Organic Unsalted Butter - 1lbs	14729221	-8.000000
Ortega	Ortega Thick & Chunky Mild Salsa	00G6ICL6V5KH315	-1.000000
Unilever	Klondike Choco Tacos Original	13752636	-1.000000
Heinz North America	Heinz Tomato Ketchup, 38oz	13400436	0.000000
Newman's Own, Inc.	Newman's Own Beef & Broccoli Complete Skillet Meal	14802441	0.000000
...	...	...	...

Note use of groupby and sort_values

https://tinyurl.com/cis545-lecture-01-31-22

Recap: Sentiment Analysis

- Typically looks for words or phrases that connote positive, neutral, or negative sentiment
  - Could be as simple as lists of words learned
- Use apply() over a column in a dataframe
- Can then use distribution of scores

https://tinyurl.com/cis545-lecture-01-31-22

Quick Review

https://canvas.upenn.edu/courses/1636888/quizzes/2771565 (05B)

- Where do we typically get positive or negative scores for word sentiment?
  a. from a master dictionary of sentiment values for the English (or other) language
  b. these are arbitrary
  c. by training machine learning algorithms on positive and negative text
  d. by flipping a coin
- Do all words have an associated sentiment?
  a. yes
  b. no
  c. only if they are adjectives
  d. only if they are verbs

https://tinyurl.com/cis545-lecture-01-31-22

Information Extraction from Text: Named Entity Extraction

Susan B. Davidson and Zachary G. Ives

University of Pennsylvania

CIS 545 – Big Data Analytics

Portions of this lecture have been contributed to the OpenDS4All project, piloted by Penn, IBM, and the Linux Foundation

https://tinyurl.com/cis545-lecture-01-31-22

Information Extraction

- We’ve seen how to do this for structured HTML: pulling dates out of Wikipedia infoboxes

- How might we do this without the cues of HTML tags?

- Need to parse and look for patterns in text
  - We’ll look at this in stages – first finding potential entities, later looking at how to assemble into relationships

https://tinyurl.com/cis545-lecture-01-31-22

Problem Formulation: Information Extraction (IE) from Text

- Input: text as sentences (or tweets etc)
- Output: data frames or equivalent
- Extract (and normalize) entities and relations between them
- Standard IE: *predefined* schema
- Open IE: entity types *not known* in advance

https://tinyurl.com/cis545-lecture-01-31-22

IE – Typical Pipeline

- Tokenize text
- Detect term boundaries
- Detect sentence boundaries
- Tag parts of speech (POS)
- Parse
- Identify named entities
- *Determine co-reference*
- *Extract entities and relations*

https://tinyurl.com/cis545-lecture-01-31-22

Sources of Raw Text
May Vary in Length, Quality, Grammar

- HTML without the template
- Tweet stream
- Voice transcriptions
- Captions on images
- PDFs
-

https://tinyurl.com/cis545-lecture-01-31-22

And potentially multiple languages!

Foundations: Natural Language Processing (NLP) Tools

NLTK

```
import nltk

nltk.download("twitter_samples")
nltk.download("punkt")
nltk.download("averaged_perceptron_tagger")
nltk.download("maxent_ne_chunker")
nltk.download("words")
nltk.download("ieer")
nltk.download("stopwords")
```

Google

- https://github.com/tensorflow/models/tree/master/research/syntaxnet

https://tinyurl.com/cis545-lecture-01-31-22

Using NLTK

- The twitter_samples corpus contains 3 files.
  - negative_tweets.json: contains 5k negative tweets
  - positive_tweets.json: contains 5k positive tweets
- tweets.20150430-223406.json: contains 20k positive and negative tweets

