Technology Fundamentals of Business Analytics

Kaggle 1

- TitanicOverview(10)
- TitanicPython(10)
- TitanicR(10)
- Comparison(10)
- ThirdOverview(10)
- ThirdSolution(10)
- Formatting(10)

Overview

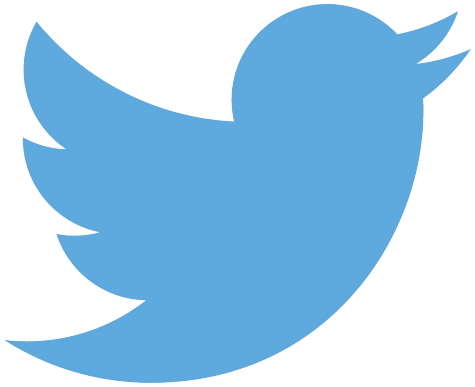
- Text mining and why it is important
- Overview of text mining
- Text mining with Twitter

http://www.chrismadden.co.uk/cartoon-gallery/turing-test-cartoon-turing-test-being-failed-by-a-human/

“The Turing test is a test of a machine's ability to exhibit intelligent behavior equivalent to, or indistinguishable from, that of a human.”

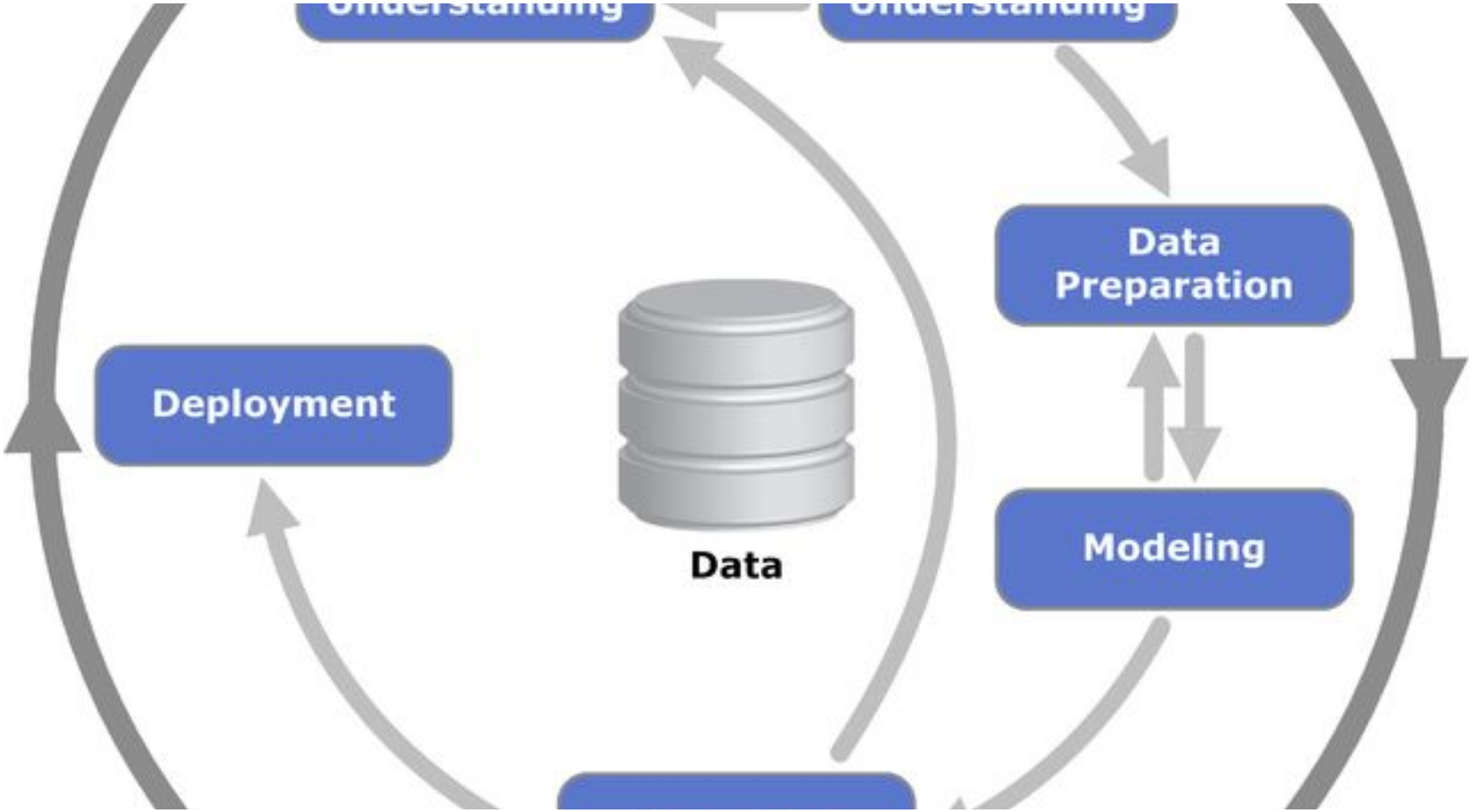
Why is text mining important?

