

Text Analytics:

The One that Introduces Things

Lecture 1: *Text Analytics for Big Data*
Mark Keane, Insight/CSI, UCD

Three Things

◆ Scene Setting:

- ◆ What is ? Text Analytics ...Big Data...

◆ Housekeeping:

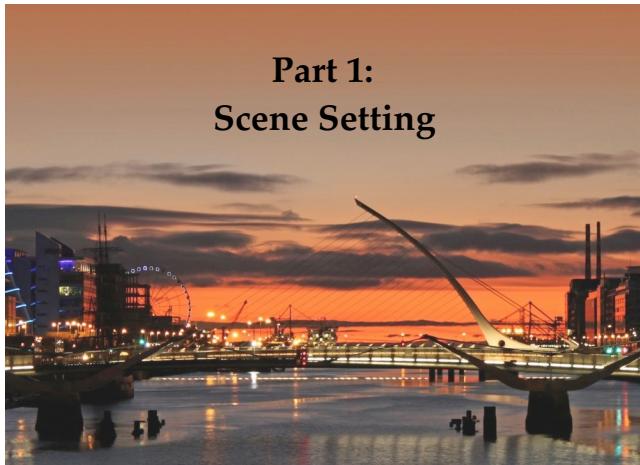
- ◆ Course Structure & Aims

◆ Tooling Up:

- ◆ Installing stuff for Practicals

Part 1:

Scene Setting



De Basics

- ◆ What is Text Analysis ?

- ◆ What is Big Data?

- ◆ What's new about all of this?

- ◆ Example of Text Analytics...

Text Analysis: What are you?

- ◆ The use of textual data to mine, reflect, understand, and represent what happens in the physical, psychological and social world
- ◆ This text comes from books, news, comments, tweets, txts, conversations...whatever is written down...once it is expressed in natural language
- ◆ So...duh...it uses a lot of Natural Language Processing (NLP) techniques !

What is Big Data?

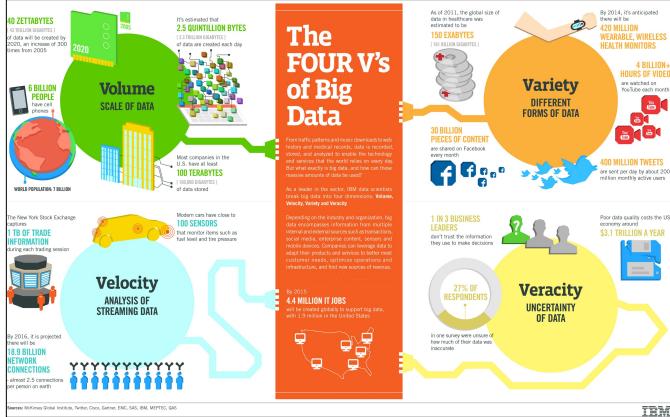
- ◆ Data that is BIG !

- ◆ Billions of items rather than millions... coming at you...at speed...in unstructured forms...

- ◆ Which demands new DBs, new data structures, new algorithms for processing

- ◆ Size, Speed and Scalability become Critical

What is Big Data? IBM~ 4Vs



What's New in Big Data?

- ❖ **Who we know**, says a lot about who we are...
 - ❖ Facebook friends, linked-in network, tweet followers
- ❖ **What we write or say**, says a lot about what we think...
 - ❖ Text in books, news, blogs, social media and so on
- ❖ **Where we located**, says a lot about us...
 - ❖ location-based sensing, GPS, IP-addresses
- ❖ **What we do**, says a lot about our decisions/interests...
 - ❖ what we buy, web-sites visited, youtube videos watched, news re-tweeted, items shared, what we comment on and so on...

What's New in Big Data? *Text Bits*

- ❖ **Who we know**, says a lot about who we are...
 - ❖ Facebook friends, linked-in network, tweet followers
- ❖ **What we write or say**, says a lot about what we think...
 - ❖ Text in books, news, blogs, social media and so on
- ❖ **Where we located**, says a lot about us...
 - ❖ location-based sensing, GPS, IP-addresses
- ❖ **What we do**, says a lot about our decisions/interests...
 - ❖ what we buy, web-sites visited, youtube videos watched, news re-tweeted, items shared, what we comment on and so on...

What's Old in Big Data?

- ❖ Good old-fashioned, data analysis
- ❖ Many statistical ideas are very familiar
- ❖ Many research problems are familiar
- ❖ Proper collection of data is important
- ❖ Proper treatment of data is critical

What's Really New?

