GENERAL ASSEMBLY

DATA SCIENCE

SYD DAT 6

Week 7 - Natural Language Processing
Monday 21st November

AGENDA

1. Revise Big Data & Spark
2. Natural Language Processing
3. Case Studies
4. Techniques for NLP
5. Lab
6. Discussion

DATA SCIENCE PART TIME COURSE

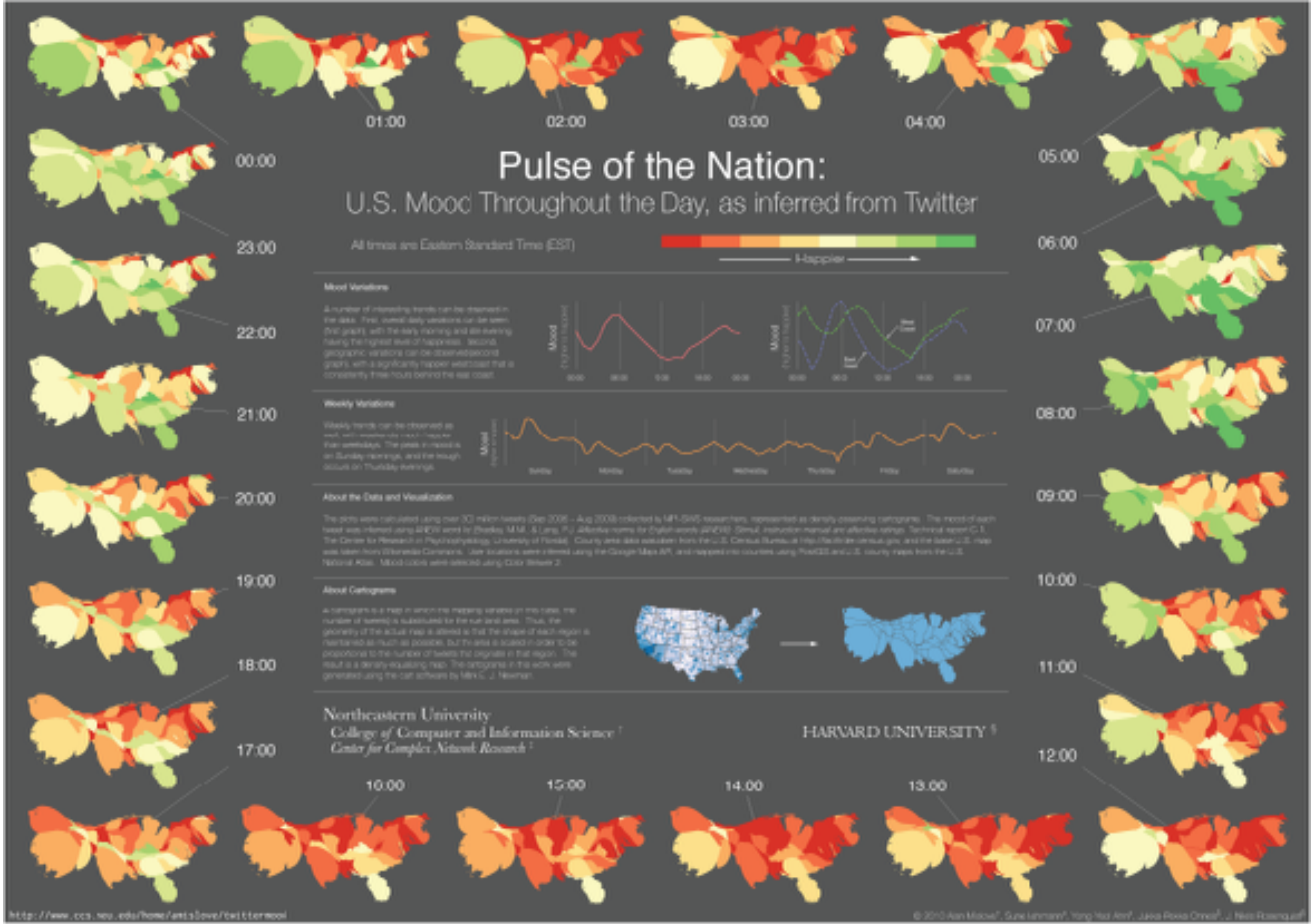
WHAT IS NATURAL LANGUAGE PROCESSING?

