GENERAL ASSEMBLY

DATA SCIENCE

SYD DAT 8

Week 7 – Natural Language Processing
Thursday 6th July

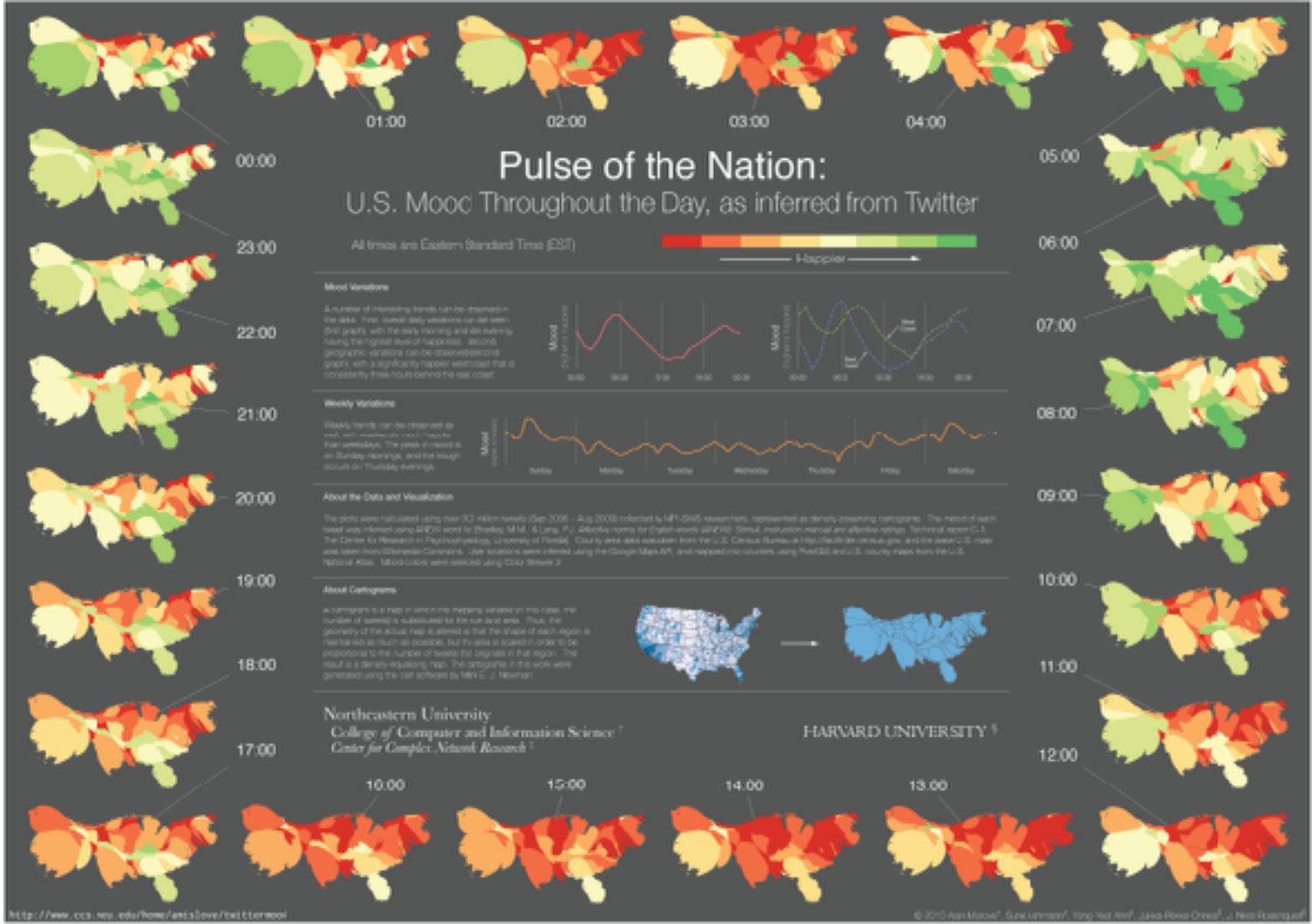
DATA SCIENCE PART TIME COURSE

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?