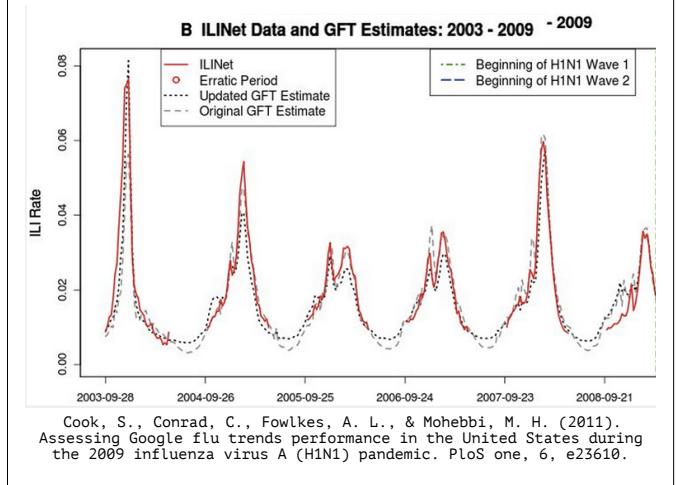
- ❖ **Tipping-point with Very Large Data Sets**
 - ❖ from 100s to 1,000,000,000s of data points
- ❖ **Unusual Types of Data**
 - ❖ video, text, thumbs-up, unstructured data
- ❖ **Non-standard Data Sources**
 - ❖ social media (FB, Tweets), news, phones
- ❖ **Data is not gathered conventionally...indirect**
 - ❖ the *sensing devices* are doing other things

Text Analytics Examples...

- ❖ Flu Epidemic Detection from Search
- ❖ Political Orientation from Twitter
- ❖ Public Sentiment on Stockmarket
- ❖ Predicting Movie Box Office Revenues
- ❖ Tracking Earthquakes and Typhoons

Search Terms & Flu Epidemics

- ◆ Google Flu Trends (GFT) aggregates search data, count in flu keywords
- ◆ US Centre for Disease Control (ILIs) tracks influenza like illnesses in outpatient data
- ◆ From 2003-2009 GFT shows high correlation with ILI stats (ILINet)
- ◆ GFT (sort of) predicts flu outbreaks & their locations

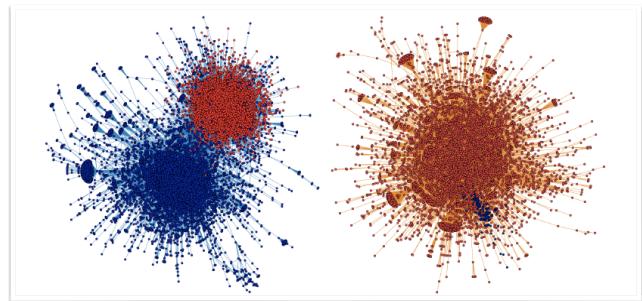


Political Views in Twitter

- ◆ People with different political biases use different hashtags in tweets (#irishwater, #dontpayforwater)
- ◆ 250K Tweets from 2010 US Congress Election
- ◆ Political retweets highly segregated, no cross-posting between left- and right-leaning users

Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., & Flammini, A. (2011, July). Political polarization on twitter. In ICWSM.

Political Polarities



Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., & Flammini, A. (2011, July). Political polarization on twitter. In ICWSM.

Sentiment in News

- ◆ Words can be classified as positive or negative in their sentiment
- ◆ If you look at these over time they can tell you a lot about what people are thinking
- ◆ Could be used to predict stock movements

Gerow, A., & Keane, M. T. (2011). Mining the web for the voice of the herd to track stock market bubbles. *IJCAI-11*. AAAI Press.