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.

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)

Mood Variations

A number of interesting trends can be observed in the data. First, overall daily variations can be seen (first graph), with the early morning and late evening having the highest level of happiness. Second, geographic variations can be observed(second graph), with a significantly happier west coast that is consistently three hours behind the east coast.

Weekly Variations

Weekly trends can be observed as well, with weekends much happier than weekdays. The peak in mood is on Sunday mornings, and the trough occurs on Thursday evenings.

About the Data and Visualization

The plots were calculated using over 300 million tweets (Sep 2006 – Aug 2009) collected by MPI-SWS researchers, represented as density-preserving cartograms. The mood of each tweet was inferred using ANEW word list (Bradley, M.M., & Lang, P.J. *Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings.* Technical report C-1. The Center for Research in Psychophysiology, University of Florida). County area data was taken from the U.S. Census Bureau at http://factfinder.census.gov, and the base U.S. map was taken from Wikimedia Commons. User locations were inferred using the Google Maps API, and mapped into counties using PostGIS and U.S. county maps from the U.S. National Atlas. Mood colors were selected using Color Brewer 2.

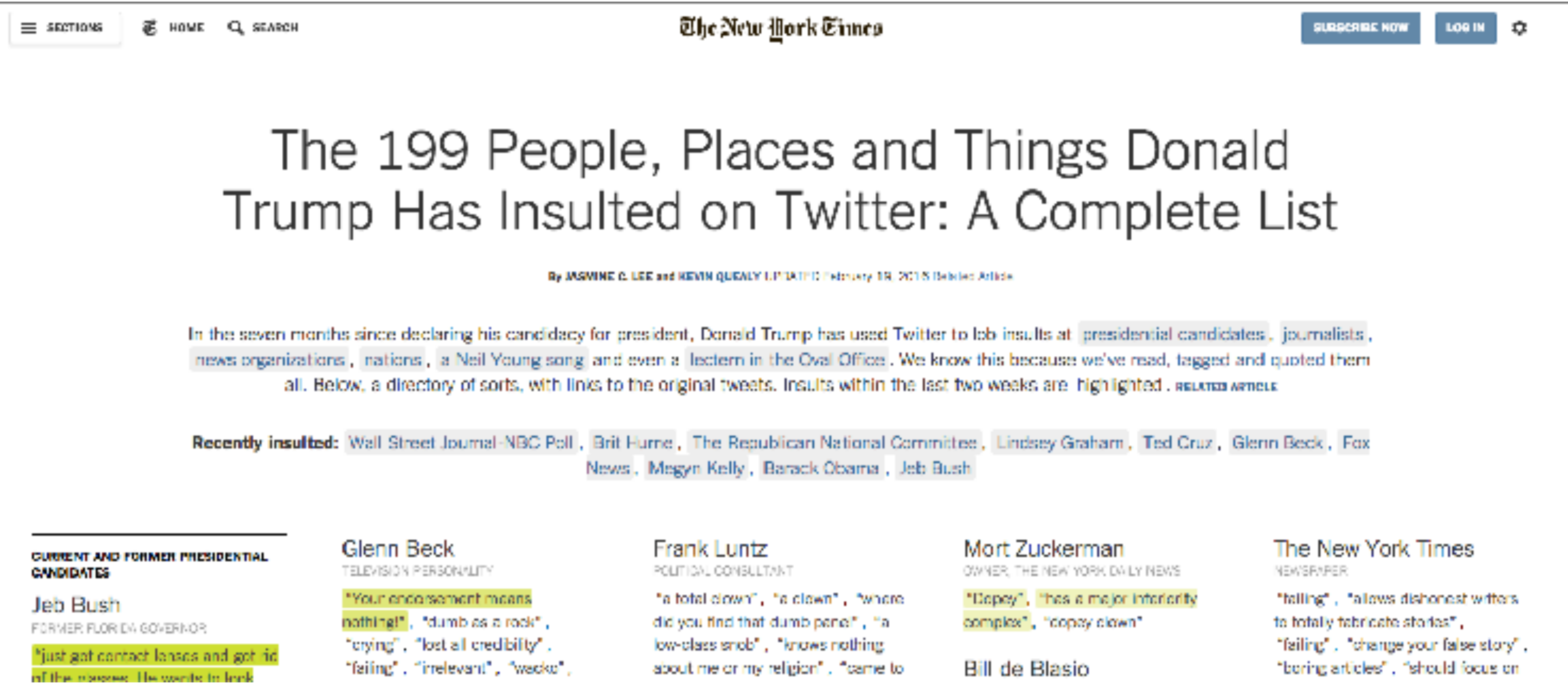
About Cartograms

A cartogram is a map in which the mapping variable (in this case, the number of tweets) is substituted for the true land area. Thus, the geometry of the actual map is altered so that the shape of each region is maintained as much as possible, but the area is scaled in order to be proportional to the number of tweets that originate in that region. The result is a density-equalizing map. The cartograms in this work were generated using the cart software by Mark E. J. Newman.

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 18, 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

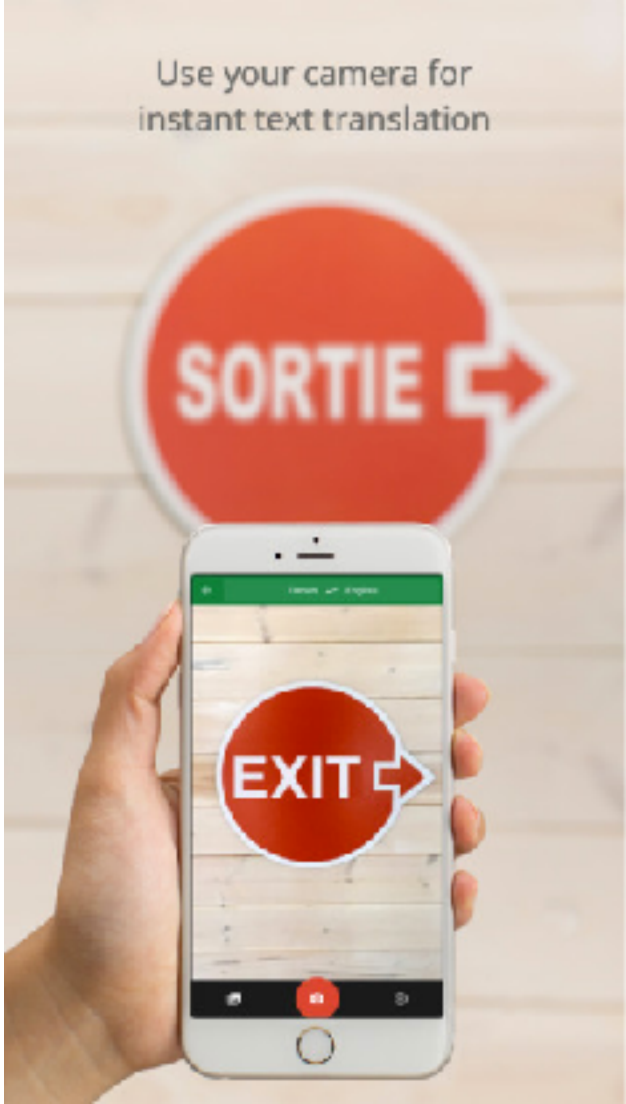
Glenn Beck
TELEVISION PERSONALITY
"Your endorsement means nothing!", "dumb as a rock", "crying", "lost all credibility", "failing", "irrelevant", "wacko",

