Text Analytics – Unstructured Data Analysis

Presented by: Derek Kane

Overview of Topics

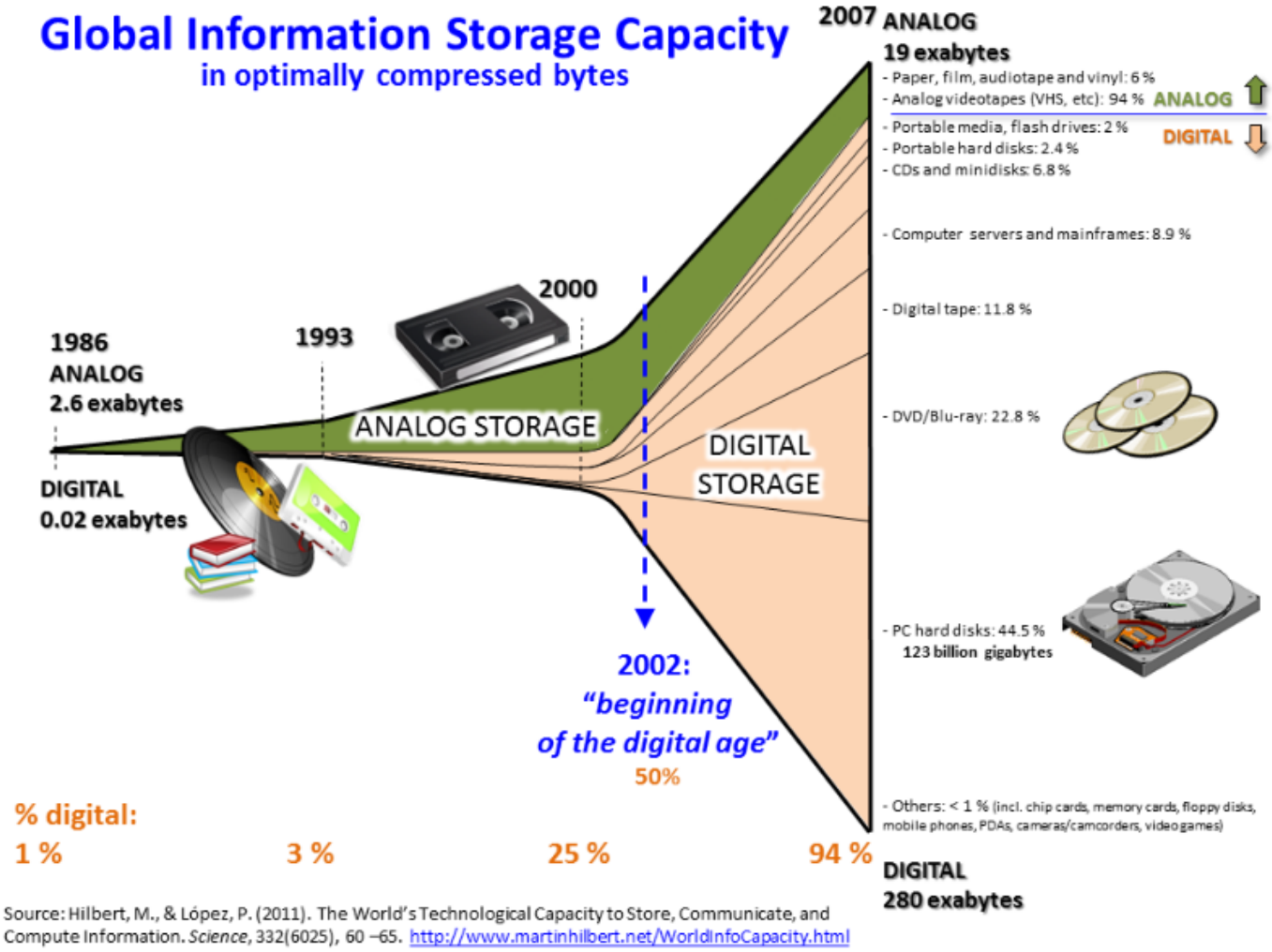
- Data Explosion in the 21st Century
- Working with Unstructured Data
- Introduction to Text Analytics Topics

- Practical Examples
  - Build a search engine
  - Categorization – 2012 Presidential Campaign
  - Categorization – The Blog Writer's Gender
  - Natural Language Processing
  - Clustering Uncategorized RSS Feeds
  - Social Media Sentiment & Network Analysis

Data Explosion in the 21st Century

- Developed economies make increasing use of data-intensive technologies.
- There are 4.6 billion mobile-phone subscriptions worldwide and there are between 1 billion and 2 billion people accessing the internet.
- Between 1990 and 2005, more than 1 billion people worldwide entered the middle class which means more and more people who gain money will become more literate which in turn leads to information growth.

Data Explosion in the 21st Century



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

Data Explosion in the 21st Century

This volume of information can be separated into 2 distinct data types: Structured and Unstructured Data.

Structured – Data is organized in a highly mechanized or manageable manner. Some examples include data tables, OLAP cubes, XML format, etc…

Unstructured – Raw and unorganized data which can be cumbersome and costly to work with. Examples include News Articles, Social Media, Video, Email, etc..

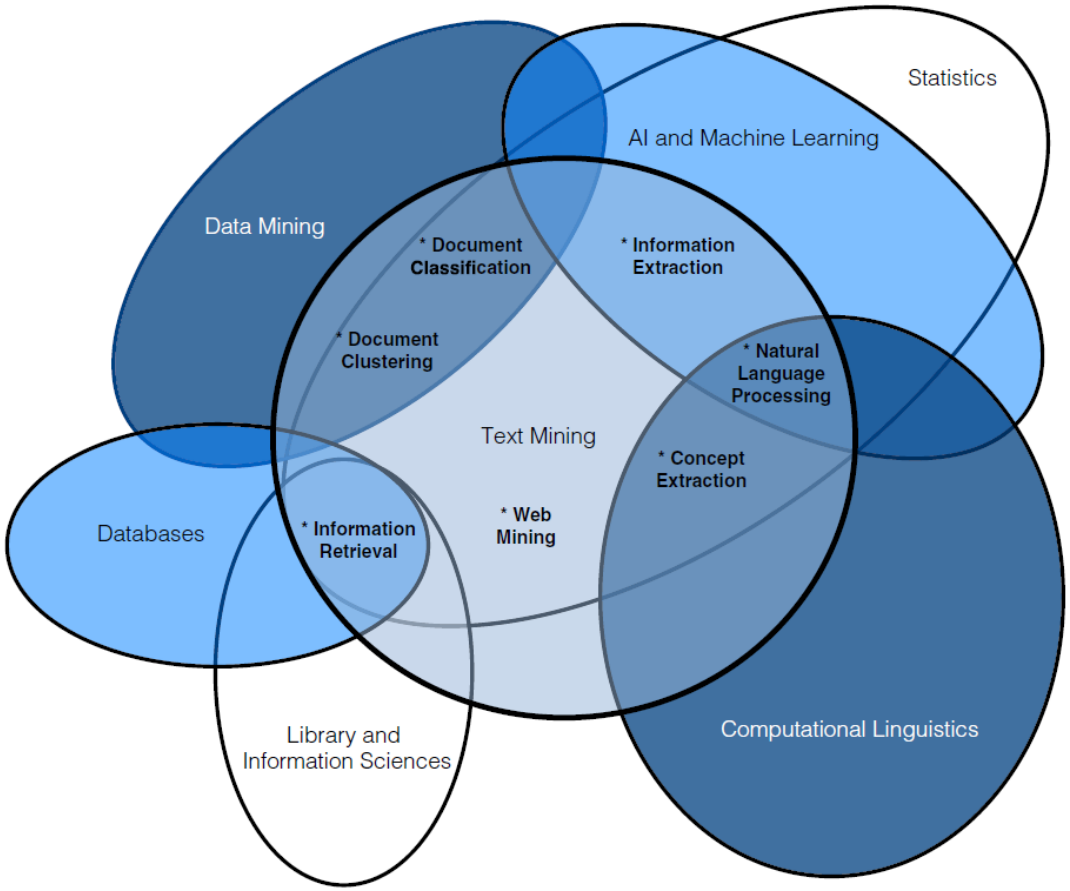
Merrill Lynch projected that 80-90% of all potential usable information exists in unstructured form. In 2010, Computer World claimed that unstructured information might account for 70-80% of all data in an organization.

The Cisco IP traffic forecast projects that video will account for 61% of the total internet data in 2015.

Working With Unstructured Data

- The presentation of data for classical data mining and text data mining is quite different.
- As such, we need to approach text analytics with different techniques in order to bridge the gap between the unstructured and structured data realm.
- The main challenge is to represent the unstructured text correctly in a structured, numerical form.
- Our goal in this presentation is to introduce some techniques that can be employed when mining with unstructured text data to demystify this process.

Text Analytics in the IT Space



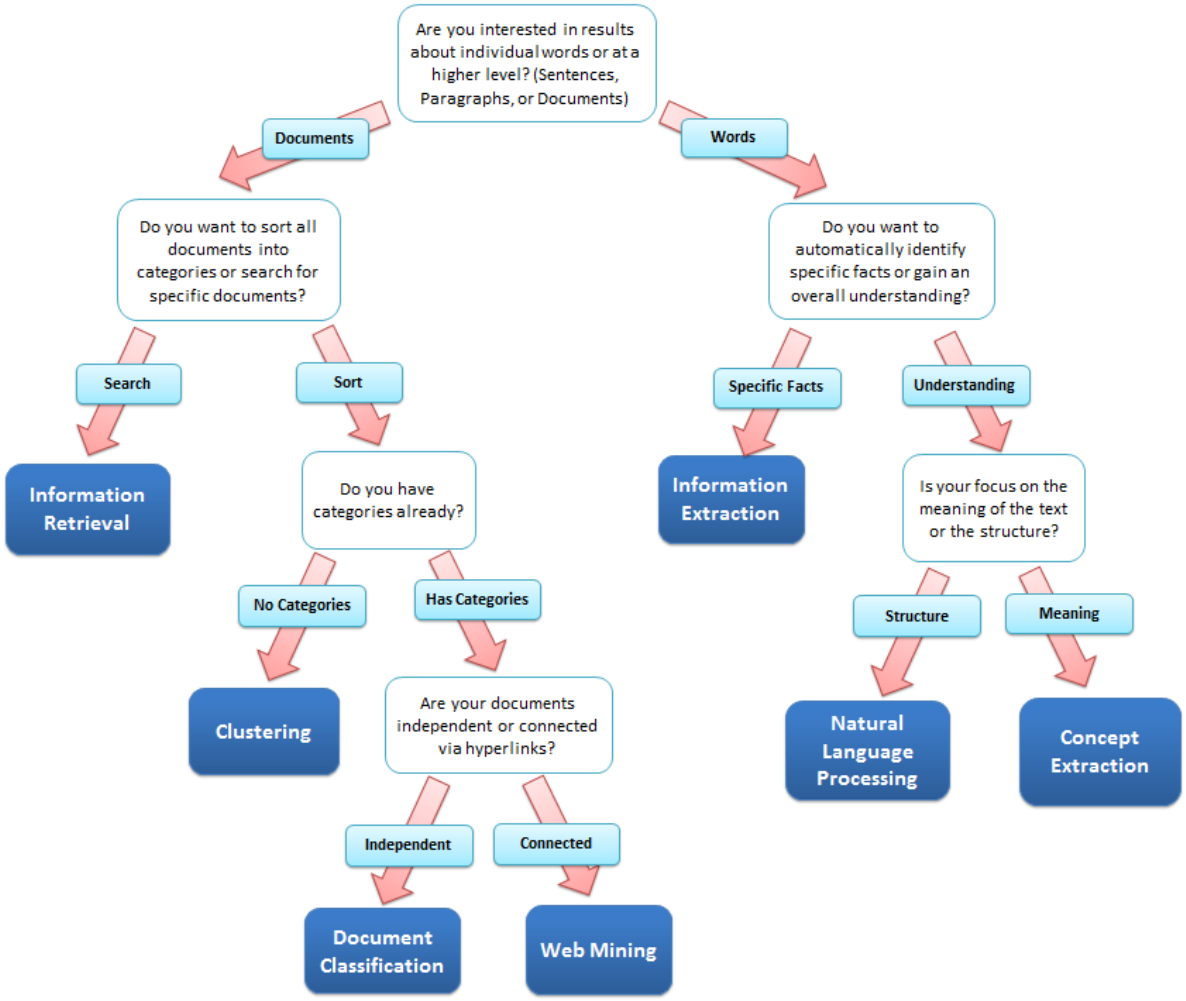
What is Text Analytics?

- Text mining and text analytics are a broad umbrella terms describing a range of technologies for analyzing and processing semi-structured and unstructured text data.
- Text mining is in a loosely organized set of competing technologies that function as analytical "city-states" with no clear dominance among them. This area is something of the Wild West of analytics; there is no clear dominant technological solutions.
- Different areas of text mining are at various levels of developmental maturity which complicates matters when deploying a comprehensive text analytics solution.
- These different areas of text mining can be separated into 7 functional practice areas.

Major Areas of Text Analytics

- Search and Information Retrieval (IR): Storage and retrieval of text documents, including search engines and keyword search.
- Document Clustering: Grouping and categorizing terms, snippets, paragraphs, or documents using data mining clustering methods.
- Document Classification: Grouping and categorizing snippets, paragraphs, or documents using data mining classification methods, trained on labeled examples.
- Web Mining: Data and text mining on the Internet, with a specific focus on the scale and interconnectedness of the web.
- Information Extraction (IE): Identification and extraction of relevant facts and relationships from unstructured text; the process of making structured data from unstructured and semi-structured text.
- Natural Language Processing (NLP): Low-Level language processing and understanding tasks (Ex. tagging part of speech); often used synonymously with computational linguistics.
- Concept Extraction: Grouping of words or phrases into semantically similar groups.