The New York Times

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 Updated 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 correct 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 pone", "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"

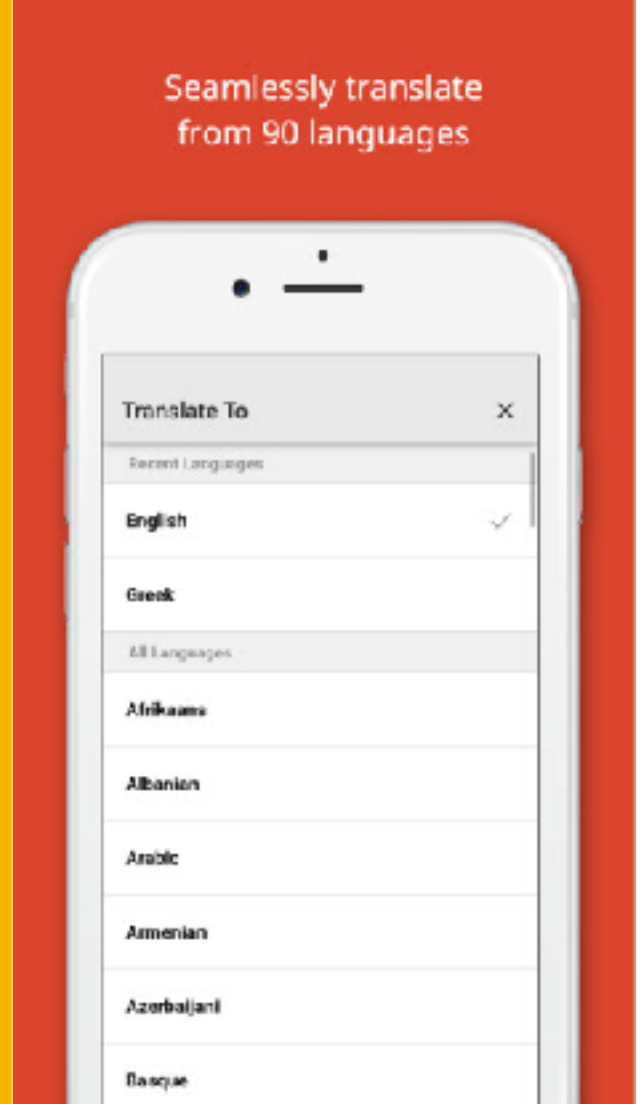
Bill de Blasio

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 bekery
Listening for both
languages
Français
English


Seamlessly translate
from 90 languages
Translate To
English
Greek
Afrikaans
Albanian
Arabic
Armenian
Azerbaijani
Basque

