GENERAL ASSEMBLY

DATA SCIENCE

11 WEEK PART TIME COURSE

Week 10 - Natural Language Processing
Monday 22nd February 2016

AGENDA

1. Guest Speaker - Dev Mookerjee
2. Tasks from last week
3. Natural Language Processing
4. Case Studies
5. Techniques for NLP
6. Lab
7. Discussion

Task List

- Download the Caret package
- Run a Random Forest Model in R
- Run Recommendation Engine code on Git on your Spark Cluster.
- Read the following articles:
  - http://www.wired.com/2016/01/googles-go-victory-is-just-a-glimpse-of-how-powerful-ai-will-be/
  - http://www.wired.com/2014/01/geoffrey-hinton-deep-learning
  - https://sites.google.com/site/deepernn/home/blog/briefsummaryofthepaneldiscussionatdlworkshopicml2015
- Install TensorFlow

Task List

- ☐ Run one of the following Tensorflow examples:
  - ☐ MNIST For ML Beginners
  - ☐ Convolutional Neural Networks
  - ☐ Vector Representations of Words
- ☐ Post a picture you generated via Google Deep Dream in Slack

☐ EMAIL ~~ONE~~ TWO OF THE GUEST PRESENTERS OF THIS CLASS

WHAT IS NATURAL LANGUAGE PROCESSING?

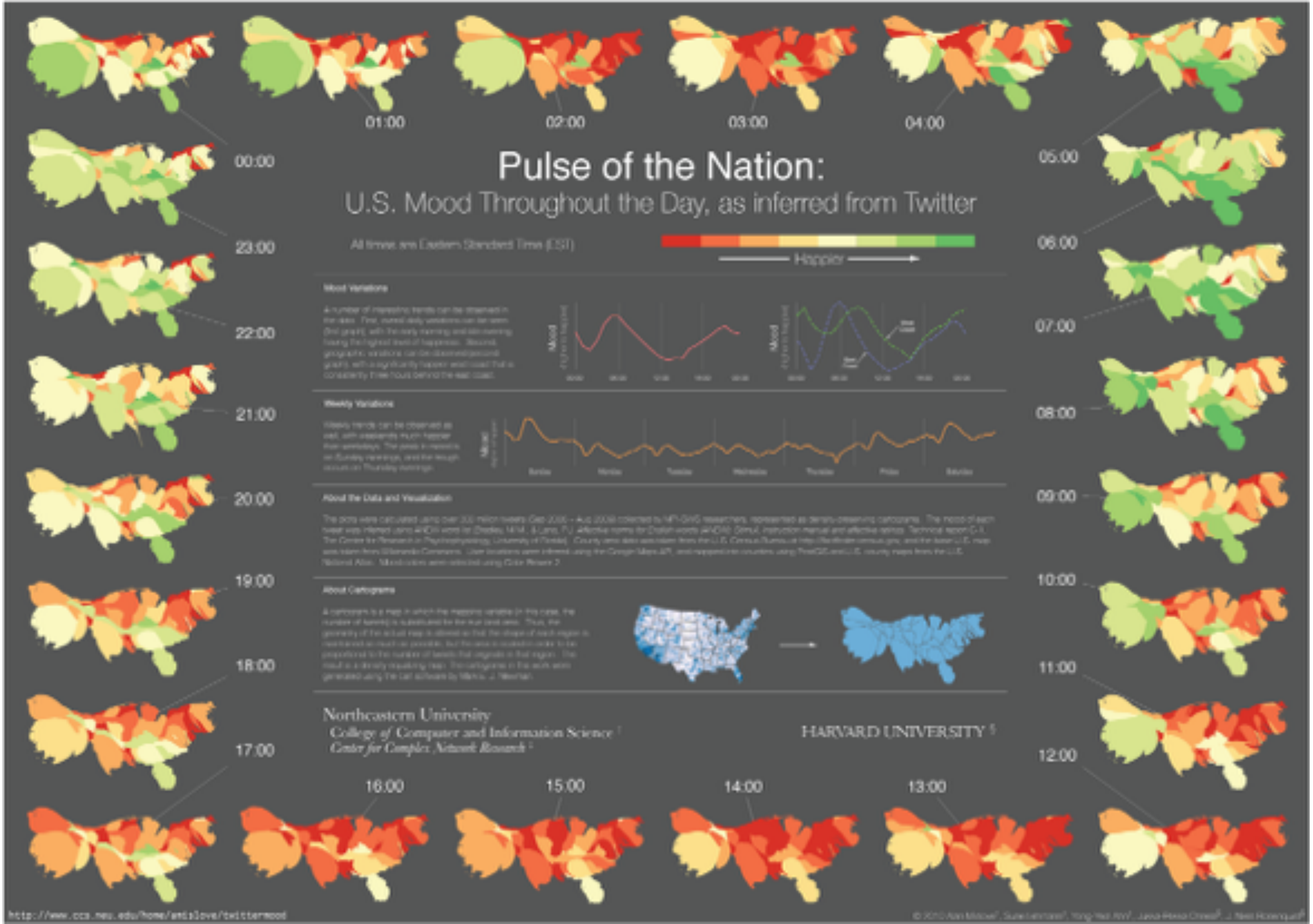
- Text is considered to be un-structured data. This means we don't have nice features we can use as inputs. We will have to construct them using a model or rules we know about language.
- Natural Language Processing is the algorithms and processing we program to interpret human language.
- It allows us to extract meaning from text as it appears in emails, articles, tweets, journal articles, books, speech, advertisements, etc in the dialect it was created in.