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

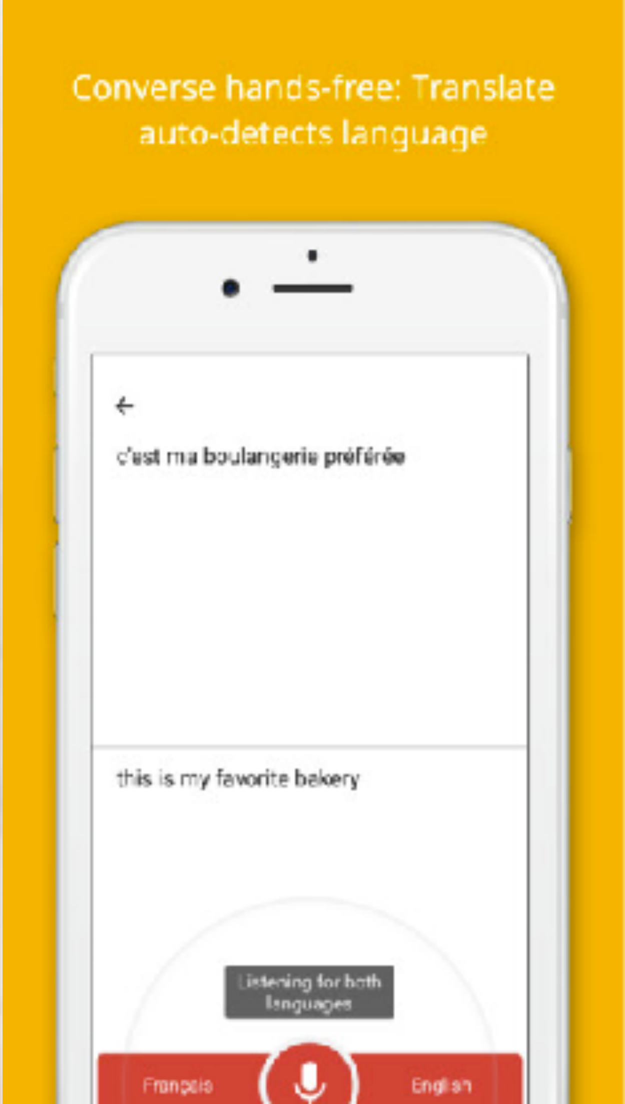
Mort Zuckerman
OWNER, THE NEW YORK DAILY NEWS
"Dopey", "has a major inferiority complex", "dopey clown"