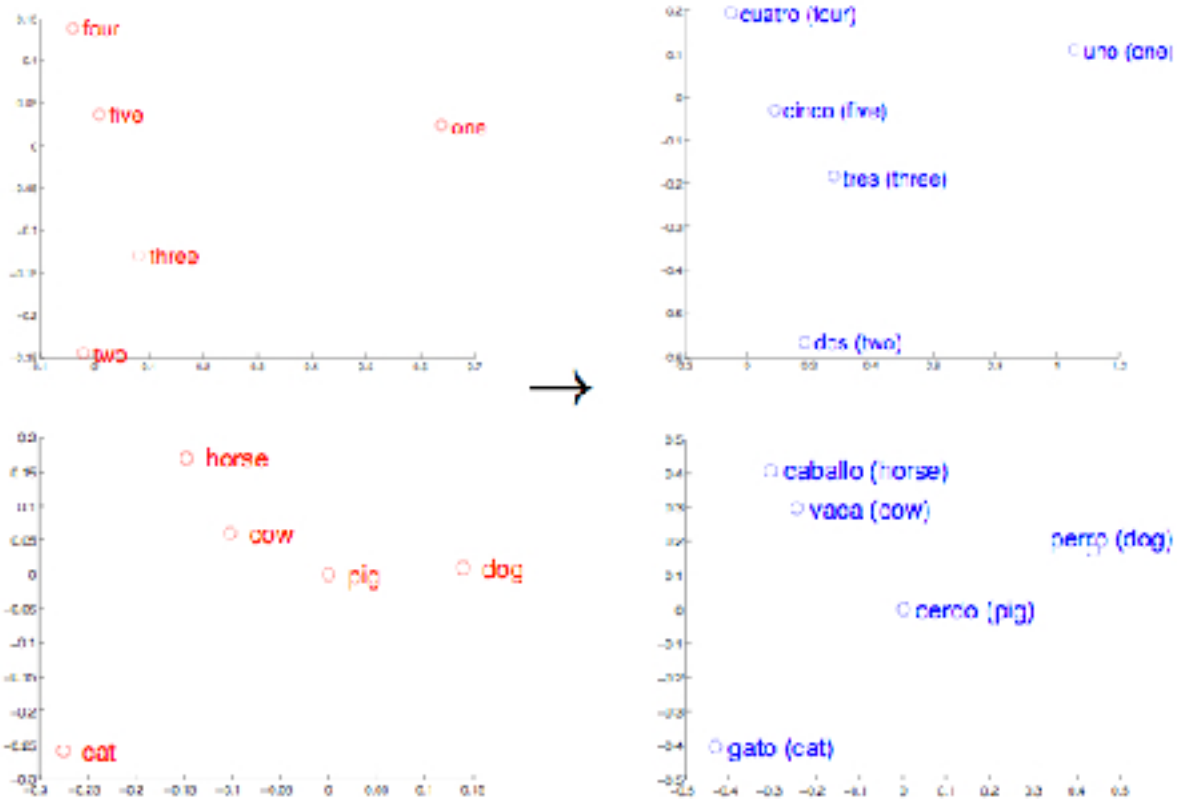
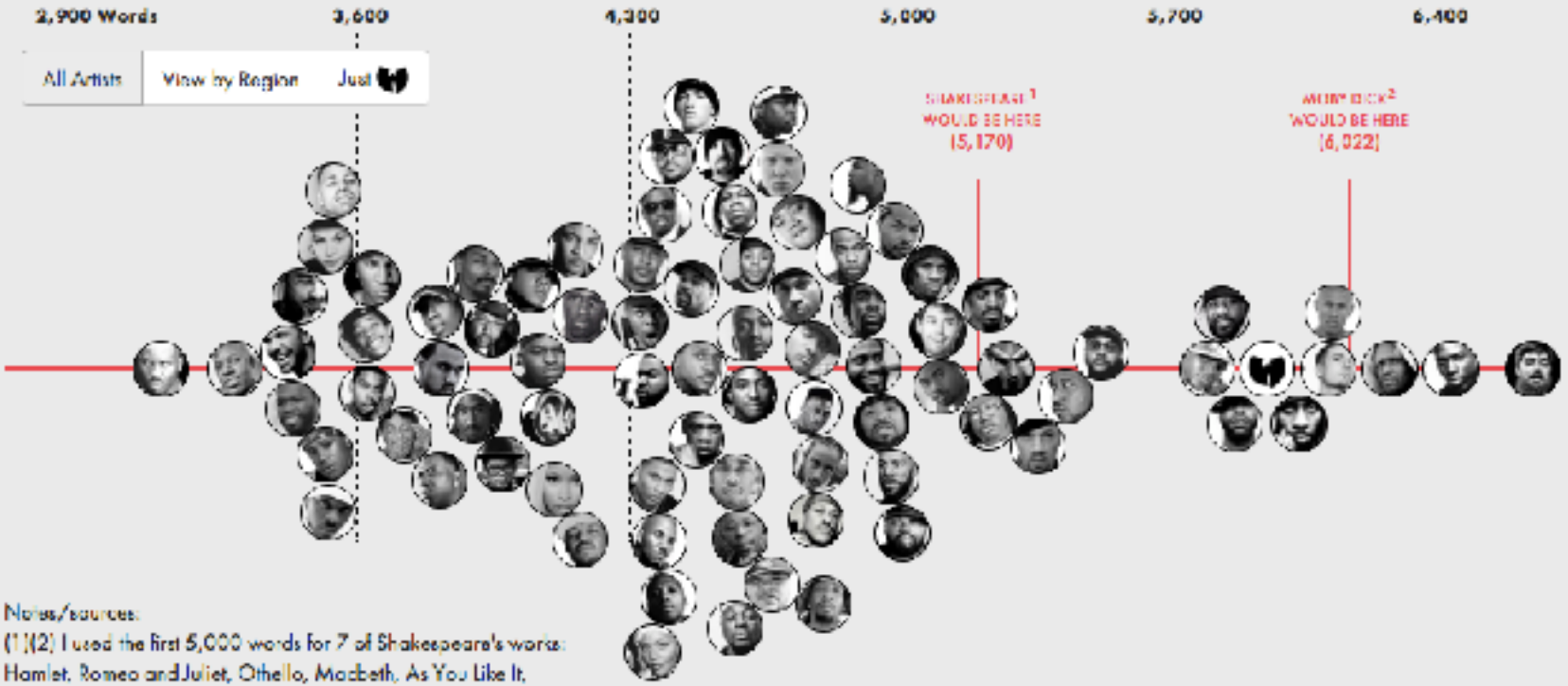