Identifying the Text Mining Task

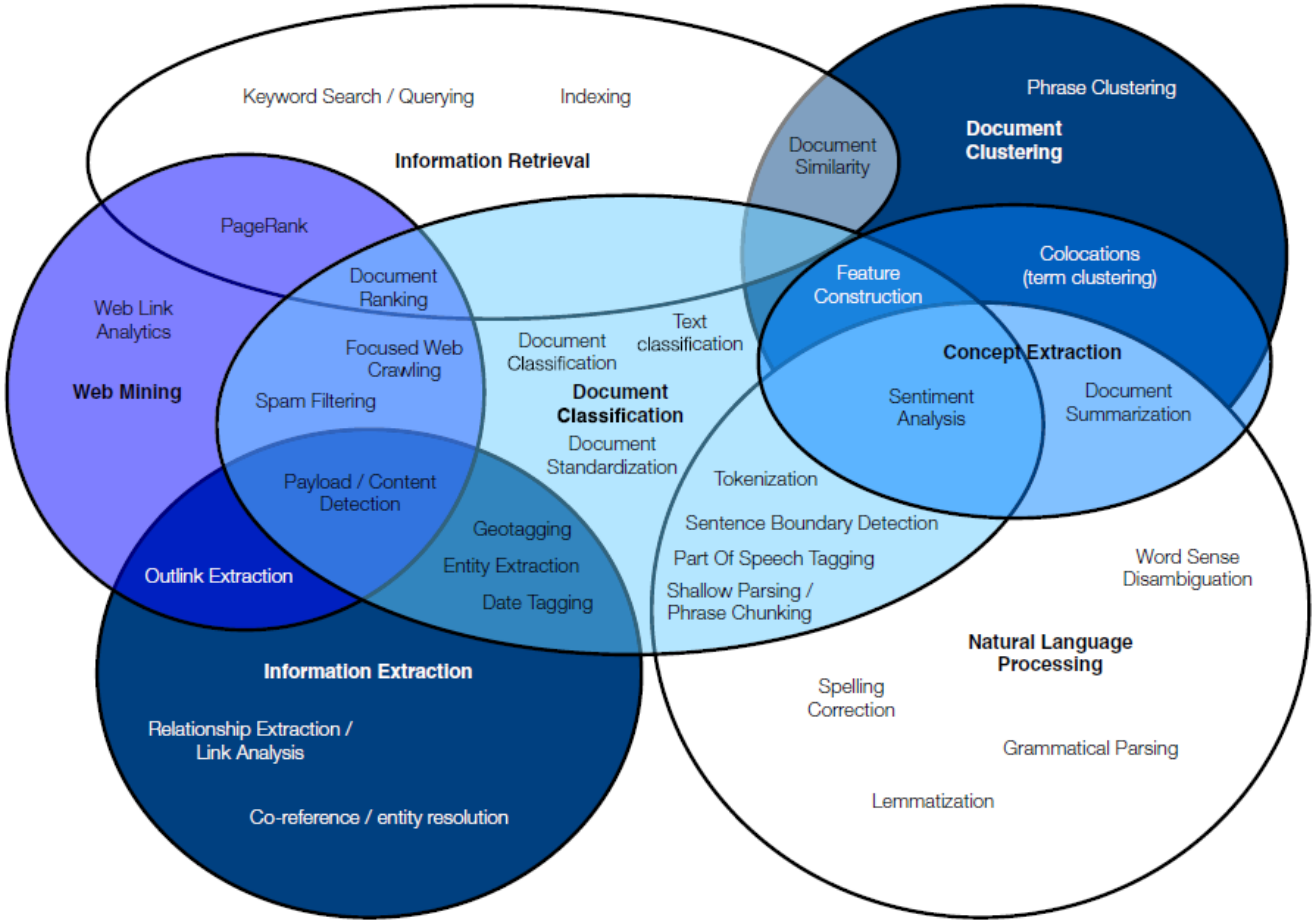


Major Areas of Text Analytics

Finding a Practice Area Based on the Desired Product of Text Mining	
Desired Application	**Practice Area**
Linguistic Structure	Natural Language Processing
Topic / Category Assignment	Document Classification
Documents that match keywo	Information Retrieval
A structured database	Information Extraction
"Needles in a Haystack"	Document Classification
List of synonyms	Concept Extraction
Marked Sentences	Natural Language Processing
Understanding of microblogs	Web Mining
Similar documents	Document Clustering

Text Mining Topics and Related Practice Areas	
Topic	**Practice Area**
Keyword Search	Search and Information Retrieval
Inverted Index	Search and Information Retrieval
Document Clustering	Document Clustering
Document Similarity	Document Clustering
Feature Selection	Document Classification
Sentiment Analysis	Document Classification
Dimensionality Reduction	Document Classification
eDiscovery	Document Classification
Web Crawling	Web Mining
Link Analytics	Web Mining
Entity Extraction	Information Extraction
Link Extraction	Information Extraction
Part of Speech Tagging	Natural Language Processing
Tokenization	Natural Language Processing
Question Answering	Natural Language Processing
Topic Modeling	Concept Extraction
Synonym Identification	Concept Extraction

Major Areas of Text Analytics



Text Analytics Information Retrieval - Search Engine

Search engine

- Web based search engines are the driver of the 21st century information infrastructure. The tech giants of our lifetime have built entire business models worth billions of dollars off of a fundamentally simple concept:

- *How can I find the information I am looking for on the internet?*

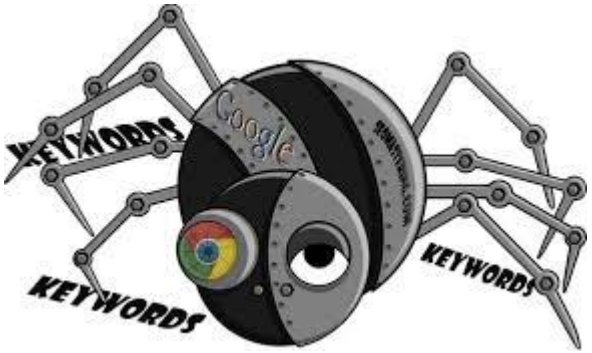
- Of course, there is more to this idea when we speak in business applications. How is the ranking determined? If someone pays google for a higher ranking, how is this incorporated into the results, etc…

Google

YAHOO!

bing

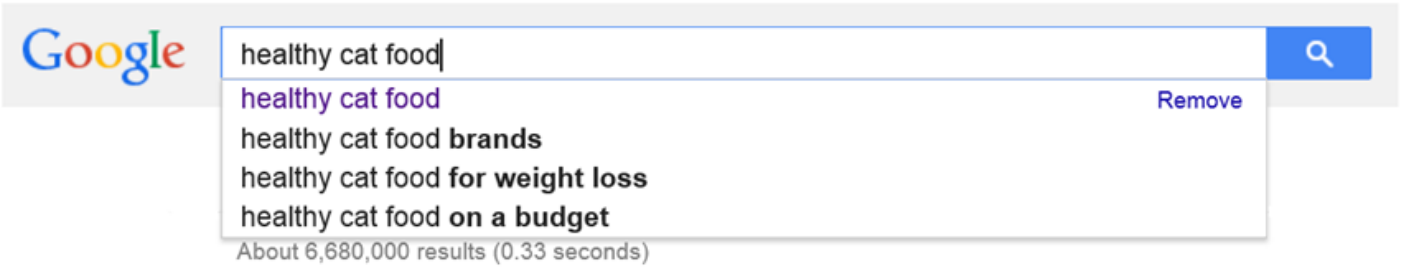
Search engine



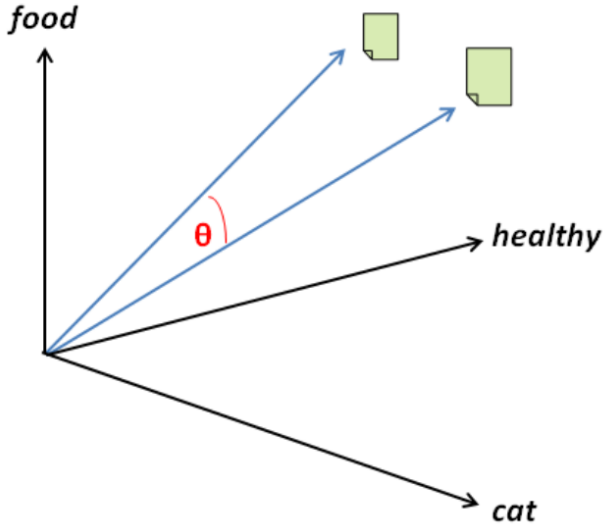
- Search engines are a practical application of text analytics which bridges the gap between unstructured and structured data analytics.
- The basic operating principle is that the search engine provider categorizes the websites (documents) of interest and indexes them using some criterion. We then specify our search parameter and pass this through the search query engine which determines a ranking of results.
- The results with the highest ranking will be the best match based upon our search algorithm.
- For our example, we're going to use a tried and true method for our search algorithm, which origins are from the 1960's. We're going to implement the vector space model of information retrieval in R.

Search Query

- We will build our search engine to find from a group of 7 websites (text documents) the best ranking in descending order.
- We will use the search criteria " healthy cat food" as the query for the analysis.



- A visualization of the vector based space information retrieval model.



Build a Corpus

- We need to first construct a corpus (a collection of texts) using the 7 various websites (documents).

- Here is the example of the unstructured text that has been indexed to apply the query results against.

Web Page	Text Field
1	"Stray cats are running all over the place. I see 10 a day!"
2	"Cats are killers. They kill billions of animals a year."
3	"The best food in Columbus, OH is the North Market."
4	"Brand A is the best tasting cat food around. Your cat will love it."
5	"Buy Brand C cat food for your cat. Brand C makes healthy and happy cats."
6	"The Arnold Classic came to town this weekend. It reminds us to be healthy."
7	"I have nothing to say. In summary, I have told you nothing."

- Most of the documents contain some reference to cats, healthy, or food with the exception of document #7.

- For simplicity sake, we are going to also include the search query “Healthy Cat Food” into the same corpus.

Preparing the Corpus for Analysis

- In order to improve the quality of our search engines results, we will need to first prepare the text data for further analysis.

- This process consists of the following steps:

  - Remove punctuation
  - Lemmatization or stemming of words (root form)
  - Shift terms to lower case
  - Remove any numbers from the text
  - Strip off any unnecessary white space

Preparing the Corpus for Analysis

- Lets take a look at the following text from our search engine.

  Stray cats are running all over the place. I see 10 a day!

- Now lets remove the punctuation.

  Stray cats are running all over the place I see 10 a day

- Stem terms to the root form.

  Stray cat are run all over the place I see 10 a day

Preparing the Corpus for Analysis

- Remove any numbers.

  Stray cat are run all over the place I see a day

- Adjust terms to lower case.

  stray cat are run all over the place i see a day

- Remove any additional white space.

  stray cat are run all over the place i see a day

Create a Term Document Matrix

Term Document Matrix	
A term-document matrix (14 terms, 8 documents)	
Non-/sparse entries	21/91
Sparsity	: 81%
Maximal term length	: 8
Weighting	: term frequency (tf)

Terms	Web Page 1	Web Page 2	Web Page 3	Web Page 4	Web Page 5	Web Page 6	Web Page 7	Query
all	1	0	0	0	0	0	0	0
and	0	0	0	0	1	0	0	0
anim	0	1	0	0	0	0	0	0
are	1	1	0	0	0	0	0	0
arnold	0	0	0	0	0	1	0	0
around	0	0	0	1	0	0	0	0
best	0	0	1	1	0	0	0	0
billion	0	1	0	0	0	0	0	0
brand	0	0	0	1	2	0	0	0
buy	0	0	0	0	1	0	0	0
came	0	0	0	0	0	1	0	0
cat	1	1	0	2	3	0	0	1
classic	0	0	0	0	0	1	0	0
columbus	0	0	1	0	0	0	0	0
classic	0	0	0	0	0	1	0	0
columbus	0	0	1	0	0	0	0	0

This row contains values from the query parameters as well.

Term Document Weights

- The values of in our document matrix are simple term frequencies.
- This is fine, but other heuristics are available. For instance, rather than a linear increase in the term frequency, tf , perhaps sqrt(tf) or log(tf) would provide a more reasonable diminishing returns on word counts within documents.
- Rare words can also get a boost. The word “healthy” appears in only one document, whereas “cat” appears in four. A word's document frequency, df, is the number of documents that contain it, and a natural choice is to weight words inversely proportional to their df’s.
- As with term frequency, we may use logarithms or other transformations to achieve the desired effect.
- Different weighting choices are often made for the query and the documents.

Term Document Weights

- For both the document and the query, we choose the following weights:

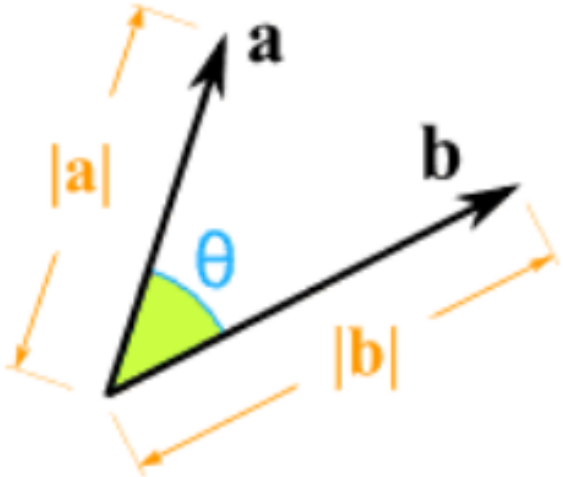
  If tf = 0, then 0, otherwise (1 + Log2(tf)) * Log2(N / df))

- We implement this weighting function across entire rows of the term document matrix, and therefore our weighting function must take a term frequency vector and a document frequency scalar as inputs.

Terms	Web Page 1	Web Page 2	Web Page 3	Web Page 4	Web Page 5	Web Page 6	Web Page 7	Query
cat	1	1	0	2	3	0	0	1

Terms	Weighting 1	Weighting 2	Weighting 3	Weighting 4	Weighting 5	Weighting 6	Weighting 7	Query
cat	0.8073549	0.8073549	0	1.61471	2.086982	0	0	0.807355

Dot Product Geometry



A benefit of being in the vector space is the use of its dot product or scalar product.

- For vectors a and b, the geometric definition of the dot product is:

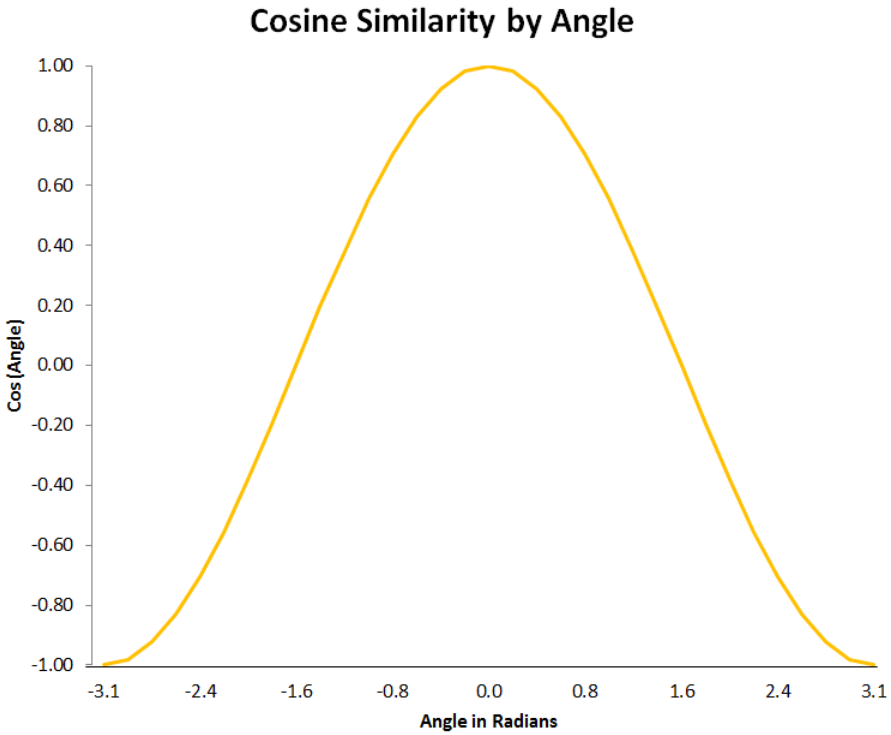
$$a \cdot b = ||a||\,||b|| cos\Theta$$

- where $\cdot$ is the Euclidean norm (the root sum of squares) and Θ is the angle between a and b.