- Let’s see if we can use some simple text processing over some of our own data…

https://tinyurl.com/cis545-lecture-01-31-22

Let’s Analyze Some Text, Starting with Finding the Sentences

```
# Text from
paragraph
   'for lea
   'to part
   'phenome
   'human m

sentences
```

```
There were 6 sentences in the paragraph!
“Data Science” is a misnomer.
Science, in general, is a set of methods for learning
about the world.
Specific sciences are the application of these methods to
particular areas of study.
Physics is a science: it is the study of physical
phenomena.
Psychology is a science: it is the study of the psyche
(i.e., the human mind).
There is no science of data.
```

https://tinyurl.com/cis545-lecture-01-31-22

Words …

```
from nltk.tokenize import word_tokenize

fo


or
```

```
['data', 'science', 'is', 'a', 'misnomer'] ['science', 'in',
'general', 'is', 'a', 'set', 'of', 'methods', 'for',
'learning', 'about', 'the', 'world'] ['specific',
'sciences', 'are', 'the', 'application', 'of', 'these',
'methods', 'to', 'particular', 'areas', 'of', 'study']
['physics', 'is', 'a', 'science', 'it', 'is', 'the',
'study', 'of', 'physical', 'phenomena'] ['psychology', 'is',
'a', 'science', 'it', 'is', 'the', 'study', 'of', 'the',
'psyche', 'the', 'human', 'mind'] ['there', 'is', 'no',
'science', 'of', 'data']
```

https://tinyurl.com/cis545-lecture-01-31-22

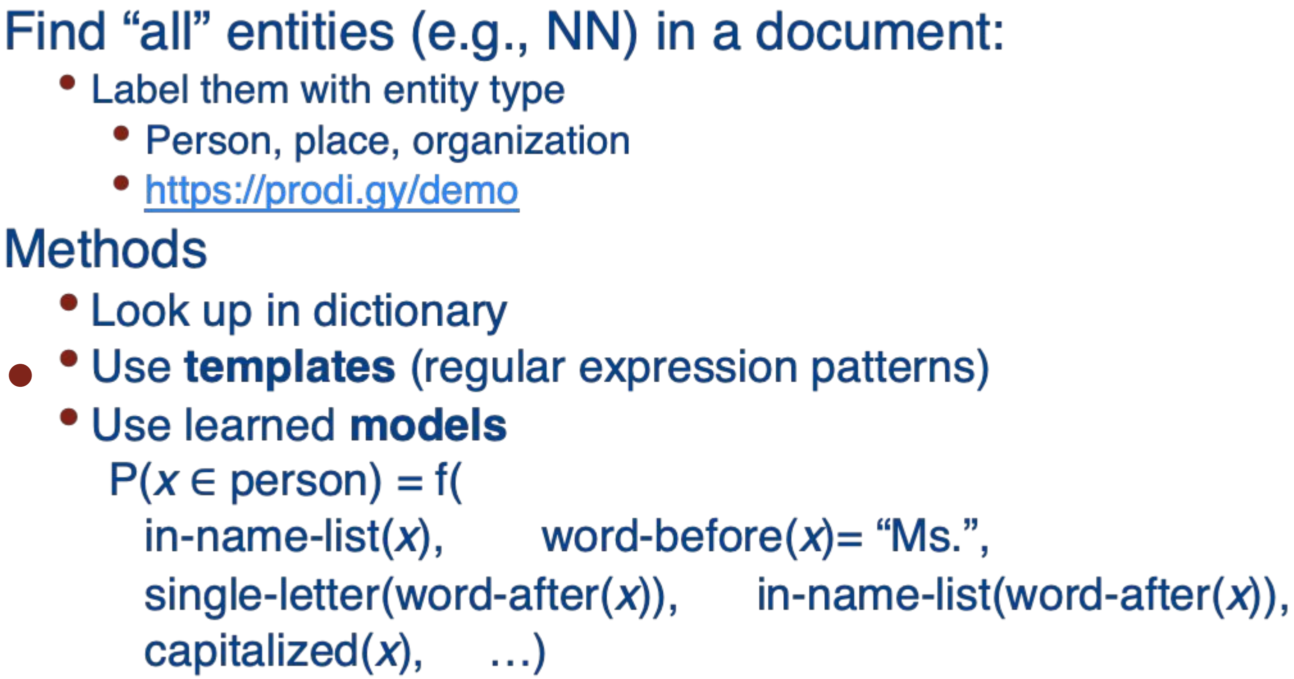
And Parts of Speech…

```
for sent in sentences:
  words = word_tokenize(sent)
  words = [word.lower() for word in words if
word.isalpha()]
  print (nltk.pos_tag(words))
```

```
[('data', 'NNS'), ('science', 'NN'), ('is', 'VBZ'), ('a',
'DT'), ('misnomer', 'NN')] [('science', 'NN'), ('in',
'IN'), ('general', 'JJ'), ('is', 'VBZ'), ('a', 'DT'),
('set', 'NN'), ('of', 'IN'), ('methods', 'NNS'), ('for',
'IN'), ('learning', 'VBG'), ('about', 'IN'), ('the', 'DT'),
('world', 'NN')] [('specific', 'JJ'), ('sciences', 'NNS'),
('are', 'VBP'),
```

https://tinyurl.com/cis545-lecture-01-31-22

Identify Named Entities



Find "all" entities (e.g., NN) in a document:

- Label them with entity type
  - Person, place, organization
  - https://prodi.gy/demo

Methods

- Look up in dictionary
- Use **templates** (regular expression patterns)
- Use learned **models**

  P(x ∈ person) = f(

  in-name-list(x), word-before(x)= "Ms.",

  single-letter(word-after(x)), in-name-list(word-after(x)),

  capitalized(x), …)

https://tinyurl.com/cis545-lecture-01-31-22

Example Named Entities

NE Type	Examples
ORGANIZATION	Georgia-Pacific Corp., WHO
PERSON	Eddy Bonte, President Obama
LOCATION	Murray River, Mount Everest
DATE	June, 2008-06-29
TIME	two fifty a m, 1:30 p.m.
MONEY	175 million Canadian Dollars, GBP 10.40
PERCENT	twenty pct, 18.75 %
FACILITY	Washington Monument, Stonehenge
GPE	South East Asia, Midlothian

http://www.nltk.org/book/ch07.html

https://tinyurl.com/cis545-lecture-01-31-22

Identify Candidates for Named Entities

```
for sent in sentences:
  words = word_tokenize(sent)
  words = [word.lower() for word in words if wo
rd.isalpha()]
  print(nltk.ne_chunk(nltk.pos_tag(words)))
```

```
(S data/NNS science/NN is/VBZ a/DT misnomer/NN)
(S science/NN in/IN general/JJ is/VBZ a/DT
set/NN of/IN methods/NNS for/IN learning/VBG
about/IN the/DT world/NN)
```

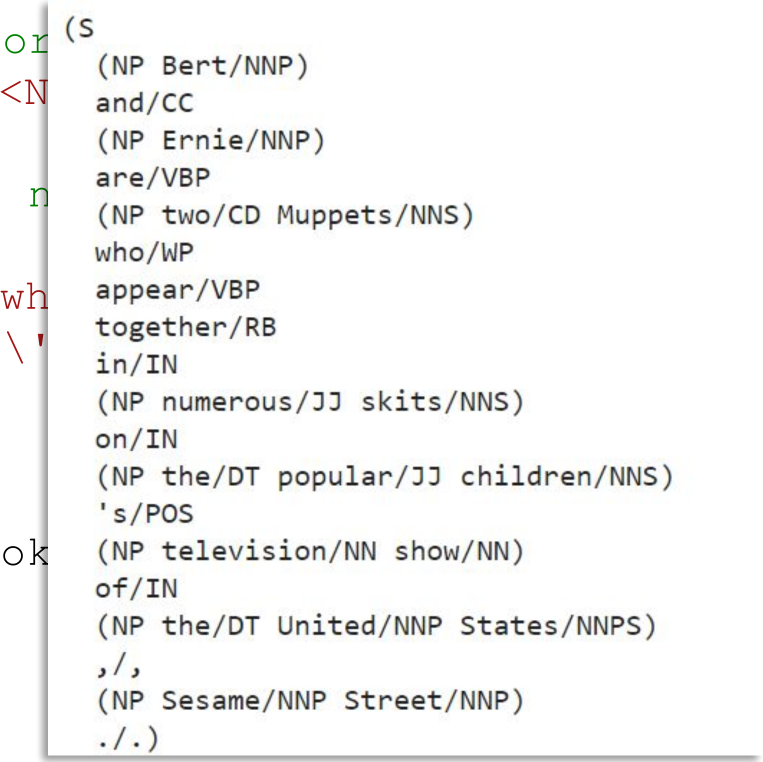
https://tinyurl.com/cis545-lecture-01-31-22