Figure 1: Distributed word vector representations of numbers and animals in English (left) and Spanish (right). The five vectors in each language were projected down to two dimensions using PCA, and then manually rotated to accentuate their similarity. It can be seen that these concepts have similar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

- 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

NLP METHODS

- 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

HOW COULD WE RUN 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

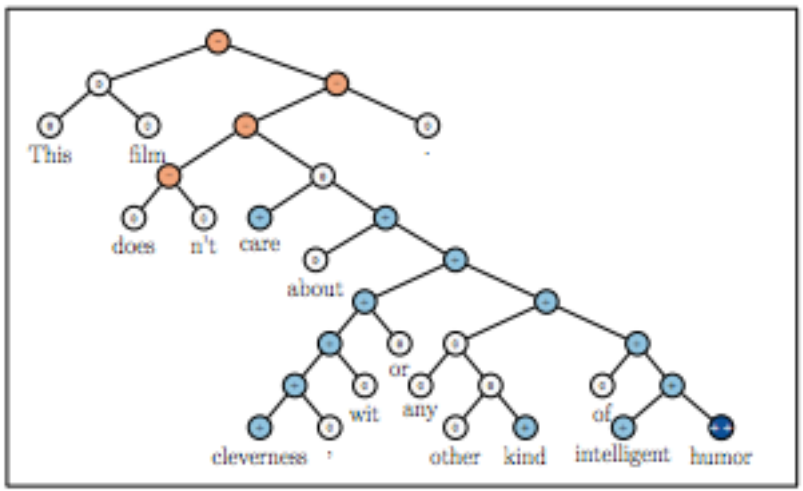


Figure 1: Example of the Recursive Neural Tensor Network accurately predicting 5 sentiment classes, very negative to very positive (– –, –, 0, +, + +), at every node of a parse tree and capturing the negation and its scope in this sentence.