Further Normalization

In fact, we can work directly with the cosine of Θ.

- For theta in the interval [-π, - π] , the endpoints are orthogonally (totally unrelated documents) and the center, zero, is complete collinear (maximally similar documents).
- We can see that the cosine decreases from its maximum value of 1.0 as the angle departs from zero in either direction.



- We may furthermore normalize each column vector in our matrix so that its norm is one.
- Now the dot product is cos Θ.

Terms	Weighting 1	Weighting 2	Weighting 3	Weighting 4	Weighting 5	Weighting 6	Weighting 7	Query
cat	0.1044566	0.1128249	0	0.2378746	0.22591472	0	0	0.347026

Matrix Multiplication

- Keeping the query alongside the other documents let us avoid repeating the same steps.
- But now it's time to pretend it was never there.

```
query.vector <- tfidf.matrix[, (N.docs + 1)]
tfidf.matrix <- tfidf.matrix[, 1:N.docs]
```

- With the query vector and the set of document vectors in hand, it is time to go after the cosine similarities. These are simple dot products as our vectors have been normalized to unit length.

- Recall that matrix multiplication is really just a sequence of vector dot products. The matrix operation below returns values of cosine Θ for each document vector and the query vector.

```
doc.scores <- t(query.vector) %*% tfidf.matrix
```

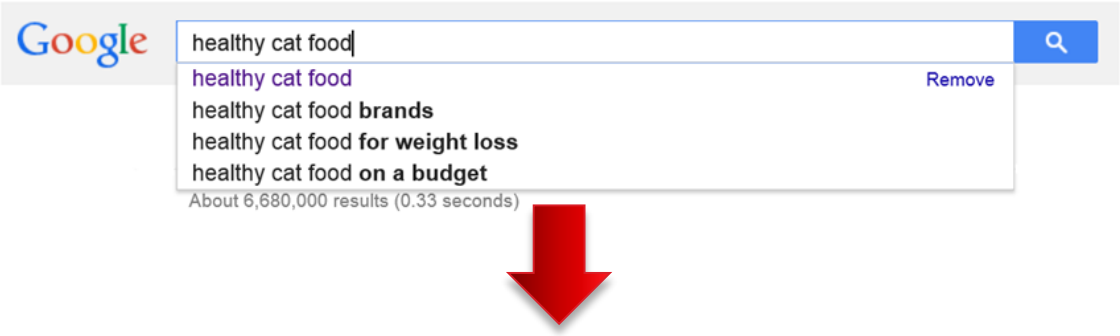
Matrix Multiplication

- With scores in hand, rank the documents by their cosine similarities with the query vector.

```
results.df <- data.frame(doc = names(doc.list), score = t(doc.scores),
                         text = unlist(doc.list))
results.df <- results.df[order(results.df$score, decreasing = TRUE), ]
```

$$\underbrace{\begin{bmatrix}1 & 2 & 3\end{bmatrix}}_{1\times 3} \cdot \underbrace{\begin{bmatrix}2 & 1 & 3\\ 3 & 3 & 2\\ 4 & 1 & 2\end{bmatrix}}_{3\times 3} = \begin{bmatrix}1\cdot 2+2\cdot 3+3\cdot 4\\ 1\cdot 1+2\cdot 3+3\cdot 1\\ 1\cdot 3+2\cdot 2+3\cdot 2\end{bmatrix} = \underbrace{\begin{bmatrix}20\\ 10\\ 13\end{bmatrix}}_{1\times 3}$$

Search Engine Results



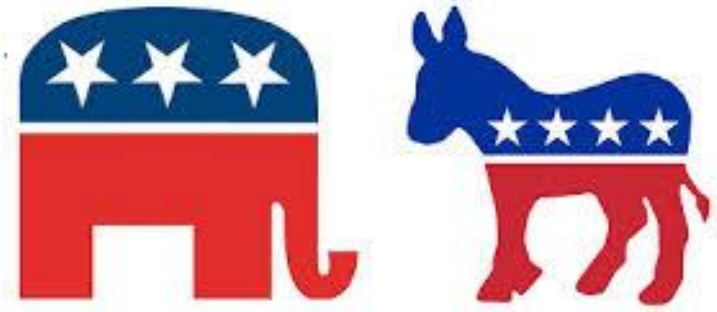
Web Page	Score	Text Field
5	0.344	Buy Brand C cat food for your cat. Brand C makes healthy and happy cats.
6	0.183	The Arnold Classic came to town this weekend. It reminds us to be healthy.
4	0.177	Brand A is the best tasting cat food around. Your cat will love it.
3	0.115	The best food in Columbus, OH is the North Market.
2	0.039	Cats are killers. They kill billions of animals a year.
1	0.036	Stray cats are running all over the place. I see 10 a day!
7	0.000	I have nothing to say. In summary, I have told you nothing.

- Our "best" document, at least in an intuitive sense, comes out ahead with a score nearly twice as high as its nearest competitor.
- Notice however that this next competitor has nothing to do with cats.
- This is due to the relative rareness of the word "healthy" in the documents and our choice to incorporate the inverse document frequency weighting for both documents and query.
- Fortunately, the profoundly uninformative document 7 has been ranked dead last.

Text Analytics Categorization Part I – 2012 Presidential Campaign

2012 Presidential Campaign

- Each of Obama & Romney's 2012 campaign speeches have been recorded and translated into digital text documents.
- These documents will be evaluated using text analytical techniques to determine if there are any patterns in their speech which can be used for categorization.
- The predictive analytics should allow for us to apply an algorithm on an unknown speech document and determine whether it is Romney or Obama's speech.
- This classification technique has applications in email spam detection, fraud, stock prediction from articles, and others.



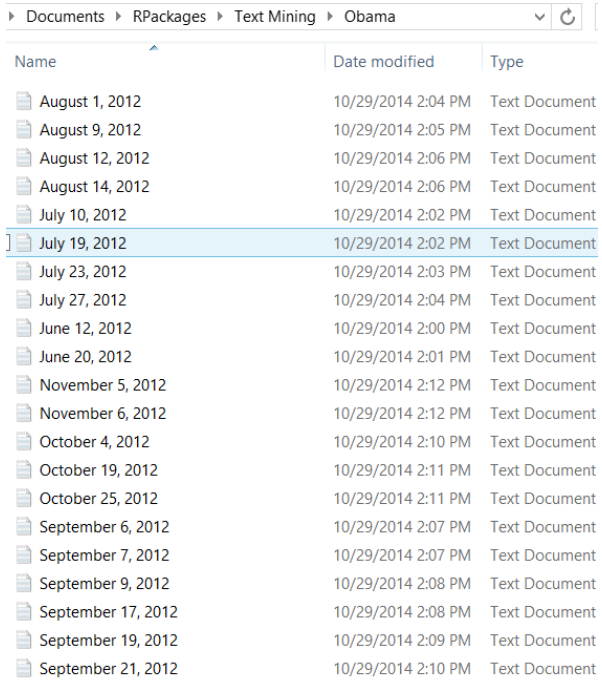
2012 Presidential Campaign

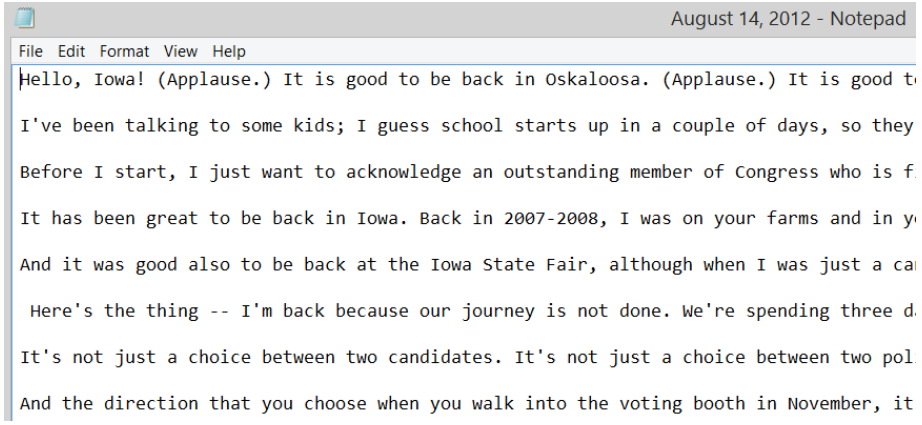
Here is the outline for the modeling task:

- Initialize the directory and files
- Clean the Text
- Build the Term Document Matrix (TDM)
- Create a holdout sample. (Training and Test Set)
- Create the KNN model.
- Assess the models performance.

2012 Presidential Campaign

- All of the speech documents are located in separate folders for Obama and Romney.
- A transcription of the text has been inserted into the .txt document with no additional scrubbing.





2012 Presidential Campaign

- Each of the folders will be processed converting the unstructured text data into a structured table with each row relating to a specific speech.

Text	Candidate
A couple of people I want to acknowledge -- first of all, please give a huge round of applause to Brenda for that great introduction. (Applause.) And go try some of her pizza if you have not tried it. (Laughter.) You got a testimony right here -- he says it's outstanding. I want to thank her so much for doing this.	Obama
There are a couple of people I want to acknowledge. First of all, DJ Vince Adams, thank you so much. (Applause.) DJ Cassidy, thank you so much. My great friend, Kal Penn, thank you for all that you do. And everybody on the Gen 44 host committee, thank you for the great job you guys did. (Applause.)	Obama
Before I start, I just want to acknowledge an outstanding member of Congress who is fighting every day on behalf of the people of his district -- Dave Loebsack is here. Give Dave a big round of applause. (Applause.) There he is. Thank you, Dave.	Obama
Now, if you guys have a seat, feel free to take a seat. That way, if it gets a little warm, I don't want anybody getting overheated. You guys are kind of out of luck. (Laughter.) So make sure you're hydrated.	Obama

Text	Candidate
In a city that is far too often characterized by pettiness and personal attacks, Paul Ryan is a shining exception. He does not demonize his opponents. He understands that honorable people can have honest differences. And he appeals to the better angels of our nature. There are a lot of people in the other party who might disagree with Paul Ryan; I don't know of anyone who doesn't respect his character and	Romney
The people I met on this tour – and the thousands of Americans I've visited in break rooms and lunch rooms, in school gymnasiums and on factory floors – are worried about their children, their jobs, their mortgages, and their future. And they are right to be worried.	Romney
Anyone who knows about the American Legion understands this is much more than an organization of veterans. Every day, you seek and you find new ways to give back to the country you love. From American Legion Baseball to the Child Welfare Foundation, your achievements are many, significant, and deeply appreciated.	Romney
Tonight I am asking you to join me to walk together to a better future. By my side, I have chosen a man with a big heart from a small town. He represents the best of America, a man who will always make us proud – my friend and America's next Vice President, Paul Ryan.	Romney

2012 Presidential Campaign

- The 2 datasets will be merged into a combined master dataset.

Text	Candidate
A couple of people I want to acknowledge -- first of all, please give a huge round of applause to Brenda for that great introduction. (Applause.) And go try some of her pizza if you have not tried it. (Laughter.) You got a testimony right here -- he says it's outstanding. I want to thank her so much for doing this.	Obama
There are a couple of people I want to acknowledge. First of all, DJ Vince Adams, thank you so much. (Applause.) DJ Cassidy, thank you so much. My great friend, Kal Penn, thank you for all that you do. And everybody on the Gen 44 host committee, thank you for the great job you guys did. (Applause.)	Obama
Before I start, I just want to acknowledge an outstanding member of Congress who is fighting every day on behalf of the people of his district -- Dave Loebsack is here. Give Dave a big round of applause. (Applause.) There he is. Thank you, Dave.	Obama
Now, if you guys have a seat, feel free to take a seat. That way, if it gets a little warm, I don't want anybody getting overheated. You guys are kind of out of luck. (Laughter.) So make sure you're hydrated.	Obama

Text	Candidate
In a city that is far too often characterized by pettiness and personal attacks, Paul Ryan is a shining exception. He does not demonize his opponents. He understands that honorable people can have honest differences. And he appeals to the better angels of our nature. There are a lot of people in the other party who might disagree with Paul Ryan; I don't know of anyone who doesn't respect his character and	Romney
The people I met on this tour – and the thousands of Americans I've visited in break rooms and lunch rooms, in school gymnasiums and on factory floors – are worried about their children, their jobs, their mortgages, and their future. And they are right to be worried.	Romney
Anyone who knows about the American Legion understands this is much more than an organization of veterans. Every day, you seek and you find new ways to give back to the country you love. From American Legion Baseball to the Child Welfare Foundation, your achievements are many, significant, and deeply appreciated.	Romney
Tonight I am asking you to join me to walk together to a better future. By my side, I have chosen a man with a big heart from a small town. He represents the best of America, a man who will always make us proud – my friend and America's next Vice President, Paul Ryan.	Romney

Text	Candidate
round of applause to Brenda for that great introduction. (Applause.) And go try some of her pizza if you have not tried it. (Laughter.)'You got a testimony right here -- he says it's outstanding. I want to thank her so much for doing this.	Obama
Adams, thank you so much. (Applause.) DJ Cassidy, thank you so much. My great friend, Kal Penn, thank you for all that you do. And everybody on the Gen 44 host committee, thank you for the great job you guys did. (Applause.)	Obama
Before I start, I just want to acknowledge an outstanding member of Congress who is fighting every day on behalf of the people of his district -- Dave Loebsack is here. Give Dave a big round of applause. (Applause.) There he is. Thank you, Dave.	Obama
Now, if you guys have a seat, feel free to take a seat. That way, if it gets a little warm, I don't want anybody getting overheated. You guys are kind of out of luck. (Laughter.) So make sure you're hydrated.	Obama
In a city that is far too often characterized by pettiness and personal attacks, Paul Ryan is a shining exception. He does not demonize his opponents. He understands that honorable people can have honest differences. And he appeals to the better angels of our nature. There are a lot of people in the other party who might disagree with Paul Ryan; I don't know of anyone who doesn't respect his character and judgment.	Romney
The people I met on this tour – and the thousands of Americans I've visited in break rooms and lunch rooms, in school gymnasiums and on factory floors – are worried about their children, their jobs, their mortgages, and their future. And they are right to be worried.	Romney
Anyone who knows about the American Legion understands this is much more than an organization of veterans. Every day, you seek and you find new ways to give back to the country you love. From American Legion Baseball to the Child Welfare Foundation, your achievements are many, significant, and deeply appreciated.	Romney
Tonight I am asking you to join me to walk together to a better future. By my side, I have chosen a man with a big heart from a small town. He represents the best of America, a man who will always make us proud – my friend and America's next Vice President, Paul Ryan.	Romney