So Much Data is Unstructured!

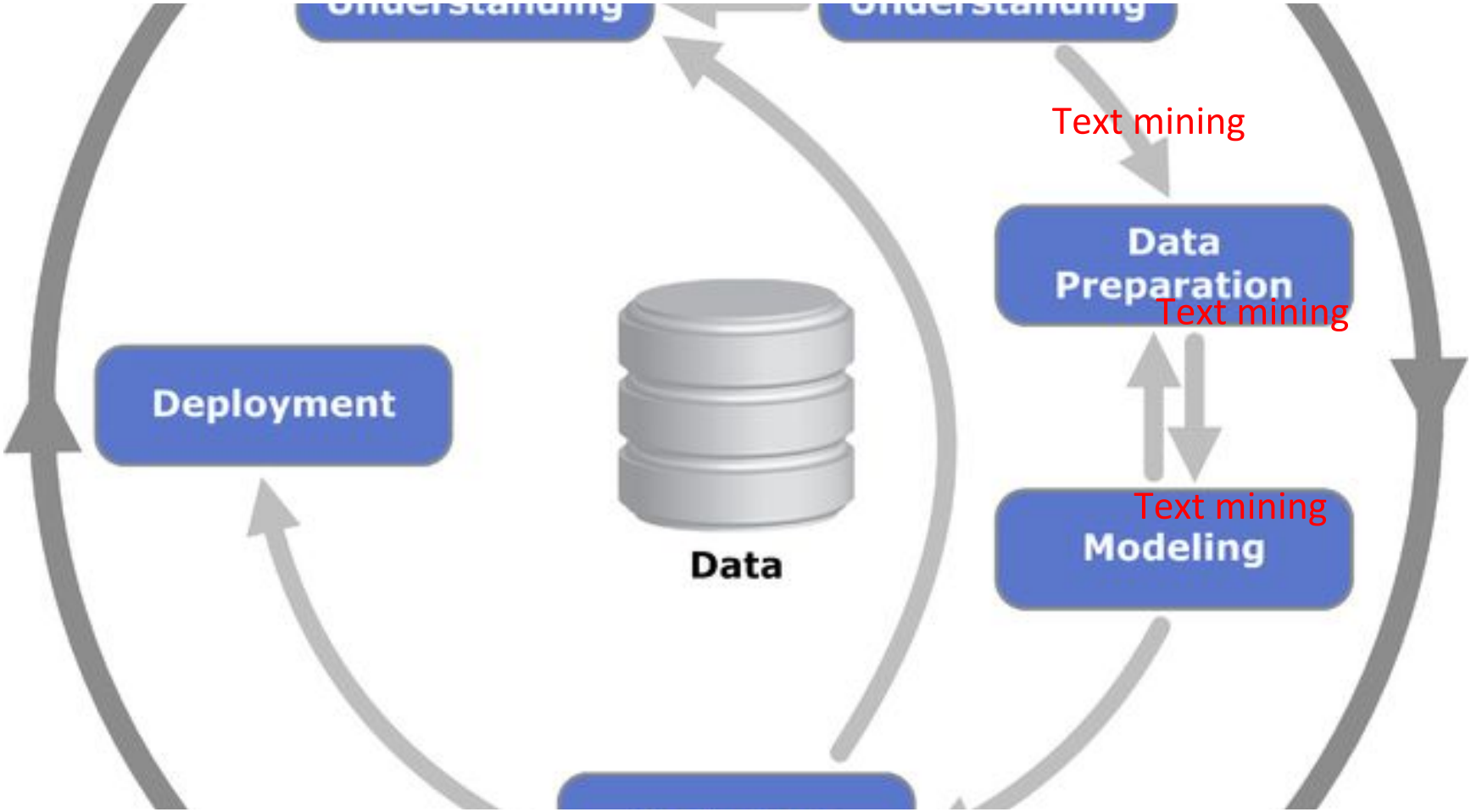


Text is knowledge, and knowledge can answer questions.

Where does text mining happen?



Where does visualization happen?



Key Text Mining Terms

- A collection of documents – corpus
- Document – a piece of text
- Terms/tokens – a word in a document
- Entity – Some type of person, place, or organization

Data Understanding

A text mining analyst typically starts with a set of highly heterogeneous input texts.

- How is the text formatted?
- What types of text data is present?
- What should be removed because it is irrelevant?
- What types of analyses are likely to be useful?

Data Preparation

- Stopword removal
- Stemming procedures
- Lemmatisation
- Misc…..
- Creation of Corpus

Stop Words

- Words that are filtered out *before* the overall analysis is conducted

"hither","home","how","howbeit","however","hundred","i","id","ie","if","i'll","im","immediate","immediately”,"importance","important","in","inc","indeed","index","information","instead","into","invention","inward","is","isn't”,"it","itd","it'll","its","itself","i've","j","just","k","keep",”keeps","kept","keys","kg","km","know","known","knows","l","largely","last","lately","later","latter","latterly”,"least","less","lest","let","lets","like","liked","likely","line","little","'ll","look","looking","looks","ltd","m","made”,"mainly","make","makes","many","may","maybe","me","mean","means","meantime","meanwhile","merely","mg”,"might","million","miss","ml","more","moreover","most","mostly","mr","mrs","much","mug","must","my","myself","n”,"na","name","namely","nay","nd","near","nearly","necessarily","necessary","need","needs","neither","never”,"nevertheless","new","next","nine","ninety","no”…..

Stemming

- The process for reducing derived words to their word stem or base

  - connection
  - connections
  - connective ---> connect
  - connected
  - connecting

Lemmatisation

- Lemmatisation involves the process of grouping together different forms of a word to capture a single meaning

  - run
  - runs -> run
  - ran

Misc Preprocessing

- Change all to lower case
  - req = tolower(req)

- Remove all numbers
- Remove punctuation
- Remove white space

Text Corpus

“In linguistics, a corpus (plural corpora) or text corpus is a large and structured set of texts (nowadays usually electronically stored and processed). They are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules within a specific language territory.”

https://en.wikipedia.org/wiki/Text_corpus

Modeling

Which pieces of text are important?

- Create Term Document Matrix
- IDF
- Create N-gram features
- Parts of speech tagging
- Bag of Words Model
- Topic Based Modeling
- Packages for NLP in R and Python

Term Frequency

- How often does a term t_i occur in the document d_j?

$$TF(i,j) = n_{ij}$$

- How often does a term occur in the dataset (normalized for length of document)

$$TF(i,j) = n_{ij} / \sum n_{kj}$$

Term Document Matrix

- “A document-term matrix or term-document matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. In a document-term matrix, rows correspond to documents in the collection and columns correspond to terms.”