Our Own Parsing for Entities

```
# Noun phrase = optional determiner, 0 or
pattern = 'NP: {<DT>?(<CD>|<JJ>)*(<NN>|<N

# Example sentence.  BERT and ERNIE are n
# not just Sesame Street characters
sent = 'Bert and Ernie are two Muppets wh
'numerous skits on the popular children\'
'the United States, Sesame Street.'

cp = nltk.RegexpParser(pattern)
print(cp.parse(nltk.pos_tag(nltk.word_tok
```



```
(S
  (NP Bert/NNP)
  and/CC
  (NP Ernie/NNP)
  are/VBP
  (NP two/CD Muppets/NNS)
  who/WP
  appear/VBP
  together/RB
  in/IN
  (NP numerous/JJ skits/NNS)
  on/IN
  (NP the/DT popular/JJ children/NNS)
  's/POS
  (NP television/NN show/NN)
  of/IN
  (NP the/DT United/NNP States/NNPS)
  ,/,
  (NP Sesame/NNP Street/NNP)
  ./.)
```

https://tinyurl.com/cis545-lecture-01-31-22

Recap

- We saw the use of NLP software to tokenize + parse
- Named entities can be extracted via:
  - Parsing and matching by parts of speech
  - Specialized templates or rules

The next challenge: resolving entities, i.e., entity matching!

https://tinyurl.com/cis545-lecture-01-31-22

Quick Review

https://canvas.upenn.edu/courses/1636888/quizzes/2771544 (05C)

- What does a word tokenizer do?
  a. converts from English to another language
  b. breaks a paragraph into characters
  c. breaks a sentence into words and other "tokens"
  d. establishes a secure connection
- Extraction of named entities typically relies on
  a. Bert and Ernie
  b. patterns in sentence parse trees
  c. specific words only
  d. all parts of speech

https://tinyurl.com/cis545-lecture-01-31-22

Entity Resolution

Susan B. Davidson and Zachary G. Ives

University of Pennsylvania

CIS 545 – Big Data Analytics

ODPi

OpenDS4All

Portions of this lecture have been contributed to the OpenDS4All project, piloted by Penn, IBM, and the Linux Foundation

https://tinyurl.com/cis545-lecture-01-31-22

Named Entities

- So far we've seen how to extract (potential) entities from text

- How do we know when they mean the same thing

https://tinyurl.com/cis545-lecture-01-31-22

Resolving Named Entities

We can use approximate string match, but it can be ambiguous or misleading:

- "Hep A" = "Hep B"
  or "Hepatitis A"?

Context and entity type help

- "Cal" = "calories"
  or "California"
  or "Univ. of California"

https://tinyurl.com/cis545-lecture-01-31-22

Entity Resolution for Certain Kinds of Data

Brand names (companies) are relatively easy

- Need to deal with abbreviations and spelling mistakes

Product models are more complex

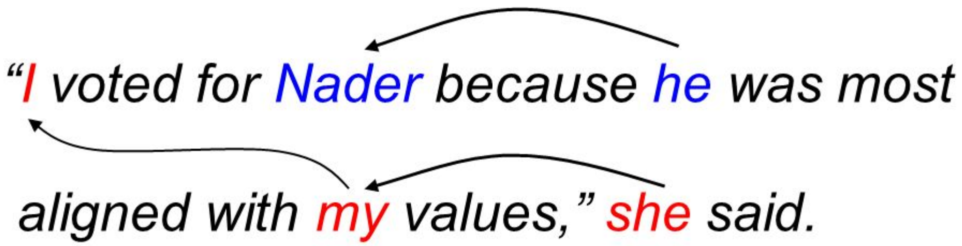
- Variations in writing styles
  - Honda Civic could be written as “Honda Civic”; “Civic”; “Honda Civic LS”; “Honda Civic LE”; “LE”; “H. Civic”; “Hondah Sivik”
  - Model numbers can be written as: 5, V, Five
  - “Asics Speedstar (both I and II), I love the I and II's and can't wait for the III's”
  - Model can be referred to as numbers but numbers do not always refer to models (e.g., “1010 for New Balance 1010”, but $1010)

City names ambiguous: Cambridge, Rochester, San Jose, Portland

Exactly the “record linking” problem we saw with our Wikipedia data wrangling example!

https://tinyurl.com/cis545-lecture-01-31-22

Coreference Resolution



"I voted for Nader because he was most aligned with my values," she said.

https://nlp.stanford.edu/projects/coref.shtml

Determining when different segments of the text are referring to the same entity

more than entity matching: pronouns, paraphrases, etc.

https://tinyurl.com/cis545-lecture-01-31-22

Coreference Resolution

Can be complicated, but relatively simple methods work OK.