Build a Loop for Processing the Corpus

- We need to first construct a corpus (a collection of texts) using the combined dataset.
- Then we will apply a looping function for the data preparation and apply the cleaning function to each speech and for both candidates.

```
cleanCorpus <- function(corpus) {
  corpus.tmp <- tm_map(corpus, removePunctuation)
  corpus.tmp <- tm_map(corpus.tmp, stripWhitespace)
  corpus.tmp <- tm_map(corpus.tmp, tolower)
  corpus.tmp <- tm_map(corpus.tmp, removeWords, stopwords("english"))
  return(corpus.tmp)
}
```

```
generateTDM <- function(candidates, pathname){
  s.dir <- sprintf("%s/%s", pathname, candidates)
  s.cor <- Corpus(DirSource(directory = s.dir, encoding = "ANSI"))
  s.cor.cl <- cleanCorpus(s.cor)
  s.tdm <- TermDocumentMatrix(s.cor.cl)
  s.tdm <- removeSparseTerms(s.tdm, 0.7)
  result <- list(name = candidates, tdm = s.tdm)
}
```

Create a Term Document Matrix

Document	Candidate	ability	achieve	across	act	advantage	afghanistan	agenda	always	america
August 1, 2012	Romney	1	1	0	1	0	0	1	0	7
August 12, 2012	Romney	0	1	2	0	1	0	3	0	19
August 14, 2012	Romney	3	1	2	1	0	1	0	0	11
August 9, 2012	Romney	0	0	0	0	1	0	0	8	53
July 10, 2012	Romney	0	0	2	0	2	0	1	2	7
July 19, 2012	Romney	2	0	1	0	2	3	0	2	17
July 23, 2012	Romney	1	1	0	0	0	1	0	1	8
July 27, 2012	Romney	0	1	2	1	0	2	0	0	5
June 12, 2012	Romney	1	0	3	0	0	0	1	1	29
June 20, 2012	Romney	0	0	1	1	0	0	0	1	18
November 5, 2012	Romney	0	1	4	0	0	0	0	0	12
November 6, 2012	Romney	0	1	6	5	1	0	1	0	9
October 19, 2012	Romney	0	0	2	0	0	0	0	0	5
August 1, 2012	Obama	0	0	1	1	0	0	0	2	12
August 12, 2012	Obama	0	0	3	3	0	1	0	5	20
August 14, 2012	Obama	0	0	9	0	0	1	0	1	27
August 9, 2012	Obama	0	0	4	2	0	1	0	0	12
July 10, 2012	Obama	0	0	5	1	0	1	0	4	15
July 19, 2012	Obama	0	0	7	3	0	1	0	4	16
July 23, 2012	Obama	0	0	3	2	0	6	0	3	20
July 27, 2012	Obama	0	0	0	2	0	1	0	3	17
June 12, 2012	Obama	0	0	7	0	0	1	0	3	12
June 20, 2012	Obama	0	0	1	0	0	1	0	5	12
November 5, 2012	Obama	0	0	3	1	0	1	0	9	8
November 6, 2012	Obama	0	0	0	0	0	0	0	4	15

- The candidate field shows the specific candidate.
- There are 1,330 terms to the right and a frequency count of each terms occurrence.
- This list of terms was pared down using a sparsity threshold parameter of 0.7.

Data Preparation for Prediction

- We will split the dataset into a training and testing sample. We will sample 70% for our training data and 30% for the validation test.

- Training Dataset

Candidate	ability	achieve	across	act	advantage	afghanistan	agenda	always	america
Romney	1	1	0	1	0	0	1	0	7
Romney	0	1	2	0	1	0	3	0	19
Romney	3	1	2	1	0	1	0	0	11
Romney	0	0	0	0	1	0	0	8	53
Romney	0	0	2	0	2	0	1	2	7
Romney	2	0	1	0	2	3	0	2	17
Romney	1	1	0	0	0	1	0	1	8
Romney	0	1	2	1	0	2	0	0	5
Romney	1	0	3	0	0	0	1	1	29
Obama	0	0	1	1	0	0	0	2	12
Obama	0	0	3	3	0	1	0	5	20
Obama	0	0	9	0	0	1	0	1	27
Obama	0	0	4	2	0	1	0	0	12
Obama	0	0	5	1	0	1	0	4	15
Obama	0	0	7	3	0	1	0	4	16
Obama	0	0	3	2	0	6	0	3	20

- Testing Dataset

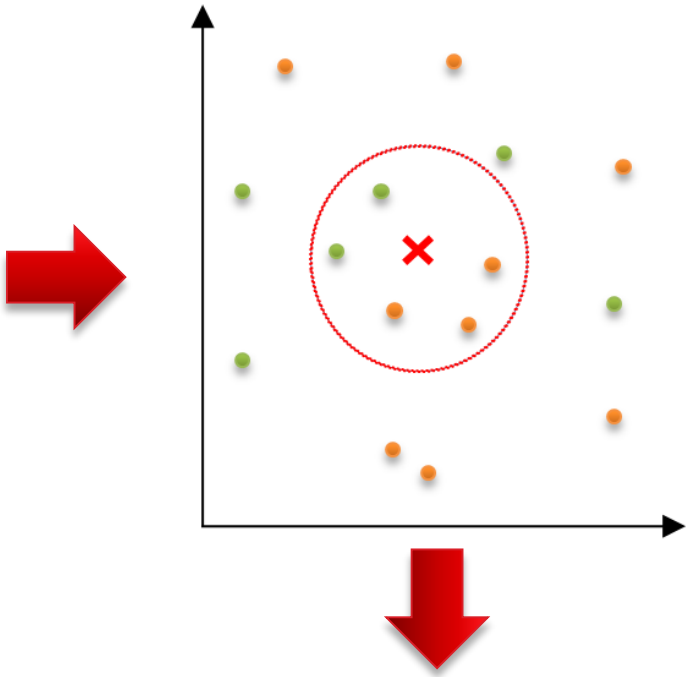
Speech	ability	achieve	across	act	advantage	afghanistan	agenda	always	america
Speech 1	0	0	1	1	0	0	0	1	18
Speech 2	0	1	4	0	0	0	0	0	12
Speech 3	0	1	6	5	1	0	1	0	9
Speech 4	0	0	2	0	0	0	0	0	5
Speech 5	0	0	0	2	0	1	0	3	17
Speech 6	0	0	7	0	0	1	0	3	12
Speech 7	0	0	1	0	0	1	0	5	12
Speech 8	0	0	3	1	0	1	0	9	8
Speech 9	0	0	0	0	0	0	0	4	15

- The Candidate column has been removed from the testing set because we will be predicting the candidate based off of the speech terms.

Predicting the Candidates

- The testing dataset was passed through a kNN classification algorithm.

Speech	ability	achieve	across	act	advantage	afghanistan	agenda	always	america
Speech 1	0	0	1	1	0	0	0	1	18
Speech 2	0	1	4	0	0	0	0	0	12
Speech 3	0	1	6	5	1	0	1	0	9
Speech 4	0	0	2	0	0	0	0	0	5
Speech 5	0	0	0	2	0	1	0	3	17
Speech 6	0	0	7	0	0	1	0	3	12
Speech 7	0	0	1	0	0	1	0	5	12
Speech 8	0	0	3	1	0	1	0	9	8
Speech 9	0	0	0	0	0	0	0	4	15



Prediction Accuracy:
91.2%

Speech	Actual	Predicted
Speech 1	Romney	Romney
Speech 2	Obama	Obama
Speech 3	Romney	Romney
Speech 4	Romney	Obama
Speech 5	Obama	Obama
Speech 6	Romney	Romney
Speech 7	Obama	Obama
Speech 8	Obama	Obama
Speech 9	Obama	Obama

Text Analytics Categorization Part II – The Blog Writer's Gender

Gender Analysis from Blog Entries

- One appealing aspect of text analytics is the ability to create a profile of characteristics for individuals related to our written and spoken text.
- We have a compiled list of list of blog entries where the gender of the person is known and we would like predict the gender of blogs where the gender is unknown.
- Understanding characteristics/ demographics of individuals has a wide variety of applications in terrorism detection, counterintelligence, education, business, and fraud.

Gender Analysis - Blog

- Each of the blogs will be processed converting the unstructured text data into a structured table with each row relating to a specific blog.

Text	Gender
I'm back from vacation, and still digging my way out of everything that's piled up while I've been offline. While I catch up, I thought I'd share with you a demo that Eric Iverson was gracious enough to share with me. It uses Yahoo! BOSS to support an exploratory search experience on top of a general web search engine. When you perform a query, the application retrieves a set of related term candidates using Yahoo's key terms API. It then scores each term by dividing its occurrence count within the result set by its global occurrence count–a relevance measure similar to one my former colleagues and I used at Endeca in enterprise contexts. You can try out the demo yourself at http://www.ittybittysearch.com/. While it has rough edges, it produces nice results–especially considering the simplicity of the approach. Here's an example of how I used the application to explore and learn something new. I started with ["information retrieval"]. I noticed "interactive information retrieval" as a top term, so I used it to refine. Most of the refinement suggestions looked familiar to me–but an unfamiliar name caught my attention: "Anton Leuski". Following my curiosity, I refined again. Looking at the results, I immediately saw that Leuski had done work on evaluating document clustering for interactive information retrieval. Further exploration made it clear this is someone whose work I should get to know–check out his home page! I can't promise that you'll have as productive an experience as I did, but I encourage you to try Eric's demo. It's simple examples like these that remind me of the value of pursuing HCIR for the open web.	Male
Who moved my Cheese??? The world has been developing in and out in all the areas and to create a difference in this competitive world... we need to change... change the way we take our things... but we rather change or atleast try to change..... we try the same routine work evryday and expect to get more and when things fail such as losing a job, loss in business we would upset, discouraged, frustrated and keep on hanging to the same thing again and start complaining. CHANGE IS GOOD.... LETS WELCOME IT..!!! wondering wat is all about Cheese?? and what actually is all about " Who Moved My Cheese??? " Well...!!!! Who moved my cheese?? is a simple parable that reveals profound thoughts. It is an enlightening story of four characters who live in a maze and look for cheese to nourish them and make them happy. The story is about two mice called "SNIFF" and "SCURRY" and two little men ... smaller in size and who were similar to us people. Their names were "HEM" and "HAW" Cheese is a metaphor for what you want to have in life - whether it is a good job, loving relationship, money or a possession, health or spiritual peace of mind.And the maze is where you look for what you want - the organisation you work in, or the family or community you live in. Everyday both the mice and men spent time in the maze looking for their own special cheese. The mice had only Rodent brains while the men used their brains, filled with many beliefs. The common thing between the rodents and these men is tat every morning they went in search for the cheese.	Female

Build a Loop for Processing the Corpus

- We need to first construct a corpus (a collection of texts) using the dataset.
- Then we will apply a looping function for the data preparation and apply the cleaning function to blog entry.

```
#################################################################################
# Corpus Cleanup
#################################################################################

Corpus.mydata <- tm_map(Corpus.mydata, removeNumbers)
Corpus.mydata <- tm_map(Corpus.mydata, tolower)
Corpus.mydata <- tm_map(Corpus.mydata, removeWords, stopwords("english"))
Corpus.mydata <- tm_map(Corpus.mydata, removePunctuation)
Corpus.mydata <- tm_map(Corpus.mydata, stripWhitespace)
```

```
# Matrix with columns as the terms and rows as the documents.
Corpus.TDM <- DocumentTermMatrix(Corpus.mydata)

# Remove Sparse Terms from Corpus.TDM
Corpus.TDM <- removeSparseTerms(Corpus.TDM, 0.95)
```

Create a Term Document Matrix

Blogger	Gender	already	also	although	always	amazing	another	anyone	anything
1	Male	0	0	0	1	0	0	0	0
2	Male	0	0	0	0	0	0	0	0
3	Male	1	2	0	0	0	0	0	0
4	Male	0	1	0	0	0	0	0	0
5	Female	0	0	0	2	0	0	0	0
6	Male	0	0	1	1	1	0	0	0
7	Female	0	0	0	0	0	2	0	0
8	Male	2	0	0	0	0	0	0	3
9	Female	0	0	0	1	0	0	0	0
10	Female	1	0	0	0	0	0	0	0
11	Male	1	2	0	0	0	1	0	0
12	Male	1	3	0	0	0	2	0	1
13	Female	2	1	0	0	1	0	0	1
14	Male	0	3	0	0	0	0	0	0
15	Male	0	5	0	1	0	0	1	1
16	Female	0	0	0	0	0	1	0	0
17	Female	0	0	0	0	0	0	0	0
18	Male	0	0	1	0	1	0	0	0
19	Male	0	1	0	0	0	0	0	1
20	Female	0	0	0	2	0	0	0	0
21	Male	0	1	0	0	0	0	0	0
22	Female	0	1	0	0	0	0	0	0
23	Male	0	0	0	0	0	0	0	2
24	Male	0	0	0	0	0	0	0	0
25	Female	0	0	0	0	0	0	0	0

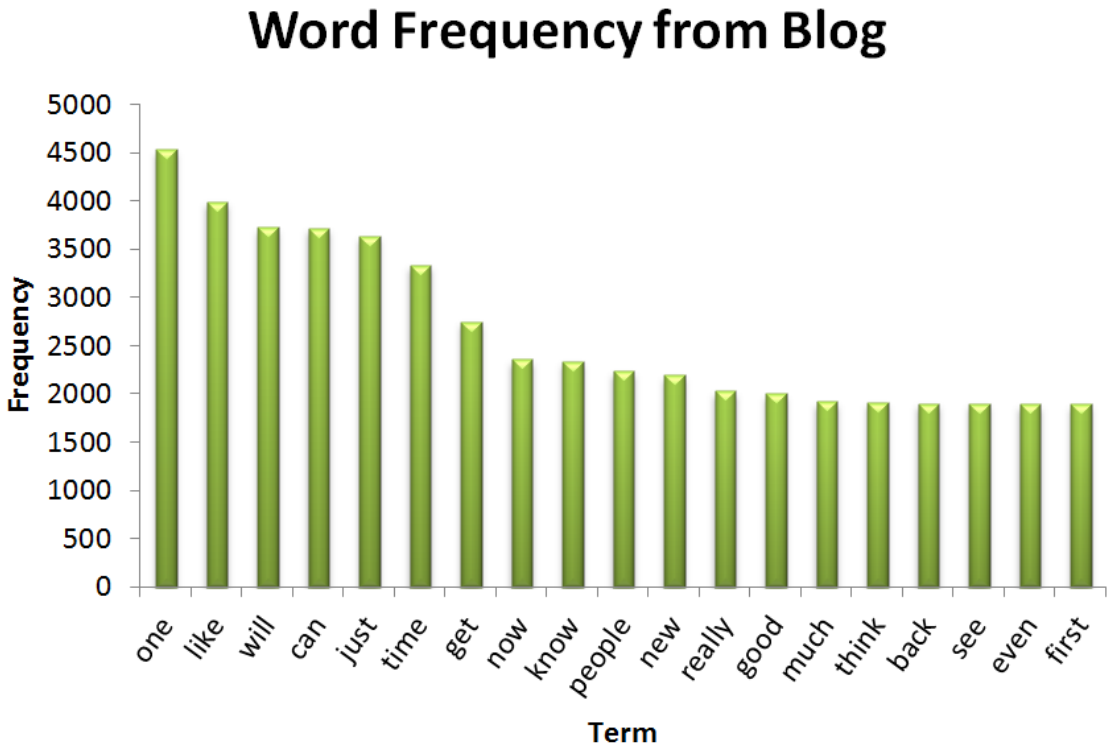
- The Gender Field contains the known gender of each blogger.
- There are 430 terms to the right and a frequency count of each terms occurrence.
- This list of terms was pared down using a sparsity threshold parameter of 0.95.