D1 = "I like databases"
D2 = "I hate databases",
then the document-term matrix would be:

	I	**like**	**hate**	**databases**
D1	1	1	0	1
D2	1	0	1	1

https://en.wikipedia.org/wiki/Document-term_matrix

Inverse Document Frequency (IDF)

- A measure of sparseness
  - Is this word in lots of different documents? Or is it a unique aspect that defines this document
    - What's cooking: Is this a unique ingredient that defines the recipe or something common

$$IDF(t) = 1 + \log(totalDocuments / NumberContainingt)$$

TFIDF

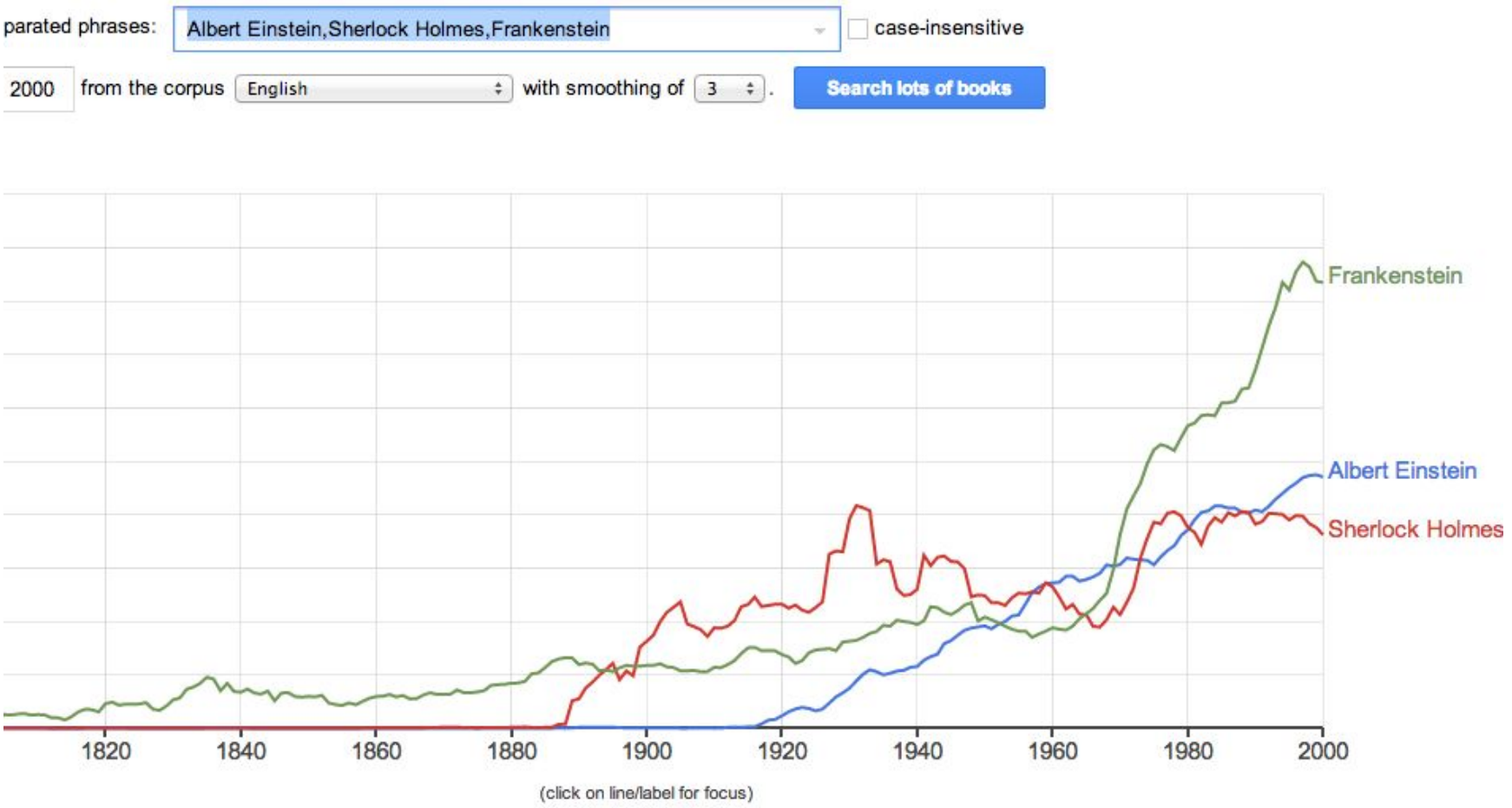
- This combines every frequency in the document and the inverse document frequency