DATA SCIENCE PART TIME COURSE

WHY BOTHER WITH NATURAL LANGUAGE PROCESSING?

???

CASE STUDIES



Pulse of the Nation:
U.S. Mood Throughout the Day, as inferred from Twitter
All times are Eastern Standard Time (EST)
Happier
Mood Variations
Weekly Variations
About the Data and Visualization
About Cartograms
00:00
01:00
02:00
03:00
04:00
05:00
06:00
07:00
08:00
09:00
10:00
11:00
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
21:00
22:00
23:00
Northeastern University
College of Computer and Information Science
Center for Complex Network Research
HARVARD UNIVERSITY
http://www.ccs.neu.edu/home/amislove/twittermood

The 199 People, Places and Things Donald Trump Has Insulted on Twitter: A Complete List

By JASMINE C. LEE and KEVIN QUEALY UPDATED February 19, 2016 Related Article

In the seven months since declaring his candidacy for president, Donald Trump has used Twitter to lob insults at presidential candidates, journalists, news organizations, nations, a Neil Young song and even a lectern in the Oval Office. We know this because we've read, tagged and quoted them all. Below, a directory of sorts, with links to the original tweets. Insults within the last two weeks are highlighted. RELATED ARTICLE

Recently insulted: Wall Street Journal-NBC Poll, Brit Hume, The Republican National Committee, Lindsey Graham, Ted Cruz, Glenn Beck, Fox News, Megyn Kelly, Barack Obama, Jeb Bush

CURRENT AND FORMER PRESIDENTIAL CANDIDATES

Jeb Bush
FORMER FLORIDA GOVERNOR

"just got contact lenses and got rid of the glasses. He wants to look

Glenn Beck
TELEVISION PERSONALITY

"Your endorsement means nothing!", "dumb as a rock", "crying", "lost all credibility", "failing", "irrelevant", "wacko",

Frank Luntz
POLITICAL CONSULTANT

"a total clown", "a clown", "where did you find that dumb panel", "a low-class snob", "knows nothing about me or my religion", "came to