Associations & Frequent Terms

- We can gain some insight and intelligence by understanding frequencies and associations of terms.

- **Example:** Which words are associated with the term "company" ?

Search Term: Company	
Related	**Association**
order	0.76
person	0.58
case	0.45
business	0.41
may	0.41
without	0.39
right	0.31
without	0.39
right	0.31



Data Preparation for Prediction

- We will split the dataset into a training and testing sample. We will sample 70% for our training data and 30% for the validation test.

- Training Dataset

Gender	already	also	although	always	amazing	another	anyone	anything
Male	0	0	0	1	0	0	0	0
Male	0	0	0	0	0	0	0	0
Male	1	2	0	0	0	0	0	0
Male	0	1	0	0	0	0	0	0
Female	0	0	0	2	0	0	0	0
Male	0	0	1	1	1	0	0	0
Female	0	0	0	0	0	2	0	0
Male	2	0	0	0	0	0	0	3
Female	0	0	0	1	0	0	0	0
Female	1	0	0	0	0	0	0	0
Male	1	2	0	0	0	1	0	0
Male	1	3	0	0	0	2	0	1
Female	2	1	0	0	1	0	0	1
Male	0	3	0	0	0	0	0	0
Male	0	5	0	1	0	0	1	1

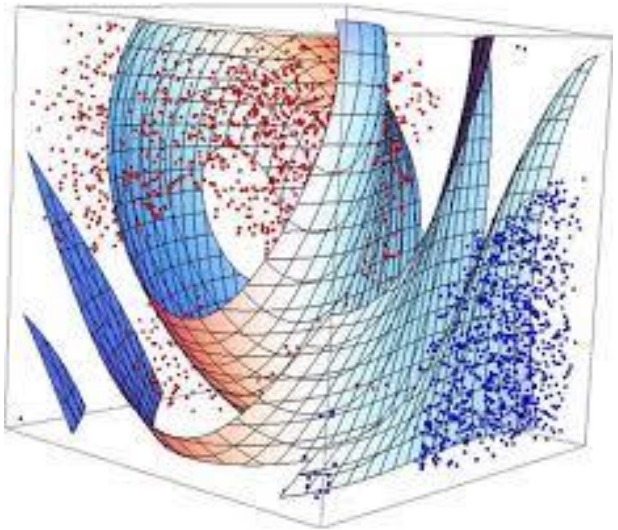
- Testing Dataset

Blogger	already	also	although	always	amazing	another	anyone	anything
1	0	0	0	0	0	1	0	0
2	0	0	0	0	0	0	0	0
3	0	0	1	0	1	0	0	0
4	0	1	0	0	0	0	0	1
5	0	0	0	2	0	0	0	0
6	0	1	0	0	0	0	0	0
7	0	1	0	0	0	0	0	0
8	0	0	0	0	0	0	0	2
9	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0

- The Gender column has been removed from the testing set because we will be predicting the gender based off of the blog entries.

Build a Support Vector Machine

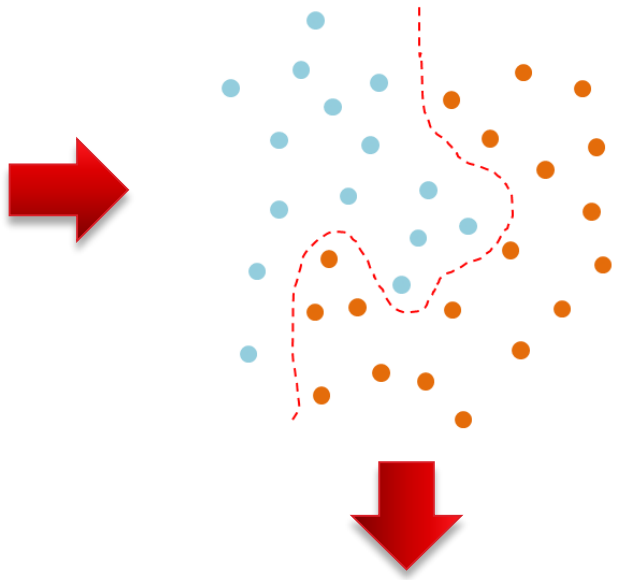
- First, lets run a tuning function to identify the best parameters to use for the SVM model.
- The process runs a 10-fold cross validation methodology and identified the following:
  - Gamma = 0.001
  - Cost = 10
- Best Performance = 0.3879642



Testing the Support Vector Machine

- We created a random test sample of the dataset and included only the measurement variables. This data was not involved in the initial training of the Support Vector Machine but will be used to validate the results from passing the data through the SVM.

Gender	already	also	although	always	amazing	another	anyone	anything
Male	0	0	0	1	0	0	0	0
Male	0	0	0	0	0	0	0	0
Male	1	2	0	0	0	0	0	0
Male	0	1	0	0	0	0	0	0
Female	0	0	0	2	0	0	0	0
Male	0	0	1	1	1	0	0	0
Female	0	0	0	0	0	2	0	0
Male	2	0	0	0	0	0	0	3
Female	0	0	0	1	0	0	0	0
Female	1	0	0	0	0	0	0	0
Male	1	2	0	0	0	1	0	0
Male	1	3	0	0	0	2	0	1
Female	2	1	0	0	1	0	0	1
Male	0	3	0	0	0	0	0	0
Male	0	5	0	1	0	0	1	1



Prediction Accuracy:

53.2%

Blogger	Gender	Prediction
1	Male	Male
2	Male	Female
3	Female	Female
4	Male	Male
5	Female	Female
6	Male	Male
7	Male	Female
8	Male	Female
9	Female	Male
10	Male	Female

Lessons Learned

- 53.2% is not a strong prediction based off of the data. It is only 3.2% better than randomly guessing the gender.
- It is important to understand that unstructured text analysis does not always yield strong predictive results.
- This is due in part of having:
  - Insufficient sample size
  - Lack of pattern in the blog entries.
  - Tuning stemming technique
  - Data dictionary definition
  - Incorrect Predictive Algorithm / parameters
- Nevertheless, we should be able to improve this result through a careful review of the modeling techniques and altering our approach when necessary. As our text analytics improve over time, so will our capabilities for developing robust profiles of individuals.

Text Analytics Example Natural Language Processing

Natural Language Processing

- Natural language processing (NLP) is a field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages.
- As such, NLP is related to the area of human–computer interaction.
- Many challenges in NLP involve natural language understanding, that is, enabling computers to derive meaning from human or natural language input, and others involve natural language generation.

Natural Language Processing in R

- There is an R package called openNLP which contains an Apache implementation of java-based NLP tools.

- This package can perform a variety of NLP functions including:
  - Sentence Splitting
  - Tokenization
  - Part of Speech Tagging (POS)
  - Named Entity Recognition
  - Chunking
  - Parsing

```
#########################################################
# Initialize the packages.
#########################################################

# Make sure that we are running 32 bit version of R!!!!

library("openNLP")
library("openNLPdata")
library("tm")
library("NLP")
library("openNLPmodels.en")

#################################################################
# Create some text to manipulate with NLP
#################################################################

s <- paste(c("Pierre Vinken, 61 years old, will join the board as
             "nonexecutive director on November 29. ",
             "Mr. Vinken is chairman of Chicago , ",
             "the Dutch publishing group."),
           collapse = "")

# s <- as.String(mydata$Message)

s <- as.String(s)

.
```

Natural Language Processing in R

Sentence splitting

- Sentence boundary = period + space(s) + capital letter

- **Example:** Unusually, the gender of crocodiles is determined by temperature. If the eggs are incubated tat over 33c, then the egg hatches into a male or 'bull' crocodile. At lower temperatures only female or 'cow' crocodiles develop.

- Unusually, the gender of crocodiles is determined by temperature.
- If the eggs are incubated tat over 33c, then the egg hatches into a male or 'bull' crocodile.
- At lower temperatures only female or 'cow' crocodiles develop.

```
##################################################
# Sentence Splitting
##################################################

# Break apart the text into separate sentances.

sent_token_annotator <- Maxent_Sent_Token_Annotator()
a1 <- annotate(s, sent_token_annotator)
```

Sentence	Text
1	Pierre Vinken, 61 years old, will join the board as a nonexecutive
2	Mr. Vinken is chairman of Chicago , the Dutch publishing group.

Natural Language Processing in R

Tokenization

- Convert a sentence into a sequence of tokens
- Divides the text into smallest units (usually words), removing punctuation.

- **Example:** A Saudi Arabian woman can get a divorce if her husband doesn't give her coffee.

- A Saudi Arabian woman can get a divorce if her husband does n't give her coffee .

```
##########################################################
# Tokenization
##########################################################

# Find the individual words in each sentence.

word_token_annotator <- Maxent_Word_Token_Annotator()
word_token_annotator
a2 <- annotate(s, word_token_annotator, a1)
a2
```

ID	Type	Start	End
3	word	1	6
4	word	8	13
5	word	14	14
6	word	16	17
7	word	19	23
8	word	25	27
9	word	28	28
10	word	30	33
11	word	35	38

Natural Language Processing in R

Part-of-speech tagging

- Assign a part-of-speech tag to each token in a sentence.
- Example: Most lipstick is partially made of fish scales

- Most/ JJS lipstick/ NN is/ VBZ partially/ RB made/ VBN of/ IN fish/ NN scales/ NNS

Tokens with POS Tag		
Pierre/NNP	Vinken/NNP	,/,
61/CD	years/NNS	old/JJ
,/,	will/MD	join/VB
the/DT	board/NN	as/IN
a/DT	nonexecutive/JJ	director/NN
on/IN	November/NNP	29/CD
./.	Mr./NNP	Vinken/NNP
is/VBZ	chairman/NN	of/IN
Chicago/NNP	,/,	the/DT
Dutch/JJ	publishing/NN	group/NN
./.		

```
##############################################
# Part of Speech Tagging
##############################################

pos_tag_annotator <- Maxent_POS_Tag_Annotator()
pos_tag_annotator
a3 <- annotate(s, pos_tag_annotator, a2)
a3

## Variant with POS tag probabilities as (additional) features.
head(annotate(s, Maxent_POS_Tag_Annotator(probs = TRUE), a2))

## Determine the distribution of POS tags for word tokens.
a3w <- subset(a3, type == "word")
tags <- sapply(a3w$features, `[[`, "POS")
tags
table(tags)

## Extract token/POS pairs (all of them): easy.
sprintf("%s/%s", s[a3w], tags)
```

Natural Language Processing in R

- Part of Speech Tags

Tag	Description	Tag	Description
CC	Coordinating conjunction	PRP$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential *there*	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition or subordinating conjunction	SYM	Symbol
JJ	Adjective	TO	to
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund or present participle
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3rd person singular present
NNP	Proper noun, singular	VBZ	Verb, 3rd person singular present
NNPS	Proper noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ending	WP$	Possessive wh-pronoun
PRP	Personal pronoun	WRB	Wh-adverb

Natural Language Processing in R

Named Entity Recognition

- Named entity recognition classify tokens in text into predefined categories such as date, location, person, time.
- The name finder can find up to seven different types of entities - date, location, money, organization, percentage, person, and time.
- Example: Diana Hayden was in Philadelphia on October 3rd.

- <namefind/person> Diana Hayden </namefind/person> was in<namefind/location> Philadelphia </namefind/location> on <namefind/date> October 3rd </namefind/date>

```
##############################################################
# Named Entity Recognition
##############################################################

# requires package openNLPmodels.en
# from http://datacube.wu.ac.at

## Entity recognition for persons.
entity_annotator <- Maxent_Entity_Annotator(kind="person")
entity_annotator
annotate(s, entity_annotator, a2)
```

ID	Type	Start	End	Features
34	entity	1	13	kind = person
34	entity	119	125	kind = location
34	entity	80	90	kind = date

Hierarchy	Text
Name	Pierre Vinken
Location	Chicago
Date	November 29

Natural Language Processing in R

Chunking (shallow parsing)

- the identification of parts of speech and short phrases (like noun phrases). Part of speech tagging tells you whether words are nouns, verbs, adjectives, etc., but it doesn't give you any clue about the structure of the sentence or phrases in the sentence.

- In NER, your goal is to find named entities, which tend to be noun phrases (though aren't always), so you would want to know that President Barack Obama is in the following sentence:

- President Barack Obama criticized insurance companies and banks as he urged supporters to pressure Congress to back his moves to revamp the health-care system and overhaul financial regulations.

- But you wouldn't necessarily care that he is the subject of the sentence.

```
##############################################################
# Chunker - Shallow Parsing
##############################################################
## Chunking needs word token annotations with POS tags.

sent_token_annotator <- Maxent_Sent_Token_Annotator()
word_token_annotator <- Maxent_Word_Token_Annotator()
pos_tag_annotator <- Maxent_POS_Tag_Annotator()
a3 <- annotate(s,
               list(sent_token_annotator,
                    word_token_annotator,
                    pos_tag_annotator))

annotate(s, Maxent_Chunk_Annotator(), a3)
annotate(s, Maxent_Chunk_Annotator(probs = TRUE), a3)
```

ID	Type	Start	End	Features
3	word	1	6	POS=NNP, chunk_tag=B-NP, chunk_prob=0.9740431
4	word	8	13	POS=NNP, chunk_tag=I-NP,chunk_prob=0.9816025
5	word	14	14	POS=,, chunk_tag=O, chunk_prob=0.9863059
6	word	16	17	POS=CD, chunk_tag=B-NP,chunk_prob=0.9926662
7	word	19	23	POS=NNS, chunk_tag=I-NP,chunk_prob=0.9854421
8	word	25	27	POS=JJ, chunk_tag=B-ADJP,chunk_prob=0.9978292
9	word	28	28	POS=,, chunk_tag=O, chunk_prob=0.9909762
10	word	30	33	POS=MD, chunk_tag=B-VP, chunk_prob=0.979816
11	word	35	38	POS=VB, chunk_tag=I-VP,chunk_prob=0.9857121
12	word	40	42	POS=DT, chunk_tag=B-NP,chunk_prob=0.9932718

Example: He reckons the current account deficit will narrow to only 1.8 billion in September

NP VP NP VP PP NP PP NP

Natural Language Processing in R

Tree Bank Parsers

- A program that works out the grammatical structure of sentences, for instance, which groups of words go together (as "phrases") and which words are the subject or object of a verb.
- Probabilistic parsers use knowledge of language gained from hand-parsed sentences to try to produce the most likely analysis of new sentences.
- Their development was one of the biggest breakthroughs in natural language processing in the 1990s.