$$TFIDF(t,d) = TF(t,d) x IDF(t)$$

N-Gram

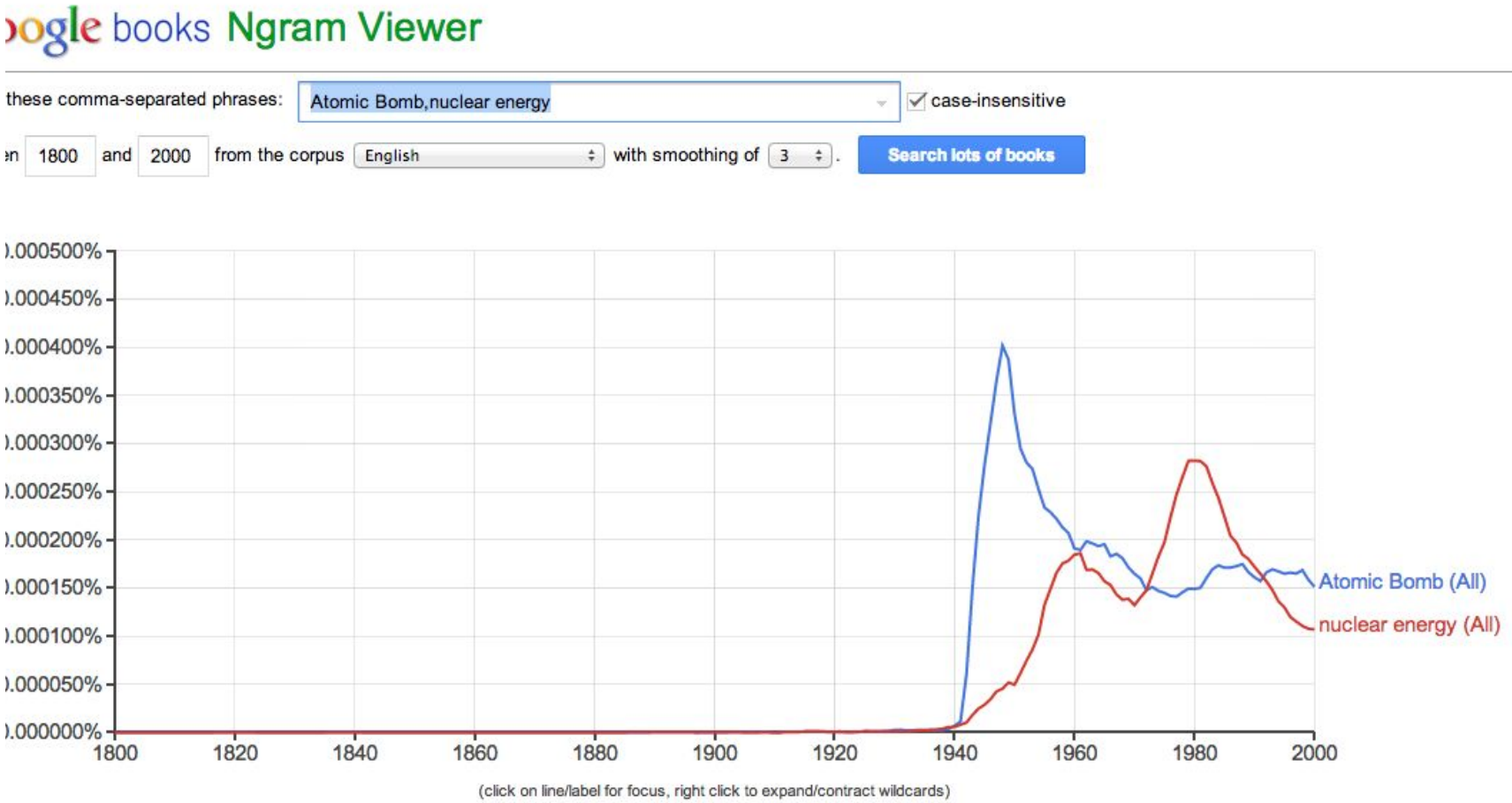
- An n-gram is a sequence of N items from a text corpus.
  - Typically works/terms….but could be syllables, letters, words or base