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

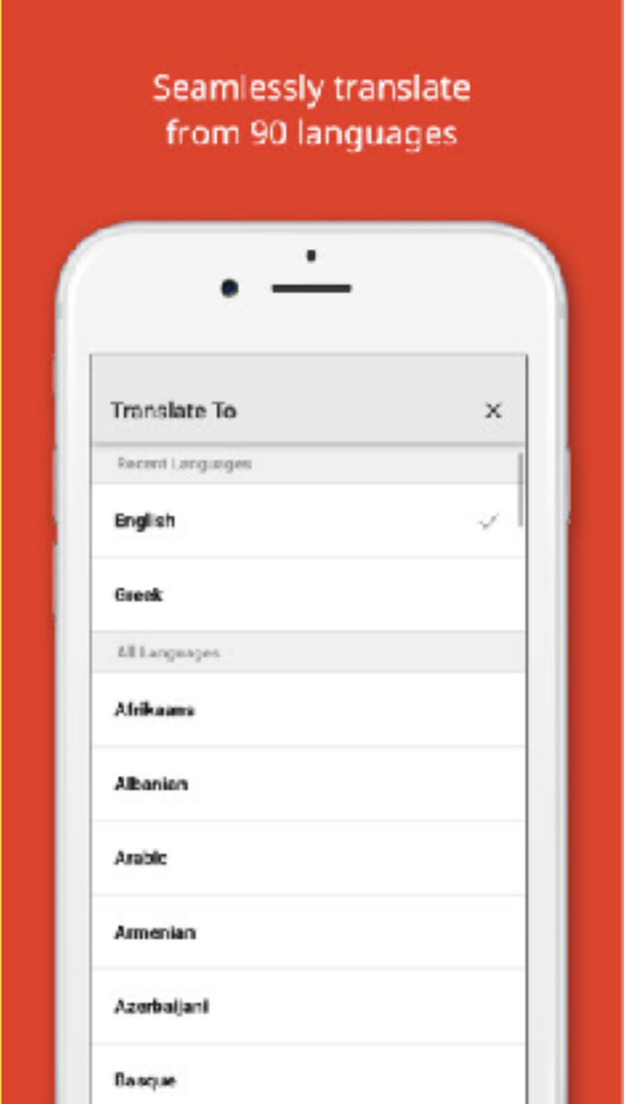


Converse hands-free: Translate auto-detects language

c'est ma boulangerie préférée

this is my favorite bakery

Listening for both languages