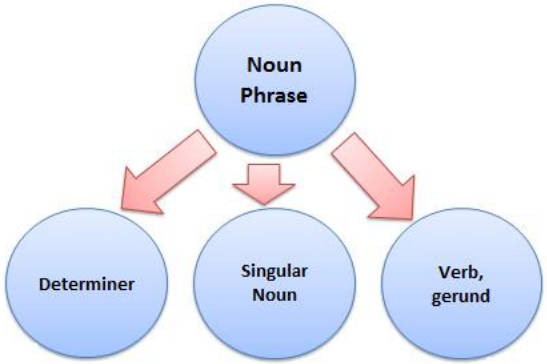
```
#####################################################################
# Parse Annotator
#####################################################################

## Need sentence and word token annotations.
sent_token_annotator <- Maxent_Sent_Token_Annotator()
word_token_annotator <- Maxent_Word_Token_Annotator()
a2 <- annotate(s, list(sent_token_annotator, word_token_annotator))

parse_annotator <- Parse_Annotator()
## Compute the parse annotations only.
p <- parse_annotator(s, a2)
## Extract the formatted parse trees.
ptexts <- sapply(p$features, `[[`, "parse")
ptexts
```

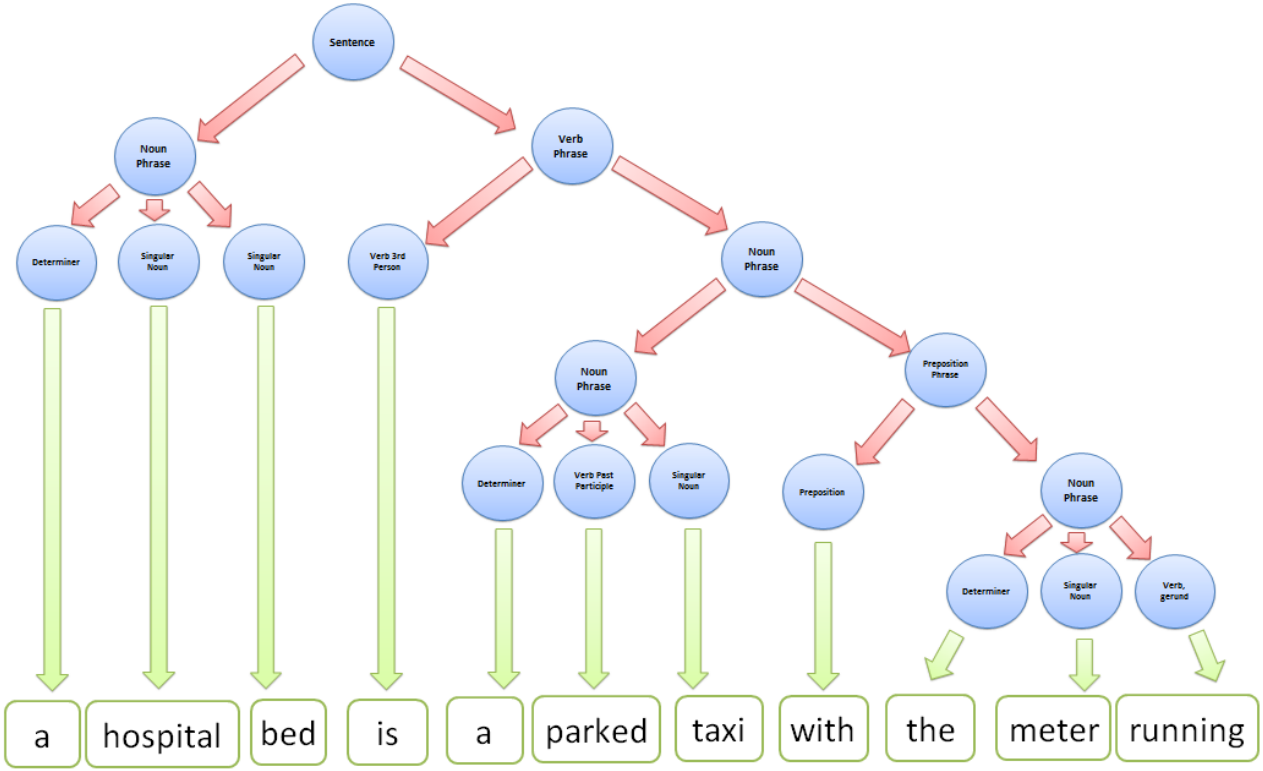
```
[1] "(TOP (S (NP (NP (NNP Pierre) (NNP Vinken))(, ,) (ADJP (NP (CD 61) (
NNS years)) (JJ old)))(, ,) (VP (MD will) (VP (VB join) (NP (DT the) (NN
board)) (PP (IN as) (NP (NP (DT a) (JJ nonexecutive) (NN director)) (PP
(IN on) (NP (NNP November) (CD 29)))))))(. .)))"
[2] "(TOP (S (NP (NNP Mr.) (NNP Vinken)) (VP (VBZ is) (NP (NP (NP (NN ch
airman)) (PP (IN of) (NP (NNP Chicago)))) (, ,) (NP (DT the) (JJ Dutch)
(NN publishing) (NN group))))(. .)))"
```

- Example: A hospital bed is a parked taxi with the meter running.

- (TOP (S (NP (DT A) (NN hospital) (NN bed)) (VP (VBZ is) (NP (NP (DT a) (VBN parked) (NN taxi)) (PP (IN with) (NP (DT the) (NN meter) (VBG running)))))))

Natural Language Processing in R






```
## Read into NLP Tree objects.
ptrees <- lapply(ptexts, Tree_parse)
ptrees
```

```
(TOP
  (S
    (NP (NNP Mr.) (NNP Vinken))
    (VP
      (VBZ is)
      (NP
        (NP (NP (NN chairman)) (PP (IN of) (NP (NNP Chicago))))
        (, ,)
        (NP (DT the) (JJ Dutch) (NN publishing) (NN group))))
    (. .)))
```

Natural Language Processing

- It's ironic that natural language, the symbol system that is easiest for humans to learn and use, is hardest for a computer to master.
- Long after machines have proven capable of inverting large matrices with speed and grace, they still fail to master the basics of our spoken and written languages.
- Eventually you will be able to address your computer as though you were addressing another person.
- This goal is not easy to reach. "Understanding" language means, among other things, knowing what concepts a word or phrase stands for and knowing how to link those concepts together in a meaningful way.
- NLP remains one the uncharted frontiers in computer science and constantly evolving.

Text Analytics Example Clustering Uncategorized RSS Feeds

News Sites on the WWW

- The content of major news outlets can be a rich source of information for text analytical work.
- Data obtained from the news can be leveraged to help predict stock movements as well as a drive new regression coefficients that can be used in data mining / predictive analytics.
- RSS feeds are an XML format designed to make the extraction of text information from the news easy for BI systems.
- This presentation will process various RSS news feeds obtained by BBC, CNN, NBC News, and NY Times and we will cluster them for categorization using an unsupervised learning technique.

Acquiring the Data

- The data has been scrapped into MS SQL through an SSIS package designed to find major RSS news feeds.


Science on NBCNEWS.com

Up and atom! Simplest clock yet tells time with single atom

Scientists say it could lead to radically new way to define mass — or even more exotic clock

Jump to discuss
Loading comments...

Below: Discuss Related

Recommend 0 Tweet 0 +1 2 Share 5

By Charles Choi
LiveScience
updated 1/11/2013 11:41:13 AM ET

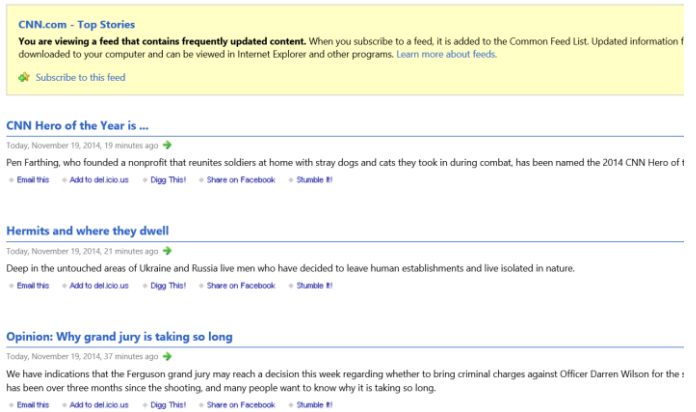
Print | Font: A A + –

A clock based on just a single atom — the simplest clock yet — has now been devised, researchers say.

This new device to measure time could help lead to a radically new way to define mass as well, scientists added.

- News Page

- RSS Feed


CNN.com - Top Stories

You are viewing a feed that contains frequently updated content. When you subscribe to a feed, it is added to the Common Feed List. Updated information f downloaded to your computer and can be viewed in Internet Explorer and other programs. Learn more about feeds.

Subscribe to this feed

CNN Hero of the Year is ...

Today, November 19, 2014, 19 minutes ago

Pen Farthing, who founded a nonprofit that reunites soldiers at home with stray dogs and cats they took in during combat, has been named the 2014 CNN Hero of t

Email this · Add to del.icio.us · Digg This! · Share on Facebook · Stumble It!

Hermits and where they dwell

Today, November 19, 2014, 21 minutes ago

Deep in the untouched areas of Ukraine and Russia live men who have decided to leave human establishments and live isolated in nature.

Email this · Add to del.icio.us · Digg This! · Share on Facebook · Stumble It!

Opinion: Why grand jury is taking so long

Today, November 19, 2014, 37 minutes ago

We have indications that the Ferguson grand jury may reach a decision this week regarding whether to bring criminal charges against Officer Darren Wilson for the s has been over three months since the shooting, and many people want to know why it is taking so long.

Email this · Add to del.icio.us · Digg This! · Share on Facebook · Stumble It!

Title	Description	News Feed	Extraction Date
Surprise! Catcher is soldier dad	Soldier Ian Jones returns from a deployment in Afghanistan	CNN	4/27/2014
Why this 23-year-old has 24 kids	It's a sunny April afternoon at the University of Rwanda Coll	CNN	4/27/2014
Check for fifties under the mattress	There are still millions of Houblon £50 notes in circulation, d	BBC	4/28/2014
It looks like a puppy, but it's not ...	A family was surprised to learn the abandoned 'puppy' they	CNN	4/28/2014
New Dove beauty ad goes too far?	Dove's latest viral ad campaign has hurt a lot of feelings. Ke	CNN	4/28/2014
Outrage: 'What's the captain doing?'	A teen's father gave a South Korean TV network footage fro	CNN	4/28/2014
A swift punishment, but is it a just on	The debate over the banning of LA Clippers' Donald Sterling	BBC	4/29/2014
Amanda Knox Rejects Court's Reason	Amanda Knox rejected an Italian court's contention Tuesday	NBCNews	4/29/2014
Cast Is Announced for Next 'Star War	The seventh installment of the space epic will feature seven	NY Times	4/29/2014
Heroin and Alcohol Led to the Deaths	Autopsies of Jeffrey Reynolds and Mark Kennedy, the forme	NY Times	4/29/2014
Moscow Journal: Amid a Revived Eas	At the Museum of the Cold War, visitors are drawn as much	NY Times	4/29/2014

Data Preparation for Prediction

- We will split the dataset into a training and testing sample. We will sample 70% for our training data and 30% for the validation test.
  - Training Dataset
  - Testing Dataset

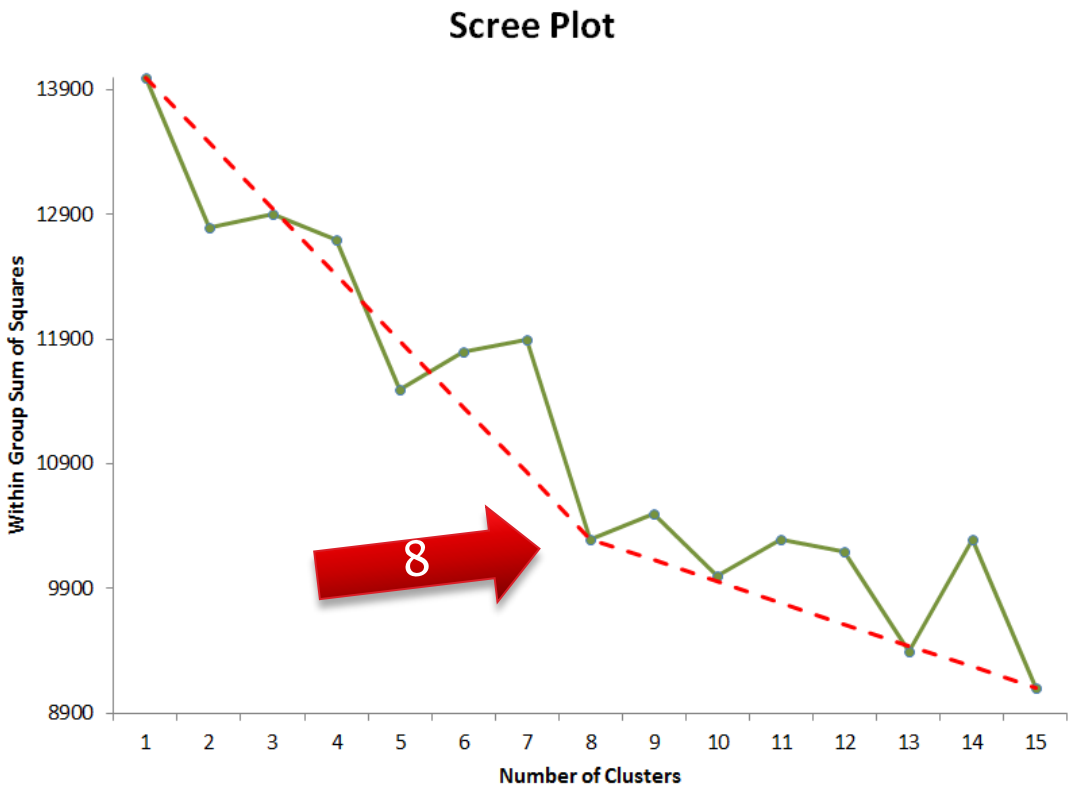
city	court	crimea	murder	officials	police	president	south	state
1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	1	1
0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
1	0	2	0	1	0	0	0	0
0	0	0	1	0	0	0	0	0
0	1	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0

city	court	crimea	murder	officials	police	president	south	state
1	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0
0	0	0	0	0	1	0	0	0
0	0	1	0	0	0	0	0	0

- We will implement a kNN algorithm to perform the clustering of the training documents.