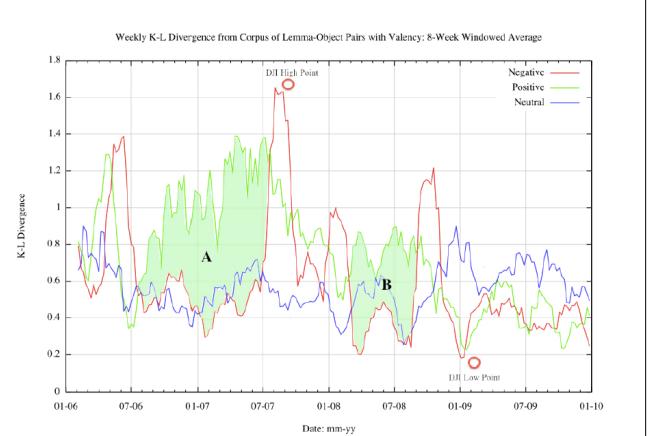
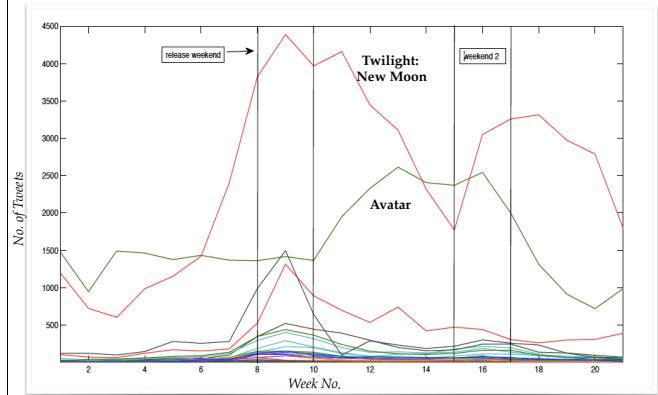


Figure 2: Symmetric K-L divergence (8-week windowed mean) of positive, negative, and neutral lemma-object pairs. Note, the two regions, A and B, of distinct positive-negative divergence preceding the 2007 crash and subsequently the beginning of the recovery in 2009.

Tweet Rate Tracks Attention

- ◆ The amount of conversation (in twitter or FB) about something can reflect attention to it, what people are talking about
- ◆ This attention can predict movie revenue
- ◆ Attention and location can track events

Movies: Shape of Attention



Tweet Rate & Revenue

- ◆ The amount of conversation (in twitter or FB) about something can reflect attention to it, what people are talking about
- ◆ This attention can predict movie revenue
- ◆ Attention and location can track events

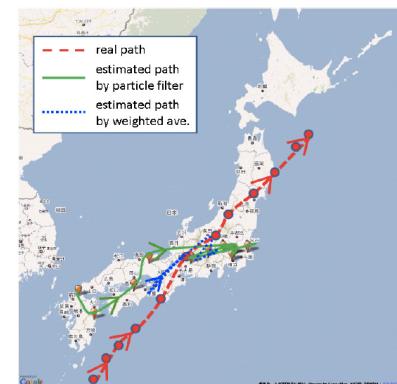
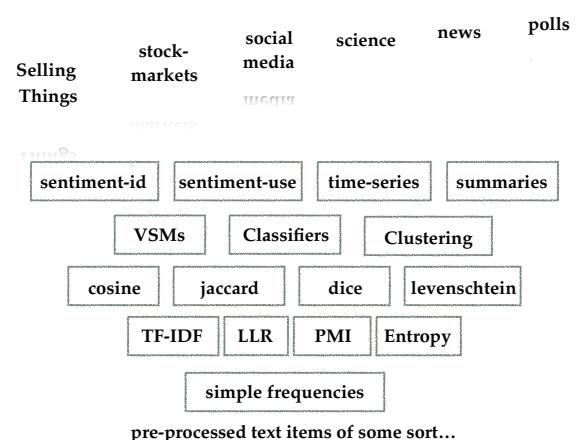
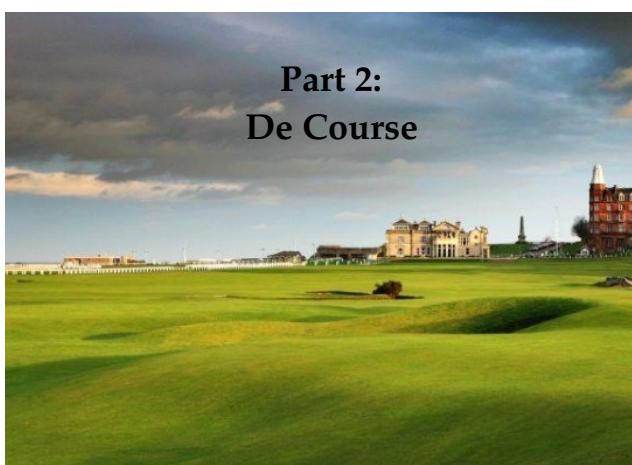


Figure 10: Typhoon trajectory estimation based on tweets.