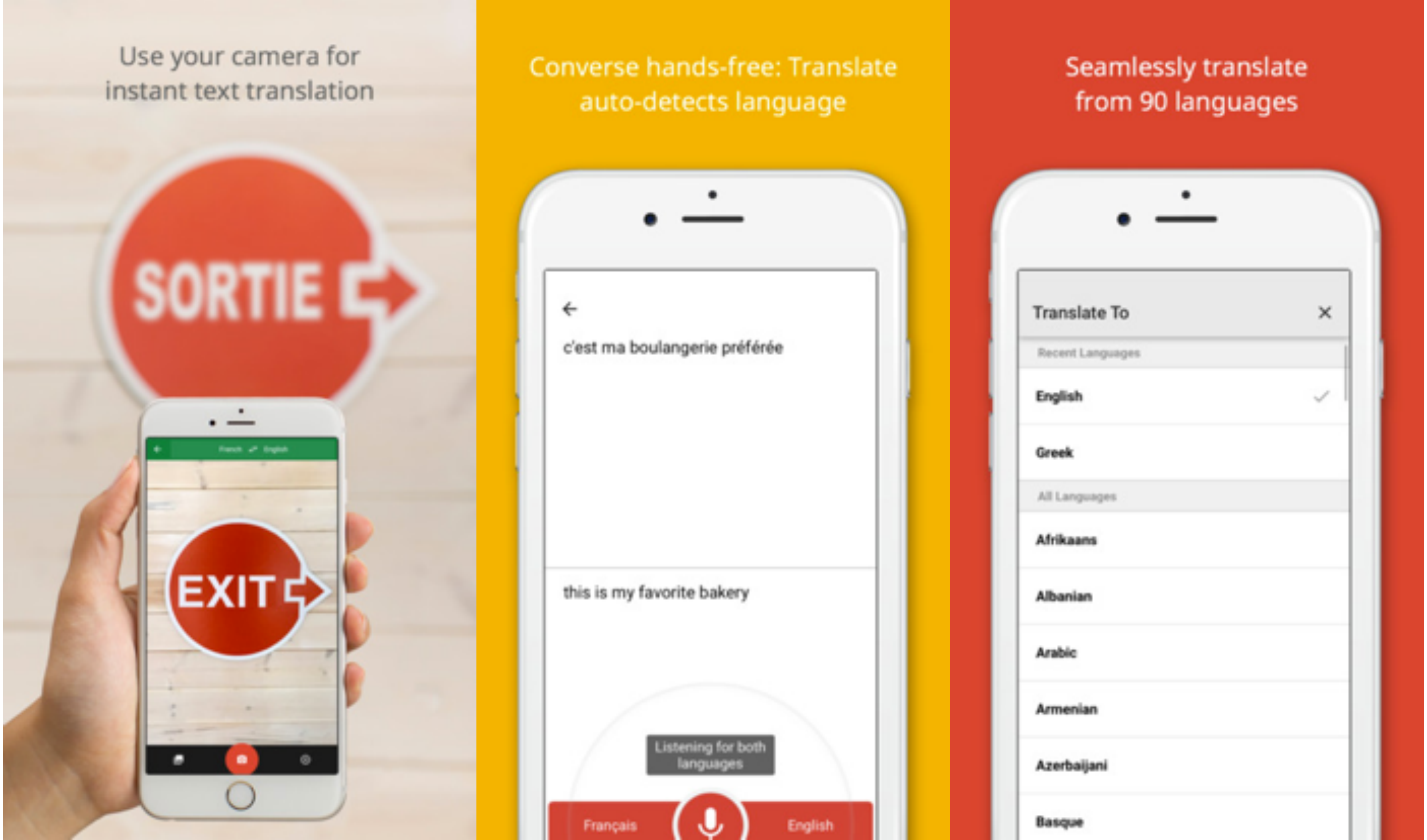
Mort Zuckerman
OWNER, THE NEW YORK DAILY NEWS

"Dopey", "has a major inferiority complex", "dopey clown"

Bill de Blasio

The New York Times
NEWSPAPER

"failing", "allows dishonest writers to totally fabricate stories", "failing", "change your false story", "boring articles", "should focus on


Use your camera for
instant text translation
SORTIE
EXIT
Converse hands-free: Translate