Lee, Peirsman et al. 2011

- Locate all noun phrases
- Identify their properties or variations
  - singular/plural, …
- Cluster them in starting with the highest-confidence rules and moving to lower-confidence ones
  - Check first for pronominal/generic-nominal references
  - Then do closest first

https://tinyurl.com/cis545-lecture-01-31-22

Co-reference resolution example

Microsoft announced it plans to acquire Visio. The company said it will finalize its plans within a week.

Mark said that he used Symlin and it caused him to get a rash. He said that it bothered him.

https://tinyurl.com/cis545-lecture-01-31-22

Summary of Entity Resolution

- A variant of the entity matching / record linking problem, and can use many of the same techniques
- General approaches work better on some domains than others
- Coreference resolution within text is more complicated due to prepositions, paraphrases – heavily based on heuristics

https://tinyurl.com/cis545-lecture-01-31-22

Brief Review

https://canvas.upenn.edu/courses/1636888/quizzes/2771543 (05D)

- Why is approximate string match tricky to use for entity resolution?
  a. abbreviations may not match closely against full words
  b. text doesn't approximately match
  c. string similarity is only defined for structured data

- Co-reference resolution looks at whether
  a. strings are similar
  b. items are appropriately cited
  c. different words or phrases represent the same thing

https://tinyurl.com/cis545-lecture-01-31-22

Relation Extraction and Part I Wrap-up

Susan B. Davidson and Zachary G. Ives

University of Pennsylvania

CIS 545 – Big Data Analytics

Portions of this lecture have been contributed to the OpenDS4All project, piloted by Penn, IBM, and the Linux Foundation

https://tinyurl.com/cis545-lecture-01-31-22

Relation Extraction

- Ultimately we want to learn more from text than the nouns

- How do they relate, can we use these to derive new facts?