Seamlessly translate from 90 languages

Translate To ×

Recent Languages

English ✓

Greek

All Languages

Afrikaans

Albanian

Arabic

Armenian

Azerbaijani

Basque

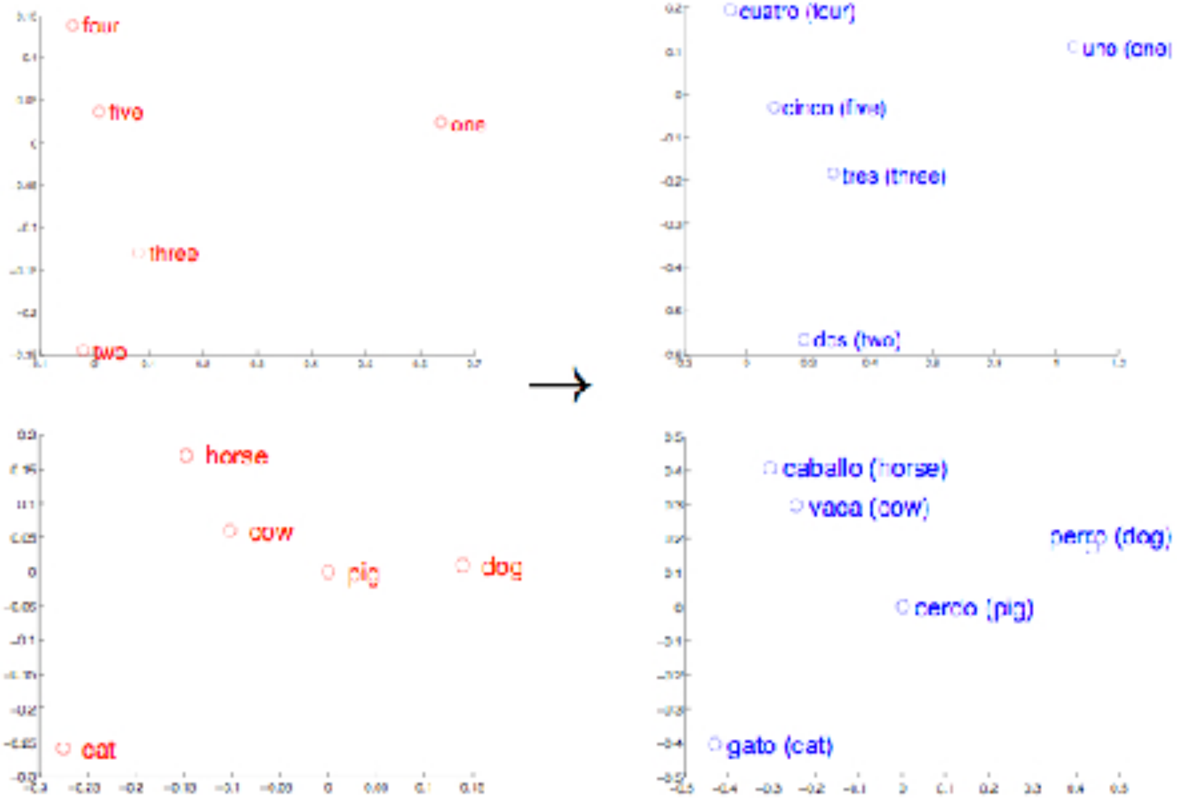
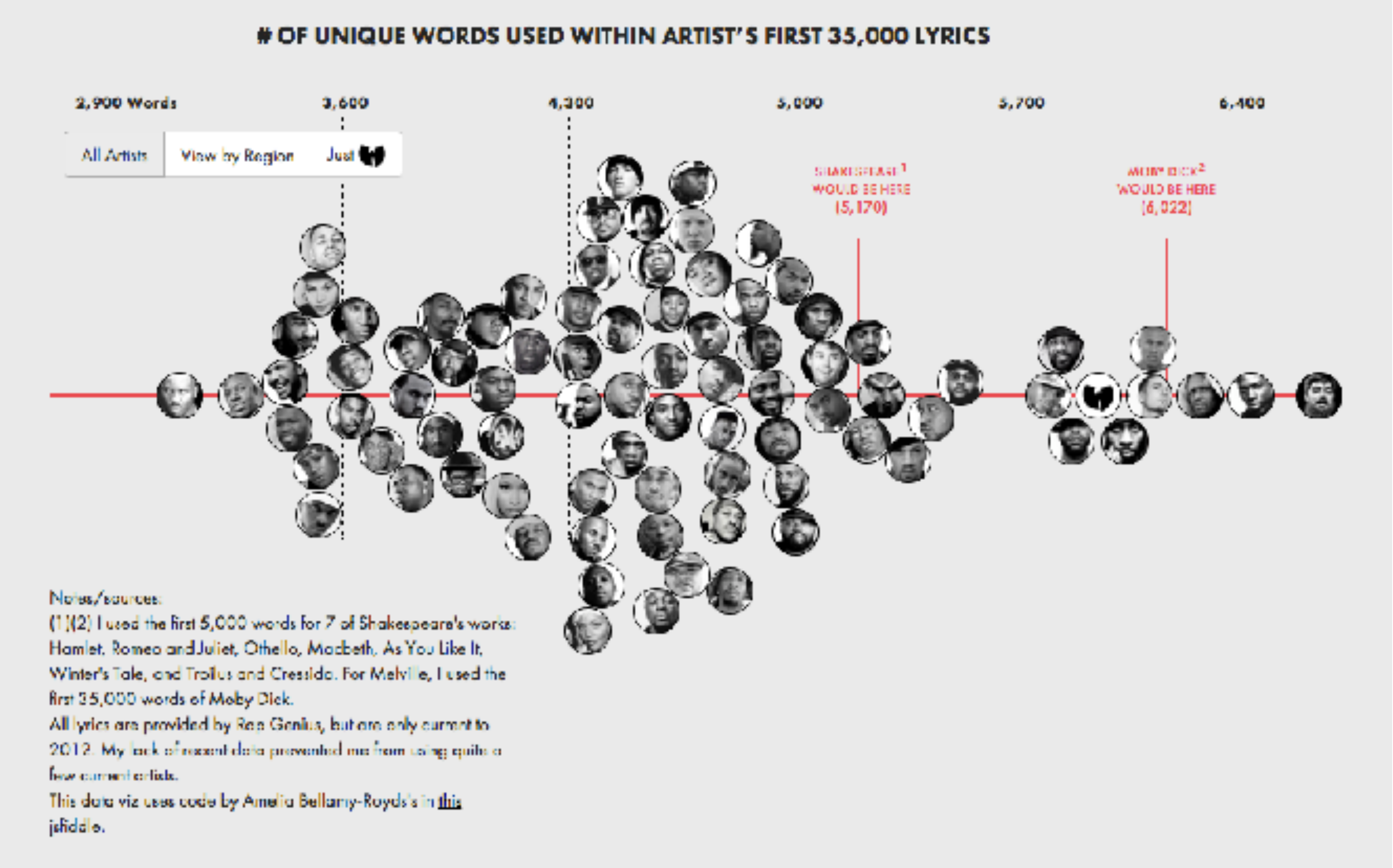