auto-detects language
c'est ma boulangerie préférée
this is my favorite bakery
Listening for both
languages
Français
English
Seamlessly translate
from 90 languages
Translate To
Recent Languages
English
Greek
All Languages
Afrikaans
Albanian
Arabic
Armenian
Azerbaijani
Basque

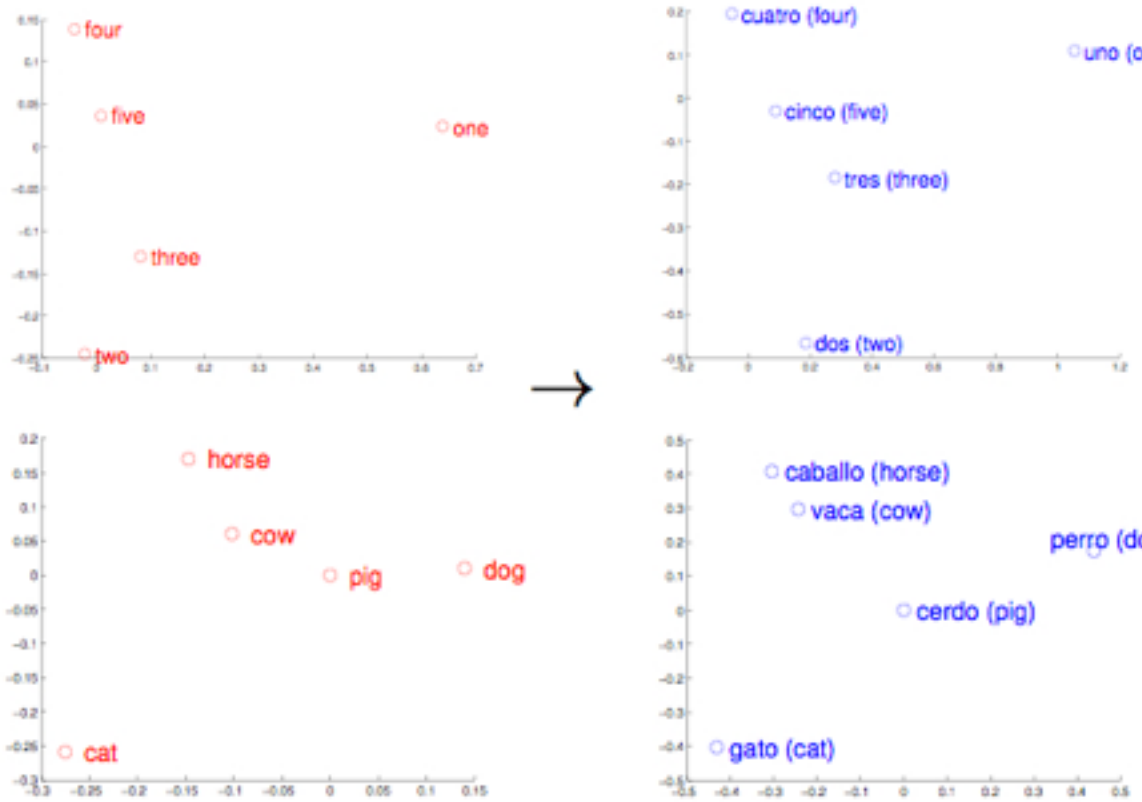
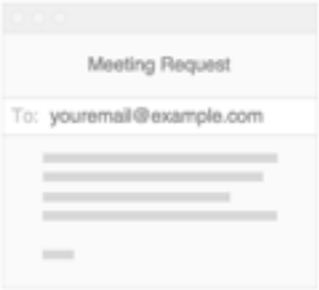
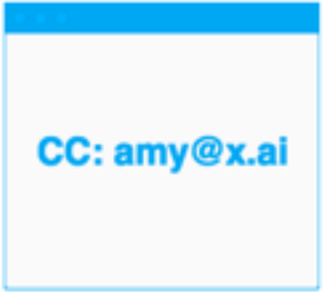