Google N-Gram Viewer



https://books.google.com/ngrams

Google N-Gram Viewer



https://books.google.com/ngrams

Bag of Words Model

- Documents are bags of words
- Ignore word order or other linguistic structure
- Each word is a feature, so the goal will be to identify relevant words
- Example:
  - We find unique ingredients for the What's cooking that drive selection of the type

Topic Modeling

- There could be a wide variety of topics that may relate to a large number of words
- There are many types of advanced models for this, but it can also be simple

Example: From Random Acts of Pizza, we may want to incorporate a variety of items related to someone's job

Entities and Text Mining

Named-entity Extraction (also known as entity identification, entity chunking and entity extraction) helps locate common entities in a document

https://en.wikipedia.org/wiki/Named-entity_recognition

Parts of Speech Tagging

- Nouns, verbs, adverbs, adjectives, articles, etc.
- Plural vs. Singular
- Tense

Relevant Packages - Python

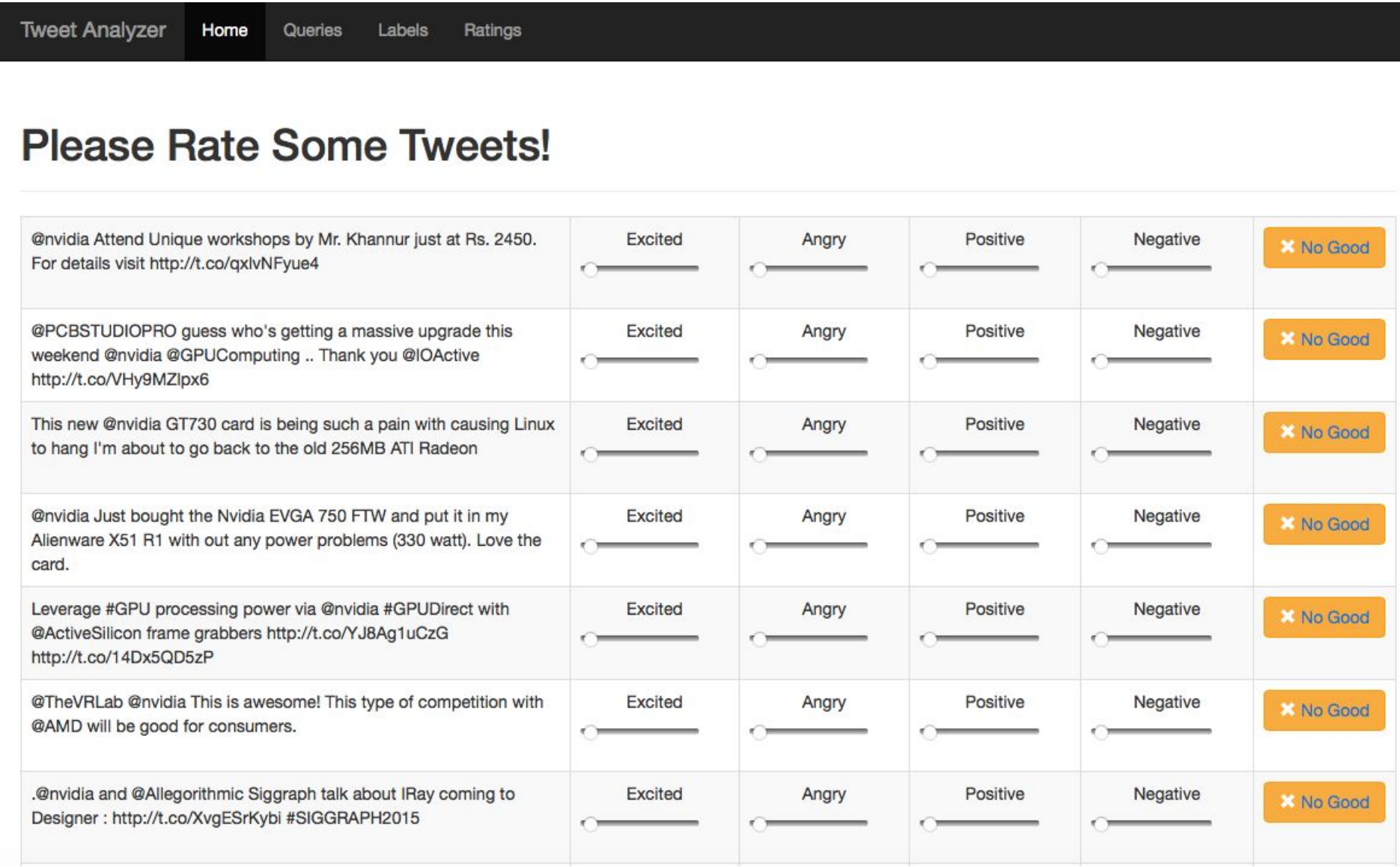
- Natural Language Tool Kit (NLTK)
  - http://www.nltk.org
- Suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning
- Free Book
  - http://www.nltk.org/book_1ed/