Part 2: De Course



Course Structure

- ◆ 12 Weeks of Lectures & Practicals
- ◆ Each week 2-hours of lectures, 1-hour practical
- ◆ Practical hour is to get advice/help with coursework which you done on your own
- ◆ You are expected to **complete** work in your own time; outside of the practical hour

Lectures: 12 Weeks (2 hours)

- ◆ Two hours (after the practical on previous week)
- ◆ Will explain main techniques cumulatively
- ◆ Will give examples from current research papers
- ◆ Aim: give you a working knowledge of area; to be able to read most papers and understand them
- ◆ Assume no prior knowledge, so you may know some areas already (sleep advised...)

Lectures: Caveats I

- ◆ If you have done lots of NLP, this may not be for you but, it could be, if it was just NLP
- ◆ If you have done lots of STATS, you may know too much already...but applications may be new
- ◆ If you have done a lot of ML, there may not be enough new stuff in this course for you
- ◆ If your profile matches, maybe choose another...

Lectures: Caveats II

- ◆ Note, this is a developing course, 3rd year, so all descriptions (aims/lectures) are *provisional*
- ◆ They may not change a lot, but I reserve the right to modify, as we see how they go
- ◆ We may have to make things harder/easier depending on how smart you are...
- ◆ If you don't like this uncertainty, choose another...

Lectures: Caveats III

- ◆ Course philosophy is, unashamedly, pragmatic:
"I want to text analyse today; how do I do it, what do I use and how do I implement that technique"
- ◆ Less on mathematical/statistical basis for techniques, less on derivations/proofs...
- ◆ Sorry, but I am simple soul...sole...soul?
- ◆ If this pragmatism does not suit, choose another...

Practicals: 12 Weeks (1 hour)

- ◆ One hour practical will precede lecture hours...someone available to answer questions about the practical work; set the week before
- ◆ You should use it to clarify any issues/problems you have encountered with the practical over the week
- ◆ Finish them, in your own time, outside class hours
- ◆ Generally, deadline for practicals will be one-week later; submit on moodle day after following week's lecture

Practicals: Wha?

- ❖ The practicals will involve a mix of different sorts of activities; programs, problems, puzzles....designed to get practical experience with the techniques
- ❖ We will use Excel, R (a bit) and Python in practicals
- ❖ NB: I will not be teaching you Python, just using existing programs and packages (nltk)
- ❖ Obviously, you can do this, but it is not required to complete the practicals

Assessment

- ❖ Assessed on Coursework and Exam
- ❖ Coursework (Practicals 10ish): 40%
- ❖ Written Exam (2hr, no books) : 60%
- ❖ Plagiarism is not tolerated

Coursework

Coursework is **40%** of total marks

Coursework made up from **10* weekly** practicals

Practicals marked on a **pass/fail** basis:

Pass: all questions answered adequately ($\approx 4\%$)

Fail: any question answered inadequately ($\approx 0\%$)

* Unless otherwise advertised during semester.

Begin Rant{

- ❖ **Post-Truth World**...No one knows anything and experts are suspect (c.f. climate change)
- ❖ **Education Commoditised**...I paid for my degree, where is it?
- ❖ **Lets Negotiate this Mark, Mark**...Can appeal procedural defects...but not dislike

}End Rant

Plagiarism

- ❖ Two Handouts are given:
 - ❖ CSI Plagiarism Procedures
 - ❖ UCD Plagiarism Policy & Procedures
- ❖ In UCD, plagiarism is not common or accepted; copyier and copyee are both guilty
- ❖ Here to learn: cheating on coursework is unfair to others and will undermine your career

Plagiarism & UCD Computer Science

- **Plagiarism is a serious academic offence**
 - [Student Code, section 6.2] or [UCD Registry Plagiarism Policy] or [CS Plagiarism policy and procedures]
- Our staff and demonstrators are **proactive** in looking for possible plagiarism in all submitted work
- Suspected plagiarism is reported to the CS Plagiarism subcommittee for investigation
 - Usually includes an interview with student(s) involved
 - 1st offence: **usually** 0 or NG in the affected components
 - 2nd offence: referred to the **University disciplinary committee**
- Student who enables plagiarism is equally responsible
 - http://www.ucd.ie/registry/academicsecretariat/docs/plagiarism_po.pdf
 - http://www.ucd.ie/registry/academicsecretariat/docs/student_code.pdf
 - <http://libguides.ucd.ie/academicintegrity>



Part 3: Tooling Up

This Week

- ◆ Establish Your Environment:
 - ◆ Install R
 - ◆ Install Python and IDLE*
 - ◆ Excel (or any spreadsheet app)

* Later we will look at a real IDE

* Consider Anaconda...Jupyter