https://tinyurl.com/cis545-lecture-01-31-22

Template-based IE for relation extraction

Write or learn templates to extract entities and relations between them

- X "was acquired by" Y
- X “in” Y
- <person> , .* inventor .* of Y

Open IE = Machine Reading

- Automatically learn templates for new relationships

https://tinyurl.com/cis545-lecture-01-31-22

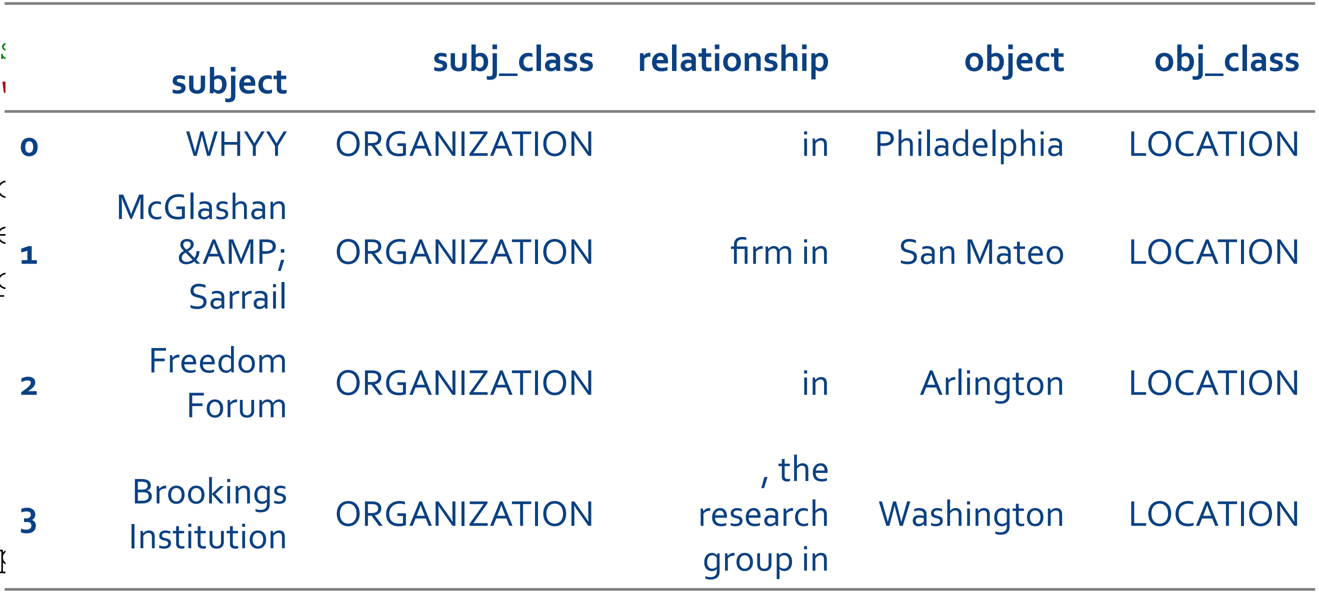
Extract Relations

http://www.nltk.org/book/ch07.html

```
# Regular expression: . means single wildcard character,
# .* means any sequence of wildcard characters, \b = blank,
#
! means negation s
IN = re.compile(r'
table = []
for doc in nltk.c
for rel in nltk.se
  corpus='ieer', p
     simple_dict =




         table.app
```



	subject	subj_class	relationship	object	obj_class
0	WHYY	ORGANIZATION	in	Philadelphia	LOCATION
1	McGlashan &AMP; Sarrail	ORGANIZATION	firm in	San Mateo	LOCATION
2	Freedom Forum	ORGANIZATION	in	Arlington	LOCATION
3	Brookings Institution	ORGANIZATION	, the research group in	Washington	LOCATION

https://tinyurl.com/cis545-lecture-01-31-22

Relation Extraction

Use templates to extract relations

- For ***Acquisition*(*Company*, *Company*)** :
  - NP2 "was acquired by" NP1
  - NP1 "'s acquisition of" NP2
- For ***MayorOf*(*City*, *Person*)**:
  - NP ", mayor of" <city>
  - <city> "'s mayor" NP
  - <city> "mayor" NP

NP = Noun Phrase

KnowItAll (Etzioni, Cafarella et al. 2005).

Impossible to guess all the possible templates – use ML!

Relation Learning (1)

Start with seed pairs of entities

- *Avatar*, *James Cameron*
- *Star Wars, George Lucas*

Find sentences that contain those entities

- *James Cameron*'s epic motion picture, *Avatar* …
- *Star Wars* director *George Lucas*
- *James Cameron*, the director of *Avatar*
- *Avatar* director *James Cameron* thinks global-warming deniers are "boneheads"

https://tinyurl.com/cis545-lecture-01-31-22

Relation Learning (2)

- Extract repeated patterns (“templates”)
  - *X* director *Y*
- Optionally, simplify them
- Use the templates to extract relations
- Reject the “bad” templates

https://tinyurl.com/cis545-lecture-01-31-22

Sample Templates:
CEO (Company/X, Person/Y)

X ceo .* Y

former X .* ceo Y

X chairman .* ceo Y

X ' s .* ceo .* Y ,

X chief executive officer Y

Y , .* ceo of .* X ,

Y , X .* ceo

Y , .* X ' .* ceo

Y , .* ceo .* X corporation

Y , .* X ceo

Y , ceo .* X ,

Y , .* chief executive officer .* of X

Y is .* chief executive officer .* of X

https://tinyurl.com/cis545-lecture-01-31-22

More Example Templates

artist - <*NAME*> , american <profession> and comedian
artist . <music> : <*NAME*> (vocals)
band currently listening <album> by <*NAME*> see
bird (<genus> rustica) <*NAME*> (
city <*NAME*> , [population (metro area)] (metro .
country held in <capital> , <*NAME*> ,
film <*NAME*> runs <length> . it
film <*NAME*> starring : <cast> ,
film watching <*NAME*> by <cast> see related
ship : <propulsion> armament : motto : <*NAME*>
wrestler - <*NAME*> , <nationality> professional wrestler

https://tinyurl.com/cis545-lecture-01-31-22

IE is hard

Language is complex

- Synonyms and Orthonyms
  - Bush, HEK
- User-generated text is rarely grammatical
- Complex structure
  - The first time I bought your product, I tried it on my dog, who became very unhappy and almost ate my cat, who my daughter dearly loves, and then when I tried it on her, she turned blue!