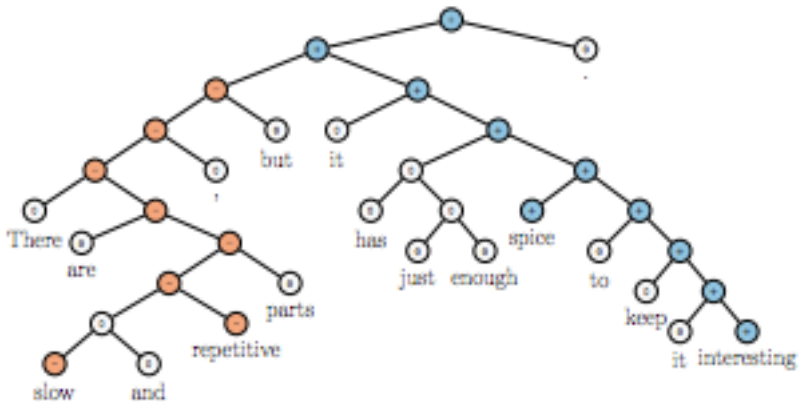
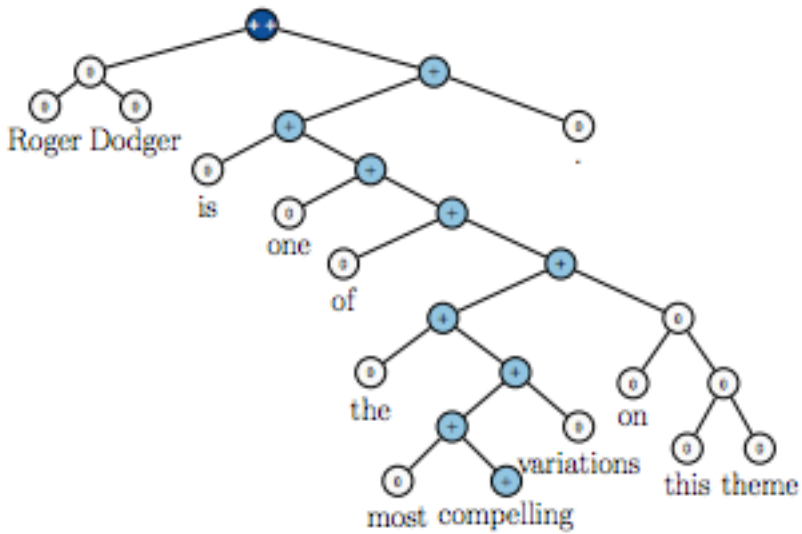
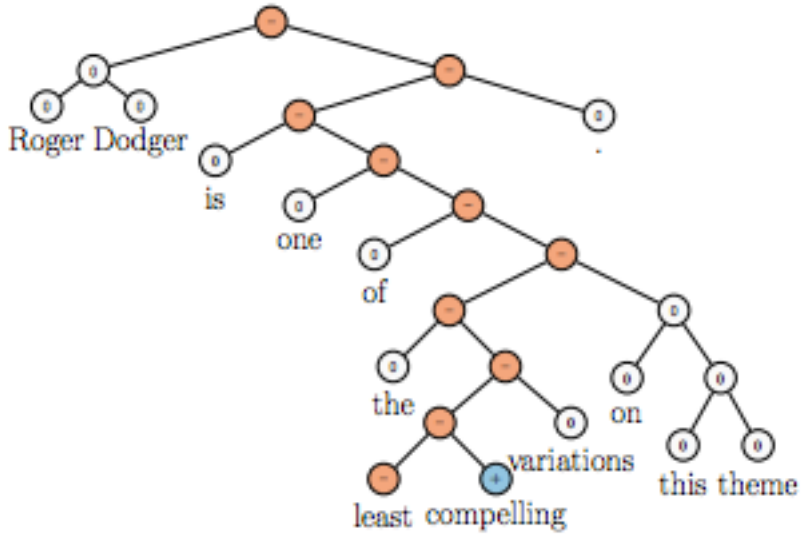


Figure 7: Example of correct prediction for contrastive conjunction X *but* Y.

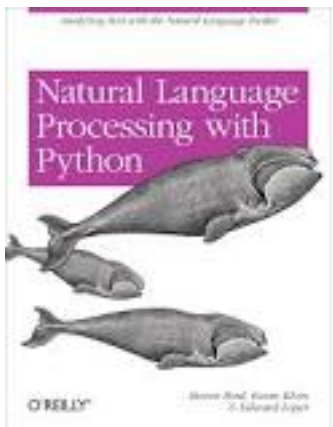

Roger Dodger
is
one
of
the
most compelling
variations
on
this theme
.


Roger Dodger
is
one
of
the
least compelling
variations
on
this theme
.

DATA SCIENCE PART TIME COURSE

TOOLS FOR TEXT ANALYSIS

- 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.



Today there are fantastic APIs available that will perform sophisticated natural language processing on text you submit. Deep learning in combination with other modelling techniques are applied to huge data sources to train the models behind the APIs.

DATA SCIENCE PART TIME COURSE

LAB A
Natural Language APIs

LAB B
Python NLTK

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

- Download and install R from CRAN website
- Download and install RStudio

Reading

- First 2 chapters of Forecasting Principles snd Practice
  https://www.otexts.org/fpp