Figure 1: Distributed word vector representations of numbers and animals in English (left) and Spanish (right). The five vectors in each language were projected down to two dimensions using PCA, and then manually rotated to accentuate their similarity. It can be seen that these concepts have similar 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.



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.

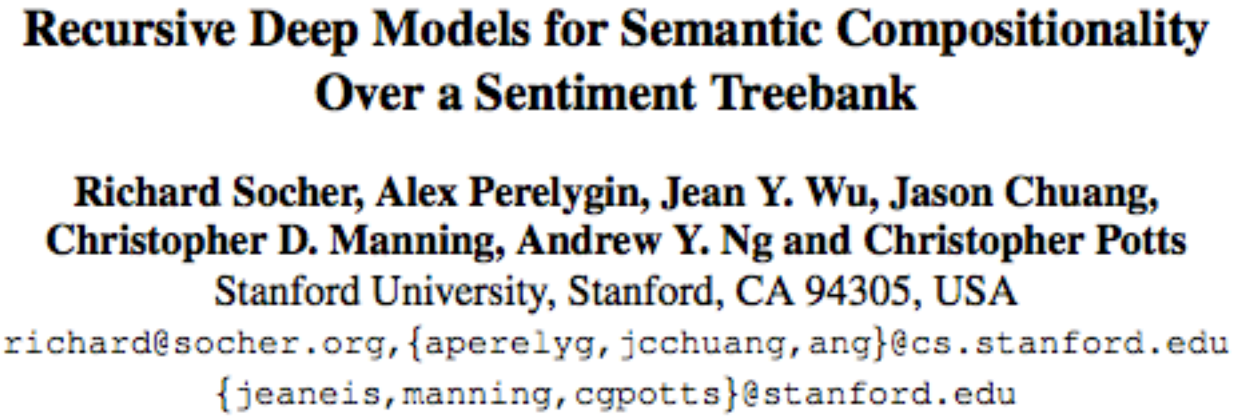
SOME TERMS

- 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.

DATA SCIENCE PART TIME COURSE

HOW COULD WE RUN SENTIMENT ANALYSIS?

EXAMPLE – 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

EXAMPLE - SENTIMENT ANALYSIS

- 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

EXAMPLE – SENTIMENT ANALYSIS

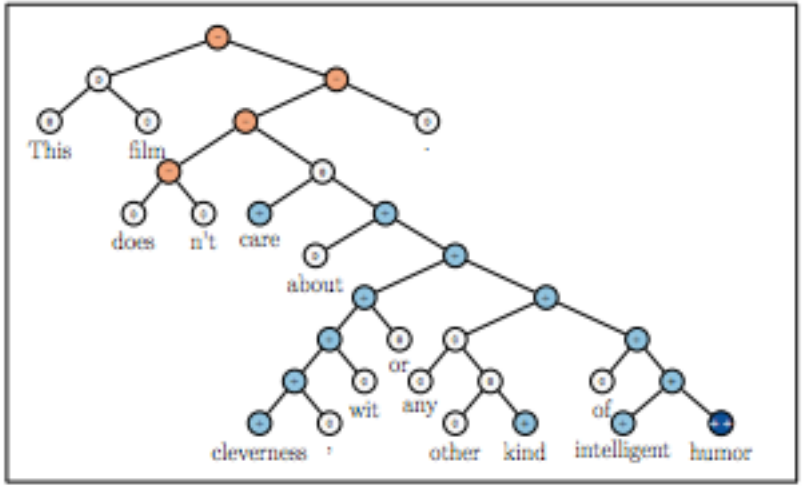


Figure 1: Example of the Recursive Neural Tensor Network accurately predicting 5 sentiment classes, very negative to very positive (– –, –, 0, +, + +), at every node of a parse tree and capturing the negation and its scope in this sentence.