https://tinyurl.com/cis545-lecture-01-31-22

Really Effective IE is Hard

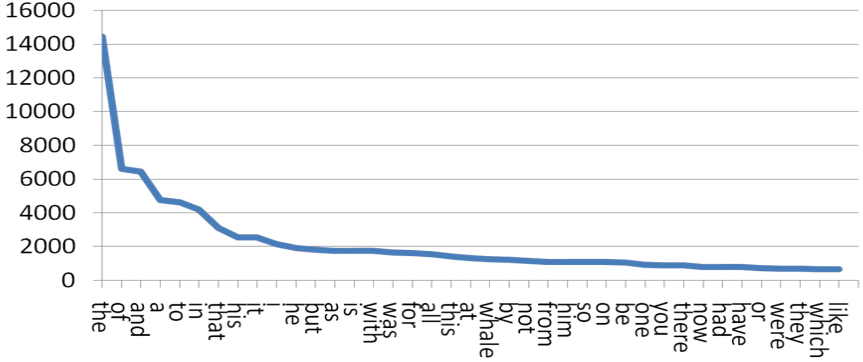
Hand-built systems give poor coverage

- Can’t manually list all patterns
- Zipf's law ensures that most words are rare

Statistical methods need training data

- Expensive to manually label data

http://searchengineland.com/the-long-tail-of-search-12198



https://tinyurl.com/cis545-lecture-01-31-22

“Adequate” IE May Be Relatively Easy

Accuracy and coverage are OK

- Typically 80% to 90% accurate
- Typically finds less than half of all mentions

Since many facts occur hundreds of times on the web, finding popular facts is easy

- Not so good if something shows up once or twice

https://tinyurl.com/cis545-lecture-01-31-22

Learning from the Web
Is Tricky

Everything on the web is NOT true

… And it’s very hard to use statistical methods to combine claims

Lots of ongoing research on copy detection, fact checking, claim provenance, …

https://tinyurl.com/cis545-lecture-01-31-22

Language-as-Data Take-aways

- Words can give insight
- Words can be features in predictive models
- Sentiment analysis and IE are useful
- IE requires named entity recognition and resolution
  - And often relation extraction
  - Tools still often give errors, and GIGO remains true

https://tinyurl.com/cis545-lecture-01-31-22

Brief Review

https://canvas.upenn.edu/courses/1636888/quizzes/2771515 (05E)

- Relation extraction depends on templates to figure out
  a. descriptions for how entities relate
  b. how to match relational tables
  c. how seed pairs relate
  d. how to make text grammatical
- Information extraction depends on what to address the issue that it has low recall (misses many mentions)?
  a. redundant information in the text
  b. fake news
  c. coregistration
  d. pronouns and antecedents

https://tinyurl.com/cis545-lecture-01-31-22

Data Science Ethics: Data Integration

Susan B. Davidson and Zachary G. Ives

University of Pennsylvania

CIS 545: Big Data Analytics

Portions of this lecture have been contributed to the OpenDS4All project, piloted by Penn, IBM, and the Linux Foundation

https://tinyurl.com/cis545-lecture-01-31-22

Privacy is not simple…

- Last time, we gave the example of researchers who publicly released data about ~70,000 OKCupid users
  - Users had consented to their data to be used only by other logged in OKCupid users
  - The researchers had not attempted to anonymize the data

- Would anonymizing (de-identifying) the dataset been enough to obfuscate the identity of the OKCupid users?
  - Masking, generalizing, or deleting both direct and indirect identifiers

https://tinyurl.com/cis545-lecture-01-31-22

De-identification is not enough

Netflix Prize Competition

- In 2006, Netflix released a de-identified data set in an open competition for the best collaborative filtering algorithm to predict user ratings for films
  - Contained information <user, movie, date_of_grade, grade>
  - Users and movies were identified by numbers assigned for the contest
- In 2010, the competition was cancelled due to privacy concerns □ **Data re-identification**
  - Researchers at UT Austin were able to link users with film ratings on the IMDB's system, where the users were identified.

https://tinyurl.com/cis545-lecture-01-31-22

De-identification is not enough

Massachusetts re-identification incident

- In the mid 1990’s, the Massachusetts Group Insurance Commissions (GIC) released de-identified health records.
  - Identifiers such as name, address and SSN were removed, however zip codes, birth date and sex were not.