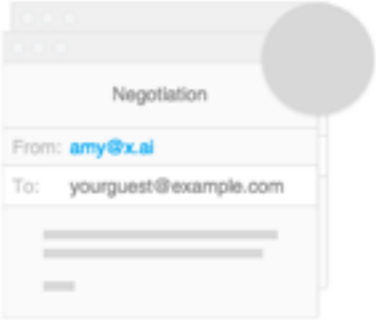
Figure 1: Distributed word vector representations of numbers and animals in English (left) and Spanish (right). The five vectors in each language were projected down to two dimensions using PCA, and then manually rotated to accentuate their similarity. It can be seen that these concepts have similar geometric arrangements in both spaces, suggesting that it is possible to learn an accurate linear mapping from one space to another. This is the key idea behind our method of translation.



You receive a meeting request, but don't want to deal with the back and forth to get it scheduled



You Cc: Amy, handing the job over to her

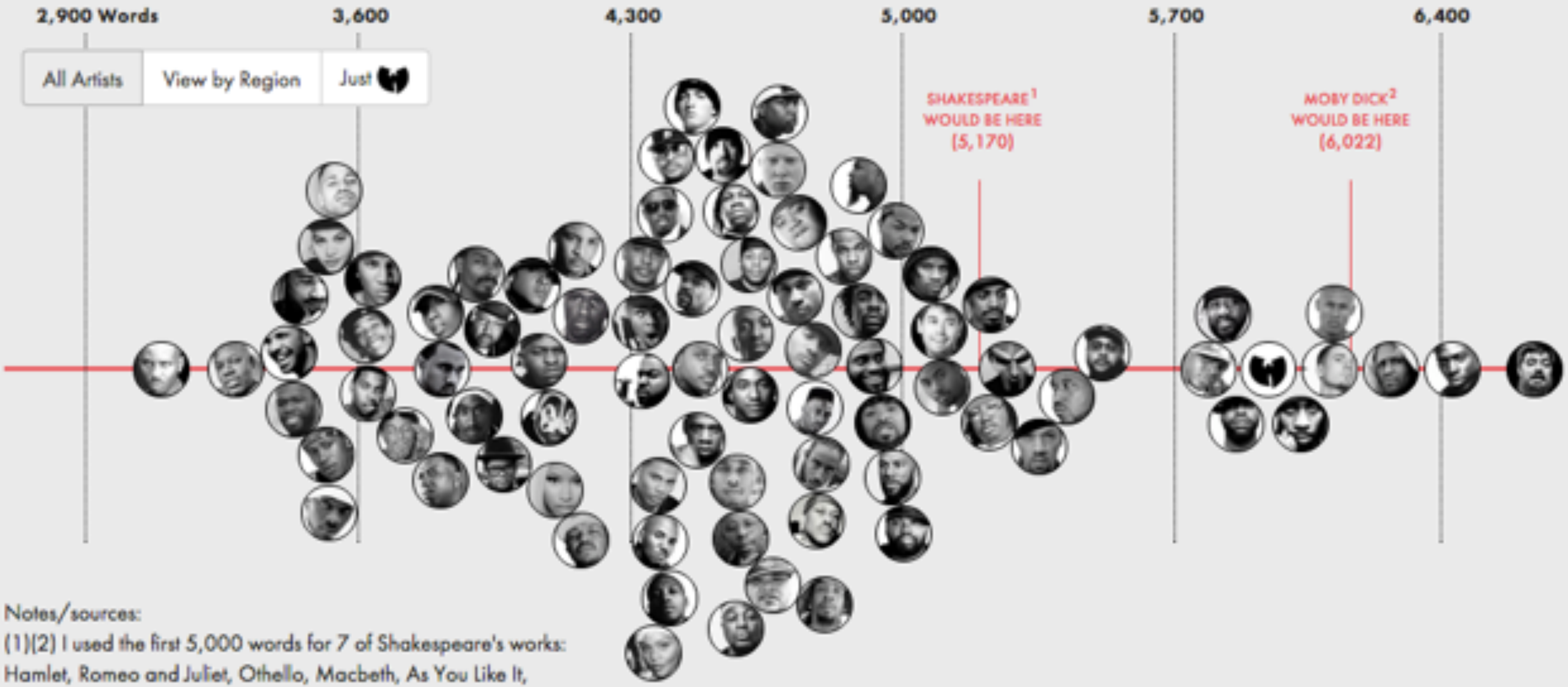


Amy emails with your guest to find the best time and location, knowing your schedule and preferences

Rank	Word Network Feature	Information Gain
1	Term frequency of the word *until*	0.621
2	Neighborhood size of the word *until*	0.611
3	Degree of the word *until*	0.610
4	Neighborhood size of the word *by*	0.576
5	Term frequency of the word *several*	0.574
6	Term frequency of the word *thus*	0.555
7	Degree of the word *thus*	0.553
8	Degree of the word *several*	0.544
9	Neighborhood size of the word *several*	0.543
10	Coreness of the word *thus*	0.538
11	Neighborhood size of the word *though*	0.524
12	Term frequency of the word *had*	0.509
13	Term frequency of the word *by*	0.507
14	Neighborhood size of the word *may*	0.505
15	Degree of the word *or*	0.499
16	Clustering coefficient of the word *said*	0.497
17	Coreness of the word *upon*	0.489
18	Coreness of the word *whom*	0.489
19	Degree of the word *by*	0.488
20	Neighborhood size of the word *returned*	0.484

Table 8: Ranking of term frequency and local word network features based on Information Gain, on Gutenberg data. We took 500 most frequent words on the whole dataset, and collected their term frequency, clustering coefficient, neighborhood size, coreness and vertex degree (for each document) in a single file. This ranking reflects the top 20 among 2,500 features in that file, along with their information gain values. Note that both term frequency as well as local word network features appeared at the top. Moreover, stopwords like *until*, *by*, *several* and *thus* are found to be important predictors of writing style.

OF UNIQUE WORDS USED WITHIN ARTIST'S FIRST 35,000 LYRICS



Notes/sources:
(1)(2) I used the first 5,000 words for 7 of Shakespeare's works: Hamlet, Romeo and Juliet, Othello, Macbeth, As You Like It, Winter's Tale, and Troilus and Cressida. For Melville, I used the first 35,000 words of Moby Dick.
All lyrics are provided by Rap Genius, but are only current to 2012. My lack of recent data prevented me from using quite a few current artists.
This data viz uses code by Amelia Bellamy-Royds's in this jsfiddle.

- Corpus, a large collection of text used for training (e.g. Gutenberg collection or scraping websites)
- Part-of-Speech tagging, understanding the nature of a word, is it a verb or a noun?
- Lexical Analysis, breaking down the structure of text (ie, Document -> Paragraph -> Sentence -> Words).
- Symbolic approach, using rules from language to parse text (can be manually written).
- Statistical approach, a sequence labelling problem, we try to infer the properties of a word by the words around it.

HOW COULD WE RUN SENTIMENT ANALYSIS?

Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng and Christopher Potts
Stanford University, Stanford, CA 94305, USA
richard@socher.org,{aperelyg,jcchuang,ang}@cs.stanford.edu
{jeaneis,manning,cgpotts}@stanford.edu

- Most sentiment prediction systems work just by looking at words in isolation, giving positive points for positive words and negative points for negative words and then summing up these points.
- The order of words is ignored and important information is lost
- It computes the sentiment based on how words compose the meaning of longer phrases. This way, the model is not as easily fooled as previous models

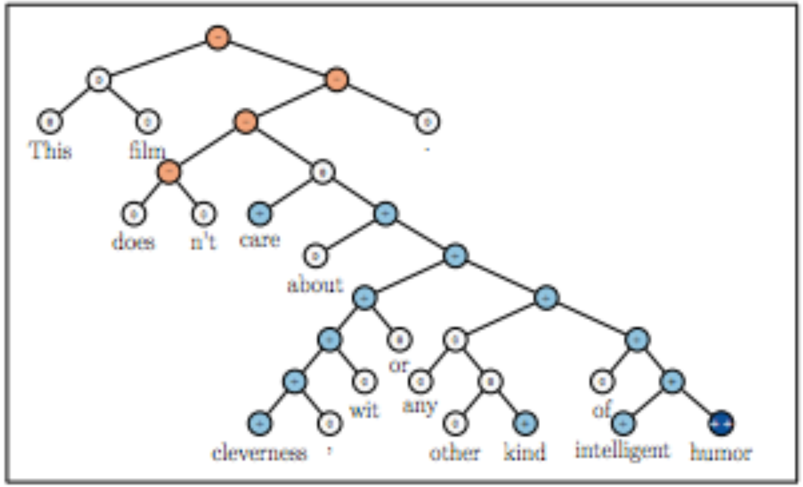


Figure 1: Example of the Recursive Neural Tensor Network accurately predicting 5 sentiment classes, very negative to very positive (– –, –, 0, +, + +), at every node of a parse tree and capturing the negation and its scope in this sentence.

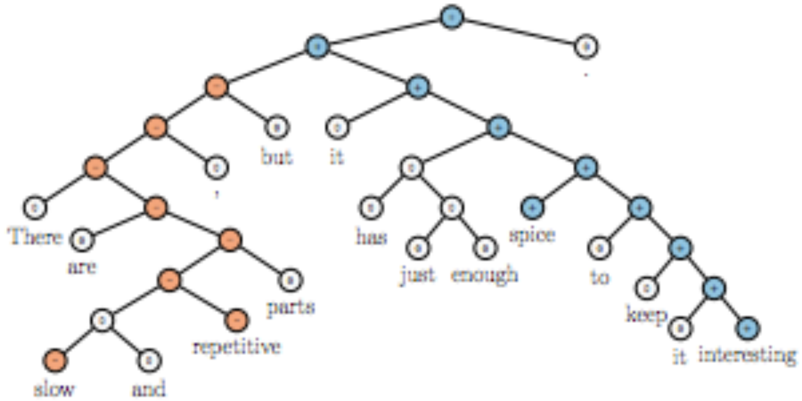
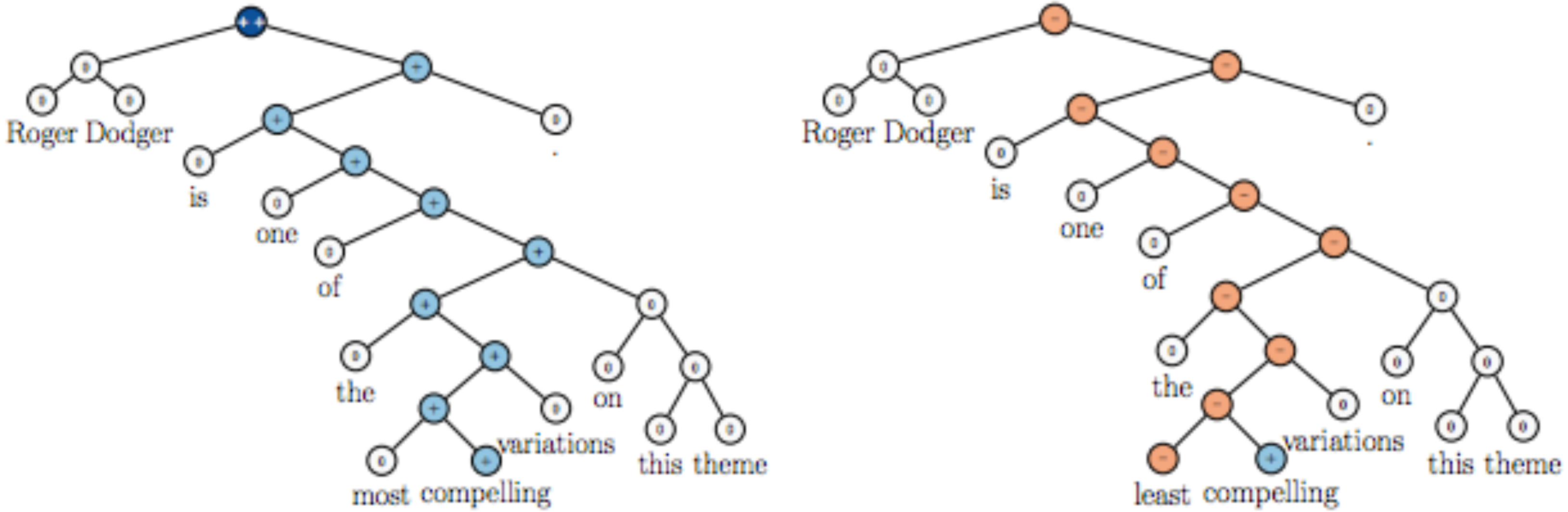


Figure 7: Example of correct prediction for contrastive conjunction *X but Y*.