Relevant Packages/Commands - R

- NLP
- PLYR (Used for manipulating data)
- gsub

Example: Channel Eyes Tweet Challenge

Tweet Rater



http://ec2-107-22-251-102.compute-1.amazonaws.com/ratings

Tweet Data

tweet	label	rating
#Fortinet : Assigned Patent http://t.co/YlFZCVghtM $FTNT	Excited	44
T http://t.co/mDvVXugFTI #antivirus #appfirewall #av #controleparentals #firewall #firewallapplicatif #fortigate #fortinet #windows2	Excited	58
T http://t.co/mDvVXugFTI #antivirus #appfirewall #av #controleparentals #firewall #firewallapplicatif #fortigate #fortinet #windows2	Angry	57
The Internet of Things could become Internet of Threats in 2015 #IoT #Fortinet - http://t.co/EfOdpVUO5f	Excited	66
#Fortinet : Joins the VMware NSX Partner Ecosystem to Further Advance Security... http://t.co/dKadmaUJdb $FTNT	Angry	58
The Convergence of Virtualization and Security, VMware NSX, and the New FortiGate-VMX http://t.co/0AxW1frtLi #Fortinet #InfoSec	Excited	46
The Convergence of Virtualization and #Security, #VMware NSX, and the New FortiGate-VMX http://t.co/f9l0FDVqHk #Fortinet	Angry	48
The Convergence of Virtualization and #Security, #VMware NSX, and the New FortiGate-VMX http://t.co/f9l0FDVqHk #Fortinet	Excited	41
Sebelumnya dipake sama SCTV.. #fortinet #security #firewall http://t.co/0wpx2QNt1K	Angry	98
Sebelumnya dipake sama SCTV.. #fortinet #security #firewall http://t.co/0wpx2QNt1K	Excited	13

Channel Eyes

- What type of problem is this?
- What type of preprocessing of the data should we do?

Bag of Words Analysis

- Twitter

https://rstudio-pubs-static.s3.amazonaws.com/92510_018db285fda546fcb89b53dd2847b5d4.html#video-4-bag-of-words

Examples Random Acts of Pizza

https://github.com/Runze/pizza

Example
What's Cooking

http://rpi-analytics.github.io/MGMT6963-2015/assets/rmarkdown/lab10_what_is_cooking.html

Example
What's Cooking

https://www.kaggle.com/manuelatadvice/whats-cooking/noname