Creating the Unsupervised Cluster

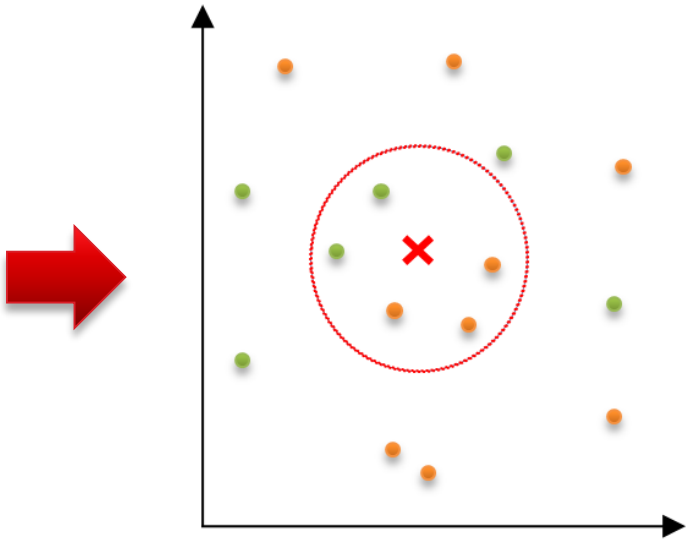
- In order to tune the classification parameter for the kNN, we will produce and review a scree plot.
- A bend at the elbow indicates the optimal parameters for the clustering procedure.
- The scree plot indicates that k = 8 might be an appropriate starting parameter.



Creating the Unsupervised Cluster

- The testing dataset was passed through a kNN classification algorithm.

city	court	crimea	murder	officials	police	president	south	state
1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	1	1
0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
1	0	2	0	1	0	0	0	0
0	0	0	1	0	0	0	0	0
0	1	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0



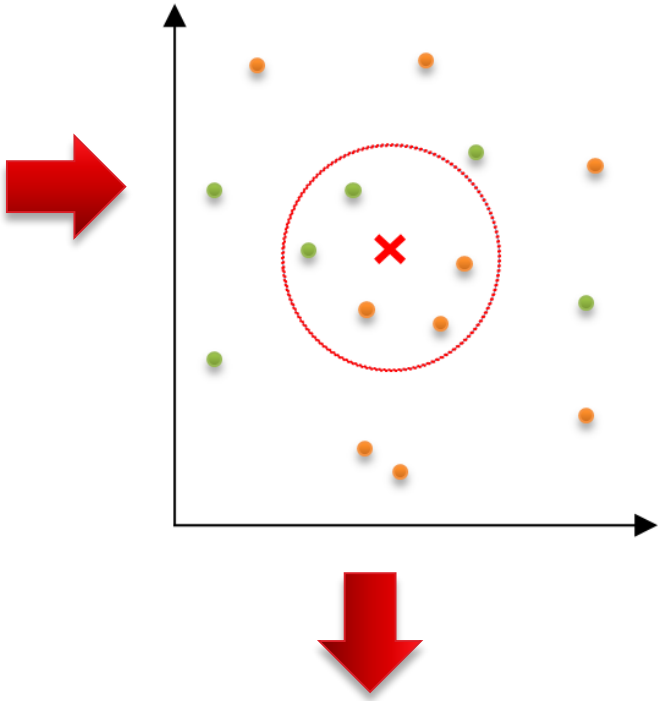
Cluster Review

- Once the 8 clusters have been produced we can review the underlying terms to detect any overarching theme(s).
- Understanding these terms will allow for us to create meaningful labels to the clusters.
- Drawing from other text data preparation techniques, such as NLP, it is possible to created unsupervised computer generated cluster labels.

Cluster	Description	Document Count	% of Total
1	Opinions	1532	12.3%
2	Global Conflict	1252	10.1%
3	General	4989	40.2%
4	US Regional	642	5.2%
5	Crime	451	3.6%
6	Coming Events	1249	10.1%
7	Police	930	7.5%
8	Political	1361	11.0%

Applying the Cluster on New RSS Feeds

city	court	crimea	murder	officials	police	president	south	state
1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	1	1
0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
1	0	2	0	1	0	0	0	0
0	0	0	1	0	0	0	0	0
0	1	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0



Cluster	city	court	crimea	murder	officials	police	president	south	state
General	1	0	0	0	0	0	0	0	0
General	0	0	0	0	0	0	0	0	1
US Regional	0	0	0	0	0	0	0	1	1
Political	0	0	0	0	0	0	1	0	0
Coming Events	0	0	0	0	0	0	0	0	0
General	0	0	0	0	0	0	0	0	0
Global Conflict	1	0	2	0	1	0	0	0	0
Crime	0	0	0	1	0	0	0	0	0
Police	0	1	0	0	0	1	0	0	0
Opinions	0	0	0	0	0	0	0	0	0

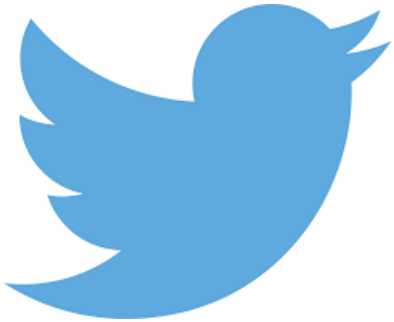
Text Analytics Example Social Media Sentiment & Network Analysis

Social Media Analysis

- The purpose of this tutorial is to showcase how we can build off of the text analytics techniques we have discussed so far and tap into a wealth of data made available from social media platforms.

- We will focus on pulling in data from Facebook and Twitter through the R interface.

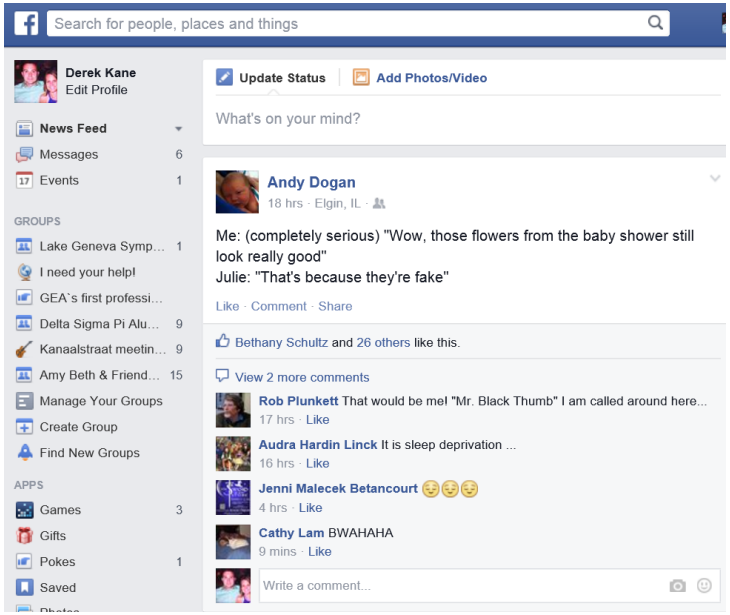
- Additionally, we will perform a couple of additional analysis including:
  - Social Network Diagrams
  - Natural Language Processing
  - Sentiment Analysis
  - Word Cloud Visualization



- Discussion of business applications for these techniques.

Acquiring the Data - Facebook

- R makes it possible to easily pull information related to Facebook and Twitter through their API. This makes R extremely appealing as the primary statistical tool to integrate unstructured and structured data analysis.



```
#############################################################
# Facebook Data Extraction
#############################################################

library(Rfacebook)
library(Rook)

fb_oauth <- fbOAuth(app_id="XXXXXXXXXXXXXXXX",
                    app_secret="YYYYYYYYYYYYYYY")
load("fb_oauth")

# Pull Friend Demogrpahics, Likes, Check-In,& News Feed

my_friends <- getFriends(token=fb_oauth, simplify=TRUE)
Likes <-getLikes(user="1075852970", n=500, token=fb_oauth)
Checkin <-getCheckins(user="1075852970", n=10,
                      token=fb_oauth, tags = FALSE)
NewsFeed <- getNewsfeed(token=fb_oauth, n=500)
```

Name	Wall Post	Gender	Birthday	Relationship Status
Kimberly Barnett	Happy Halloween! Great time with family and Friends	Female	NA	NA
Russ Kelsey	Gonna be a long night preparing for russells 1st bday party...hope to see everyon	Male	5/25/14	Married
James Martin	#stupidrandomfacti can do 22 handstand pushups before i collapse upon myself	Male	12/31/82	Engaged
Eric Alford	Hawks score and my buzzer remote takes a crap.	Male	9/17/80	Married
Stephen Lejeune	Looking forward to seeing my art buddys and some sweet ass bands at my favori	Male	NA	Married
Rob Kolb	Thanks to all for the birthday wishes! The love of my life, Marcey Bender Moorm	Male	10/14/65	In a relationship
JoAnne Serowka	OMG! They are reporting SNOW in McHenry & Woodstock- Ugh- here come the	Female	3/3/55	Married
Michael Schindler	The most impressive building I have ever seen	Male	7/14/14	NA
Tommy Brodie	Happy Halloween, especially to CPD who wouldn't let me surf.	Male	9/8/14	NA
Kelly Wulf Kellerman	Happy Birthday!!! Hope you have a great day today!! Good luck on your run!!	Female	9/16/81	Married
Eric Morgenstern	Gotta love it, my 3 year old son gets in the car in the morning and wants to listen	Male	10/8/81	Married
Susan L. Tarson	Passed recertification so I am once again a National Board Certified Teacher! Th	Female	8/5/14	In a relationship
Marcy Tunison	I'm blaming this blustery snowy Halloween on all the little Elsa's out there today	Female	5/28/85	Married
Greg Franczyk	So Dealer day at the show was interesting. I ran into a guy who collects Disney W	Male	NA	Single
April Stoltman	It kind of saddened me to not see kids out today.....we only passed 2 groups of k	Female	4/18/14	Married
April Stoltman	It's my favorite day of the year! Have a safe fabulous candy filled holiday!	Female	4/18/14	Married
Jennifer Murphy	Update on the car - the insurance adjusted told us today that they are going to g	Female	5/3/72	Married

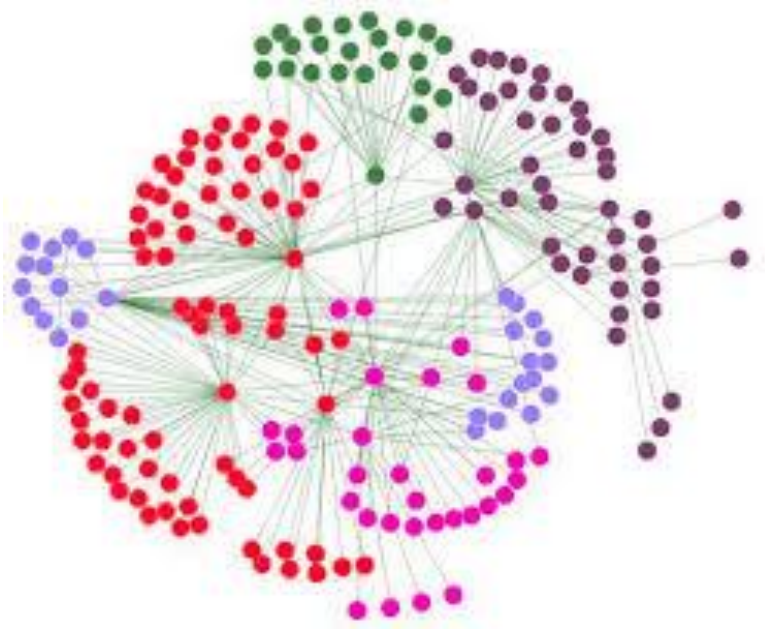
Acquiring the Data - Twitter


```
############################################################
# Twitter Data Extraction
############################################################

library("twitteR")

requestURL <- "https://api.twitter.com/oauth/request_token"
accessURL <- "https://api.twitter.com/oauth/access_token"
authURL <- "https://api.twitter.com/oauth/authorize"
consumerKey <- "XXXXXXXXXXXXXXXXX"
consumerSecret <- "YYYYYYYYYYYYYYYYYYYYYYYYY"
Cred <- OAuthFactory$new(consumerKey=consumerKey,
                         consumerSecret=consumerSecret,
                         requestURL=requestURL,
                         accessURL=accessURL,
                         authURL=authURL)

# Pull Tweets with #Adrian Peterson

Tweets <- searchTwitter('#AdrianPeterson',
                        n=100, cainfo="cacert.pem")
```

Date	Name	Tweet	retweetCount
11/18/14	DariusHoward_PR	#AdrianPeterson is suspended for the remainder of the season w/o pay. I	0
11/18/14	Young_Hypocrite	RT @TheRoot: #AdrianPeterson is suspended from the @NFL without pay	11
11/18/14	stroker66ace	@FCC please look into the actions of @1057FMTheFan. I believe they are	0
11/18/14	Adorable_Mikey	Watched my buddy hit his wife and get fired. Hit my kid and got the same	0
11/18/14	LilBill2345	Whether you agree with physical child discipline or not you have to admit	0
11/18/14	DJRobertHorry	homeboy is a dum-dum, but the NFLPA is going to sue the fuck out of the	0
11/18/14	bwolfe23	RT @AceKlubKasanova: My thoughts #NFL #AdrianPeterson http://t.co/S	1
11/18/14	legalspeaks	RT @BringMN: All the developments in the #AdrianPeterson suspension,	2
11/18/14	ItsShanaRenee	NEW! New Rules: #RogerGoodell doesn't fight fair in #AdrianPeterson's s	0
11/18/14	Krismaer53	#AdrianPeterson Not sure having an angry unemployed without pay man	0
11/18/14	Dj_Ango_	Ha "@fishsports: Do we grasp the irony in #AdrianPeterson thinking he's	0
11/18/14	ReddFoxxThePoet	Hm rt"@thetoyman1: #AdrianPeterson has been suspended without pay	0
11/18/14	The_Rob_Wagner	The only people dumber than #AdrianPeterson are the morons defending	0