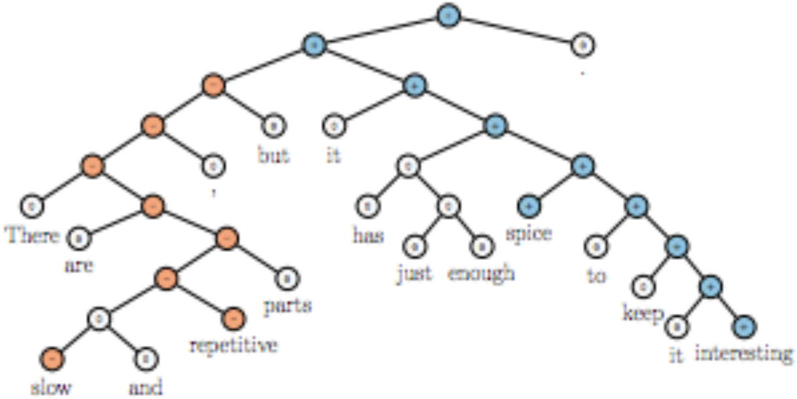
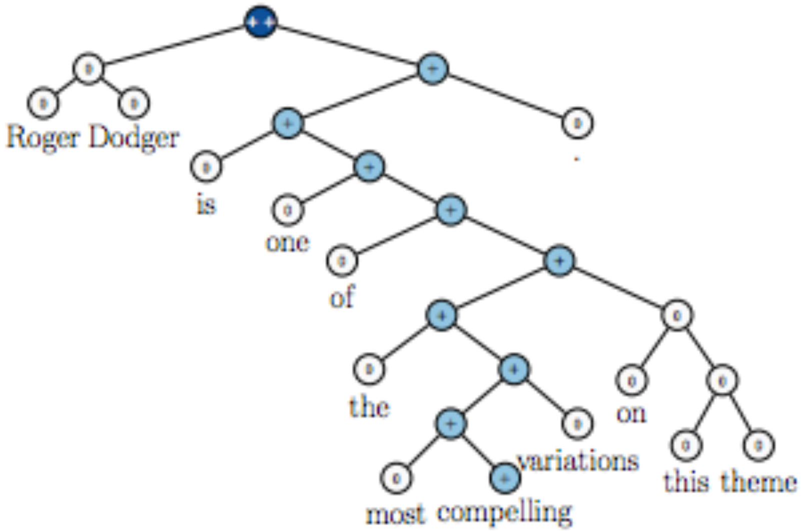
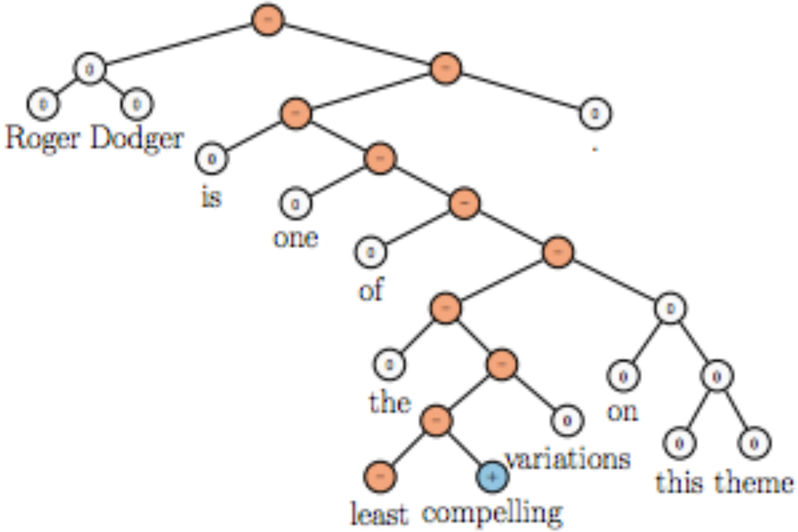


Figure 7: Example of correct prediction for contrastive conjunction *X but Y*.

EXAMPLE – SENTIMENT ANALYSIS





DATA SCIENCE PART TIME COURSE

TOOLS FOR TEXT ANALYSIS

WHAT IS TEXT ANALYSIS?

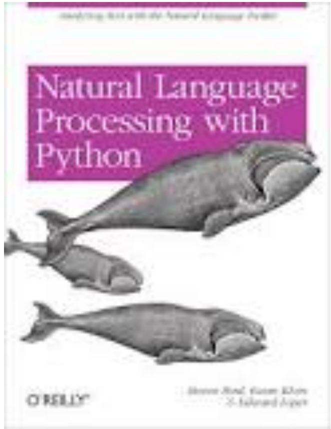
- Entity Extraction
- Sentiment Analysis
- Keyword Extraction
- Concept Tagging
- Relation Extraction
- Taxonomy Classification
- Author Extraction
- Language Detection
- Text Extraction
- Microformats Parsing
- Feed Detection
- Linked Data Support

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

SYNCHING YOUR FORK WITH THE COURSE REPO

1. re-name your labs with lab_name.<yourname>.ipynb (to prevent a conflict)
2. cd <path to the root of your SYD_DAT_6 local repo>
3. commit your changes ahead of sync
    - git status
    - git add .
    - git commit -m “descriptive label for the commit”
    - git status
4. download new material from official course repo (upstream) and merge it
    - git checkout master (ensures you are in the master branch)
    - git fetch upstream
    - git merge upstream/master

HOMEWORK

Homework

- Finish the Lab
- Complete the Mid course survey if you have not already
- Experiment with Alchemy API

Reading

- https://linkurio.us/panama-papers-how-linkurious-enables-icij-to-investigate-the-massive-mossack-fonseca-leaks/