- Latanya Sweeney combined the GIC data with the voter database of Cambridge, MA discovered the identity of then-Governor William Weld and located his health record.

https://tinyurl.com/cis545-lecture-01-31-22

De-identification is not enough

Massachusetts re-identification incident

- In the mid 1990's setts Group Insurance Commissions (GI dentified health records.
  - Identifiers such s and SSN were removed, however zip co d sex were not.

- Latanya Sweeney combined the GIC data with the voter database of Cambridge, MA discovered the identity of then-Governor William Weld and located his health record.

https://tinyurl.com/cis545-lecture-01-31-22

Correlating data

- Even if data is de-identified, entries can be correlated (i.e. linked) with entries in other datasets to make informed guesses as to identity

Correlating data

- Even if data is de-identified, entries can be correlated (i.e. linked) with entries in other datasets to make informed guesses as to identity

- **Problem: “Sparsity” of data**
  - In Netflix data, no two profiles are more than 50% similar.
  - If a Netflix profile is more than 50% similar to a profile in IMDB, then there is a high probability that the two profiles are of the same person

- **87% of the U.S. population can be identified using a combination of their gender, birthdate and zip code.**

A. Narayanan and V. Shmatikov, “Robust de-anonymization of large sparse datasets …,”
Proc. 29th IEEE Symp. Security and Privacy, 2008.

Individual versus statistical information

- When do you feel safe releasing personal information, e.g. completing a survey about your tastes in movies?

https://tinyurl.com/cis545-lecture-01-31-22

Individual versus statistical information

- When do you feel safe releasing personal information, e.g. completing a survey about your tastes in movies?
    - My answers have no impact on the privatized released result?
    - With high probability, an attacker looking at the privatized released result cannot learn any new information about me?
    - **These are not achievable.**

https://tinyurl.com/cis545-lecture-01-31-22

Differential Privacy

- **Differential privacy** aims to maximize the accuracy of queries from statistical databases while minimizing the chances of identifying its records – it adds noise and provides guarantees against a “privacy budget”.

Harm to individual

Benefit to society

https://tinyurl.com/cis545-lecture-01-31-22

Differential Privacy

- **Differential privacy** aims to maximize the accuracy of queries from statistical databases while minimizing the chances of identifying its records – it adds noise and provides guarantees against a "privacy budget".

- "Algorithmic Foundations of Differential Privacy," Foundations and Trends in Theoretical Computer Science (2014).
    - Penn CIS Professor Aaron Roth, and Turing Award winner Cynthia Dwork (Harvard University)

https://tinyurl.com/cis545-lecture-01-31-22

Differential Privacy

- **Differential privacy** aims to maximize the accuracy of queries from statistical databases while minimizing the chances of identifying its records – it adds noise and provides guarantees against a “privacy budget”.

- A very accessible video on the topic by Cynthia Dwork is linked to the course webpage.

https://tinyurl.com/cis545-lecture-01-31-22

Summary

- De-identification is not enough to ensure the privacy of individuals
- Only providing statistical summaries does not guarantee that no information will not be leaked about individuals

Harm to individual

Benefit to society

https://tinyurl.com/cis545-lecture-01-31-22

Brief Review

https://canvas.upenn.edu/courses/1636888/quizzes/2779755 (05F)

- Which of the following are correct (select all that apply)?

a) Privacy can be guaranteed by removing identifying information from a dataset.

b) Privacy can be guaranteed by only providing answers to statistical queries over a dataset.

c) Differential privacy aims to maximize the accuracy of queries from statistical databases while minimizing the chances of identifying its records.

https://tinyurl.com/cis545-lecture-01-31-22

Part I Wrap-up

- We’ve talked about how to extract relevant content into dataframes (relation), and how to process them
  - Project, filter, rename, apply, join, groupby, …
- A key question: how do we decide what our dataframes should look like? Next time!

https://tinyurl.com/cis545-lecture-01-31-22