EXAMPLE - SENTIMENT ANALYSIS



DATA SCIENCE PART TIME COURSE

TOOLS FOR TEXT ANALYSIS

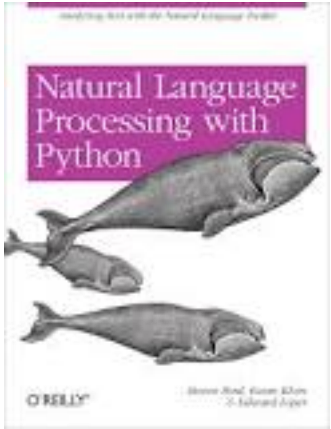
WHAT IS TEXT ANALYSIS?

- Entity Extraction
- Sentiment Analysis
- Keyword Extraction
- Concept Tagging
- Relation Extraction
- Taxonomy Classification
- Author Extraction
- Language Detection
- Text Extraction
- Microformats Parsing
- Feed Detection
- Linked Data Support

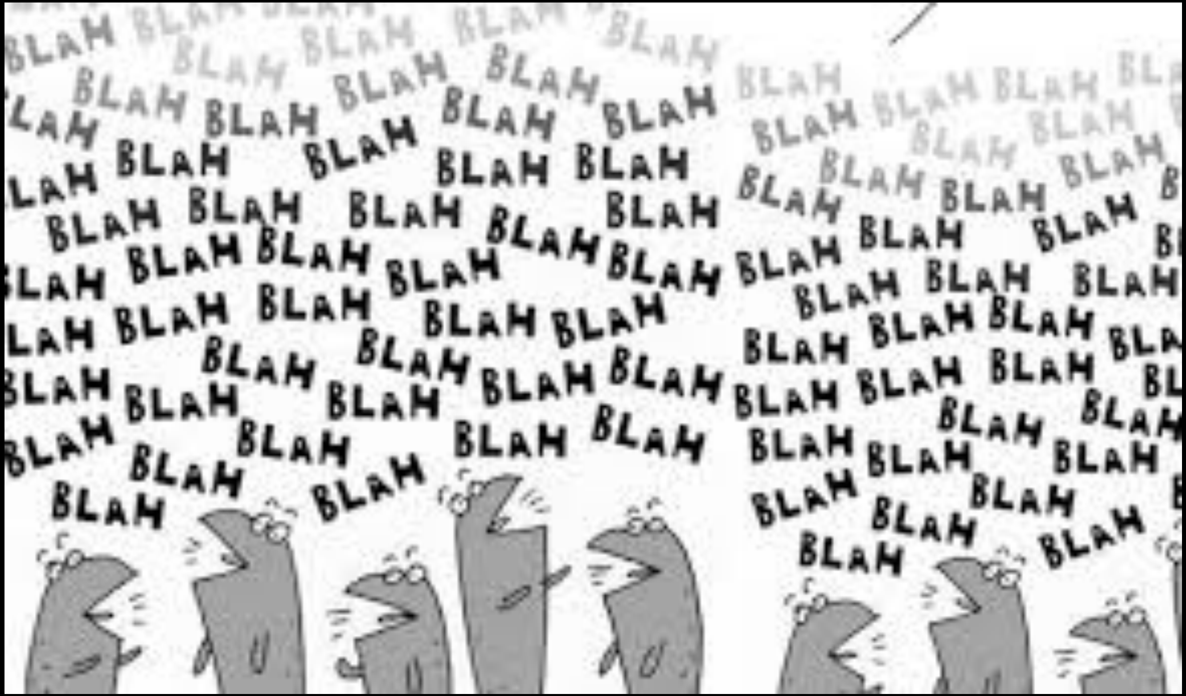
AlchemyAPI™
An IBM Company

NATURAL LANGUAGE TOOLKIT (NLTK) - Python

- NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.



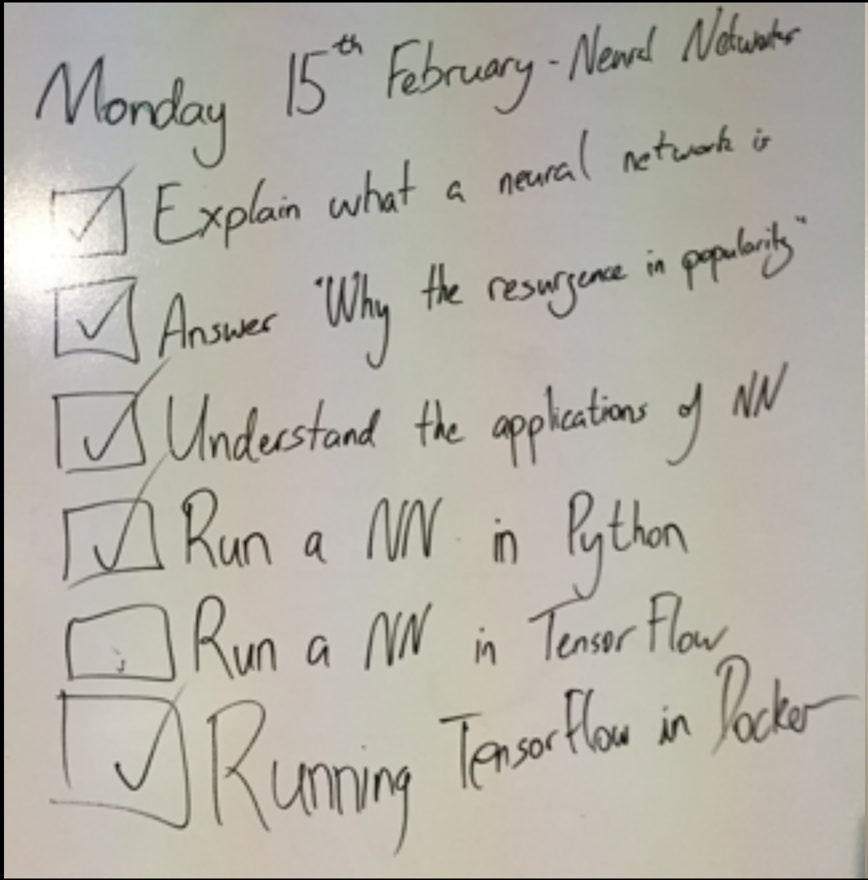
LAB

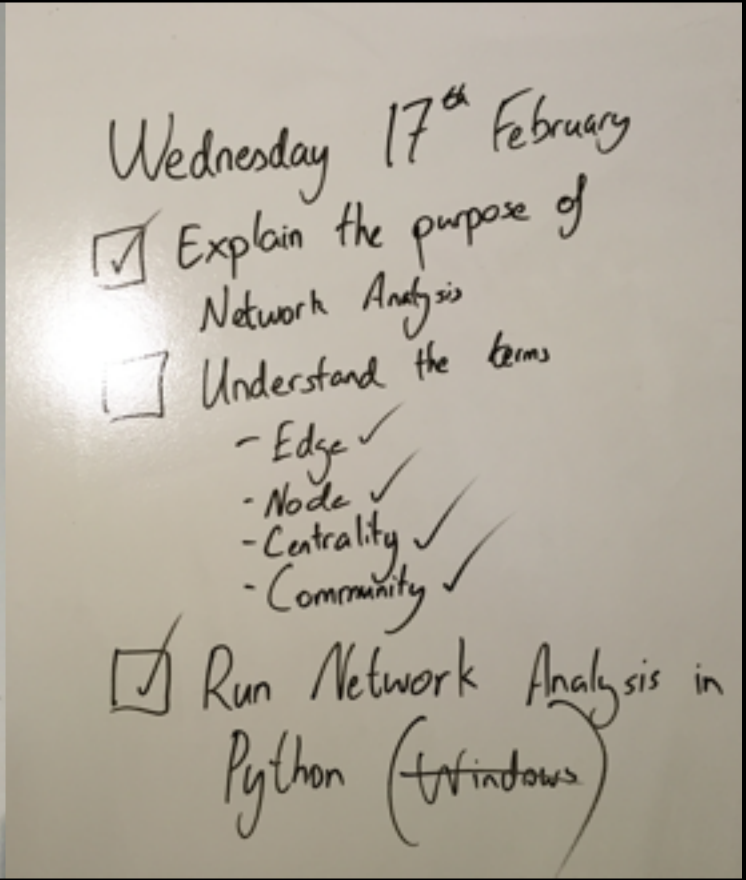


DISCUSSION

- Review of Last Week
- Tasks
- Remaining Course and beyond

DATA SCIENCE – Week 9 Review





Task List

- ☐ Get a list of things you'd like to review on Wednesday
- ☐ Read and Review the Sentiment Analysis paper and answer:
    - ☐ How would you implement this in Tensorflow?
        - ☐ What data would you use?
        - ☐ What model would you use?
        - ☐ How would you validate your model?