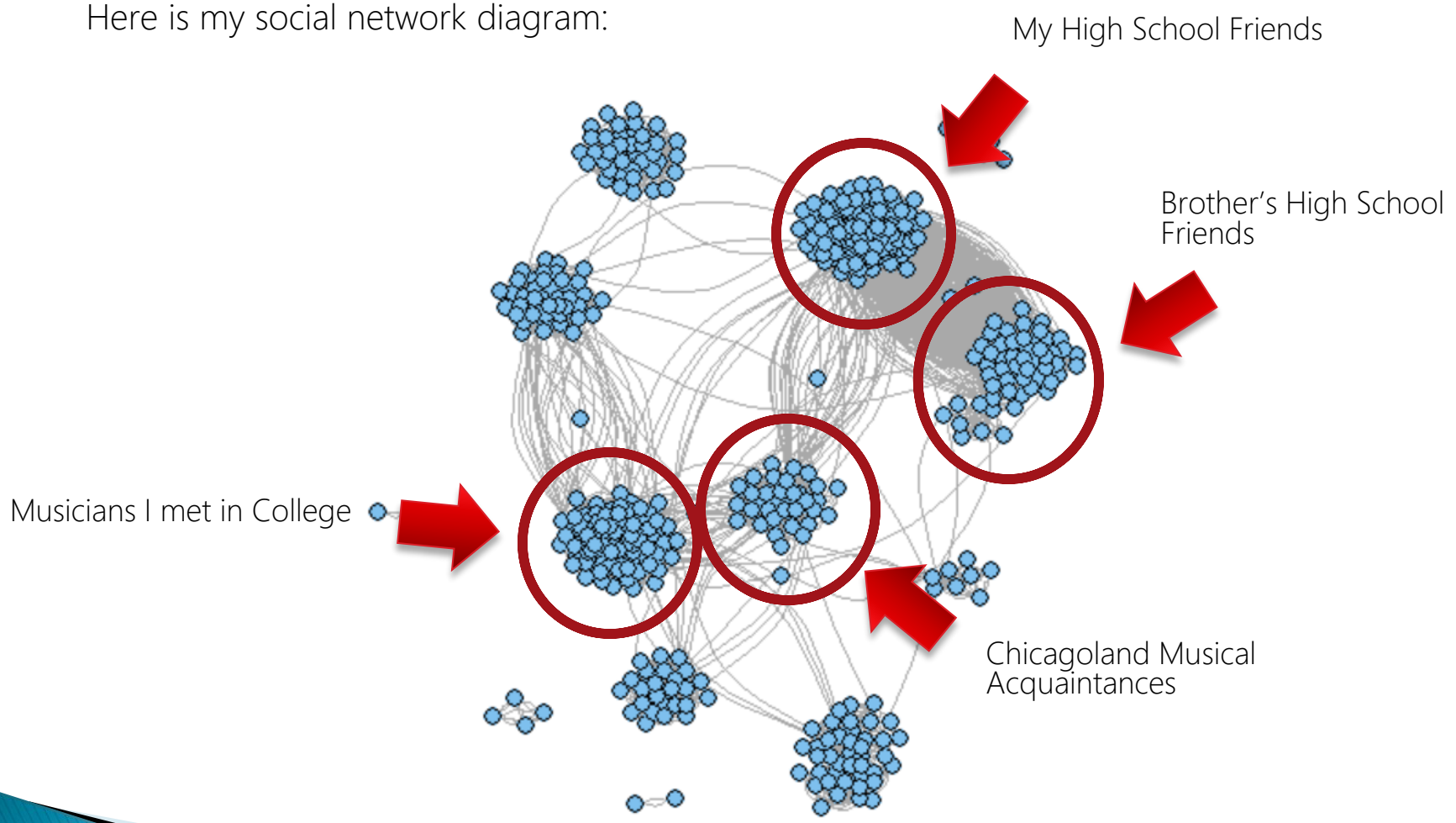
Social Network Analysis - Facebook

- Now that we have some data to analyze. Lets start off by creating a Social Network Analysis and diagram.
- Social network analysis (SNA) is the use of network theory to analyze social networks.
- Social network analysis views social relationships in terms of network theory, consisting of nodes, representing individual actors within the network, and ties which represent relationships between the individuals, such as friendship, kinship, organizations and sexual relationships.



Facebook Network Analysis

Here is my social network diagram:



What is Sentiment Analysis?

- Sentiment analysis is software for automatically extracting opinions, emotions, and sentiments in text.
- It allows for us to track attitudes and feelings on the web. People write blog posts, comments, reviews, and tweets about all sorts of different topics.
- We can track products, brands, and people for example and determine if they are viewed positively or negatively on the web.

Sentiment Analysis

- It allows for businesses to track:
  - Flame Detection (bad rants)
  - New Product Perception
  - Brand Perception
  - Reputation Management
- It allows individuals to get
  - An opinion on something (reviews) on a global scale.

Why Use Sentiment Analysis?

- According to a presentation by NetBase CMO Lisa Joy Rosner, the average consumer mentions specific brands over 90 times per week in conversations with friends, family, and co-workers.
- In addition, 53% of people on Twitter recommend companies and/or products in their tweets, with 48% of them delivering on their intention to buy the product.
- This means that Twitter and other social media are a perfect complement to traditional market research – especially has usage has spread through more demographics (social networking use among internet users aged 50+ has nearly doubled to 42% last year).
- You get unbiased, more truthful thoughts and opinions, and the target consumers come to you, naturally, and for free.

Sentiment Analysis – Multiple Areas

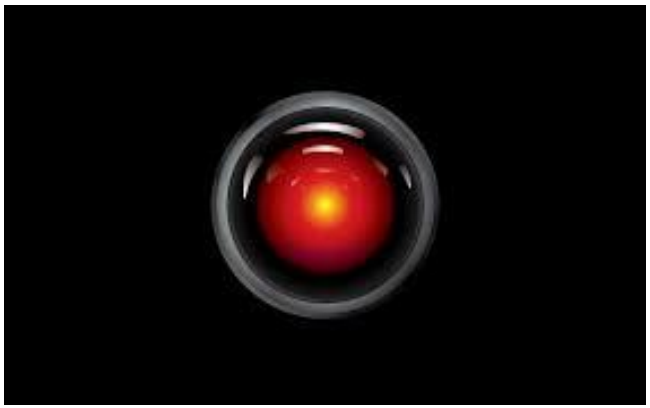
Natural Language Processing

- NLP deals with the actual text element. It transforms it into a format that the machine can use.

Artificial Intelligence

- It uses information given by the NLP and uses a lot of maths to determine whether something is negative or positive; it is used for clustering.

Sentiment Analysis



The problem has several dimensions:

- How does a machine define subjectivity & sentiment?
- How does a machine analyze polarity (negative/positive)?
- How does a machine deal with subjective word senses?
- How does a machine assign an opinion rating?
- How does a machine know about sentiment intensity?

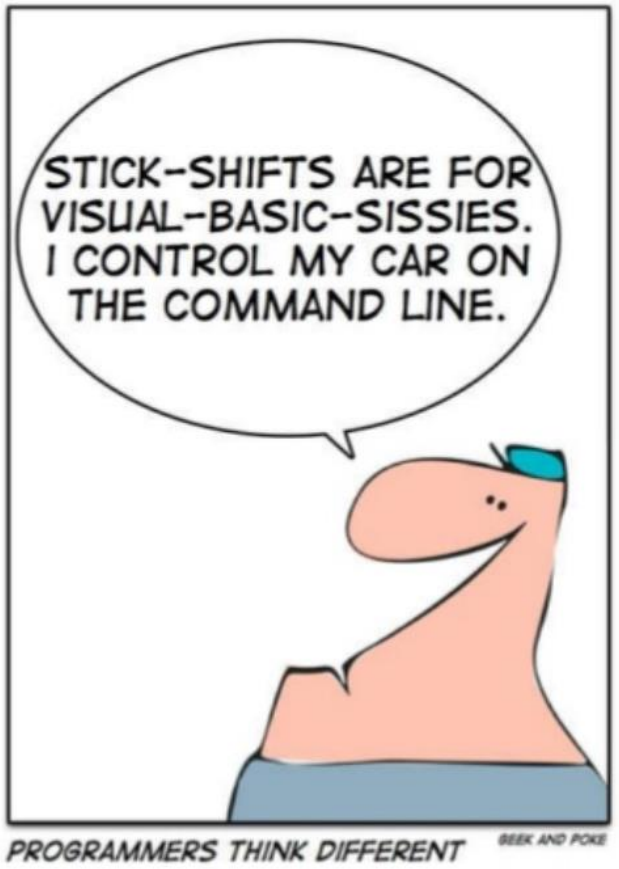
Sentiment Analysis

- It is not always easy to differentiate between fact and opinion.
- An opinion to a machine is called the "quintuple" (Bing Liu).

$$(o_j, f_{jk}, so_{ijkl}, h_i, t_i)$$

- o_j = The thing in question (ex. Product)
- f_{jk} = a feature of o_j
- so_{ijkl} = the sentiment value of the opinion of the opinion holder hi on feature f_{jk} of object o_j at time t_i

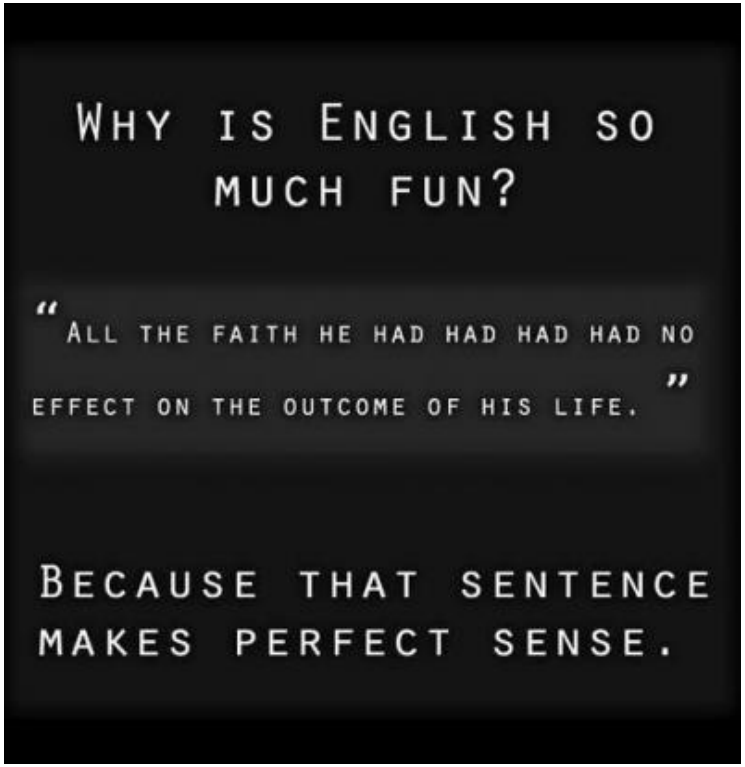
- These 5 elements have to be identified by the machine.
- All of these problems are unresolved by computer science and are open areas ripe for advancement.



Sentiment Analysis

Language is ambiguous. Consider the following:

- "The watch isn't water resistant" – In a product review this could be negative.
- "As much use as a trap door on a lifeboat" – negative but not obvious to the machine.
- "The canon camera is better than the Fisher Price on" – comparisons are hard to classify.
- "imo the ice cream is luuuurrrrrrrrvely" – slang and the way we communicate in general needs to be processed.

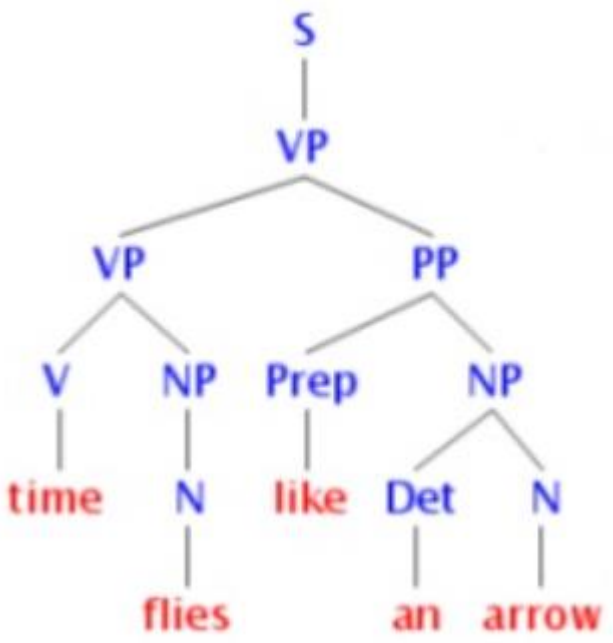


Natural Language Processing

Part of Speech Tagging

- The word in the text (or the sentence) at tagged used a POS tagger so that it assigns a label to each word, allowing the machine to do something like this:

S = subject
VP = Verb Phrase
V = Verb
N = Noun
NP = Noun Phrase
PP = Preposition
Det = Determiner



- Then we extract defined patterns like [Det] + [N] for example

Natural Language Processing cont'd

- We also look at the sentiment orientation (SO) of the patterns we extracted. For example, we may have extracted:

- Amazing + phone

which is:

- [JJ] + [NN] (or adjective followed by a noun in human)

- The opposite might be "Terrible" for example. In this stage, the computer tries to situate the words on an emotive scale.

Sentiment Analysis Scoring

The average sentiment orientation of all the phrases we gathered is computed.

- For a particular entry (Facebook Post / Tweet) the entry will be positive if the total count of positive terms is greater than the count of negative terms and vice versa.
- The sentiment scores can then be aggregated to calculate the score of the entire corpus:

Corpus Score = Positive Instances / Total Instances

- This allows the machine to say something like:

- "Generally people like the new iphone" ---> They recommend it.
- "Generally people hate the new iphone" ---> They do not recommend it.

Does Sentiment Analysis Work?

- The wider you throw your net and the more complex the language, the less accurate the system will be. This is simply due to the complexity the machine has to deal with.
- If you want to classify sentiments into + / - groups, then you are more likely to get a good result than if you are trying to classify into more exact groups (Excellent, incredible, good, etc…).
- More granularity requires more accuracy and this in turn requires a deeper understanding of the human language.

Facebook Sentiment Analysis

How is the sentiment determined for these Facebook posts?

Name	Wall Post	Gender	Birthday	Relationship Status
Kimberly Barnett	Happy Halloween! Great time with family and Friends	Female	NA	NA
Russ Kelsey	Gonna be a long night preparing for russells 1st bday party...hope to see everyon	Male	5/25/14	Married
James Martin	#stupidrandomfacti can do 22 handstand pushups before i collapse upon myself l	Male	12/31/82	Engaged
Eric Alford	Hawks score and my buzzer remote takes a crap.	Male	9/17/80	Married
Stephen Lejeune	Looking forward to seeing my art buddys and some sweet ass bands at my favori	Male	NA	Married
Rob Kolb	Thanks to all for the birthday wishes! The love of my life, Marcey Bender Moorm	Male	10/14/65	In a relationship
JoAnne Serowka	OMG! They are reporting SNOW in McHenry & Woodstock- Ugh- here come the	Female	3/3/55	Married
Michael Schindler	The most impressive building I have ever seen	Male	7/14/14	NA
Tommy Brodie	Happy Halloween, especially to CPD who wouldn't let me surf.	Male	9/8/14	NA
Kelly Wulf Kellerman	Happy Birthday!!! Hope you have a great day today!! Good luck on your run!!	Female	9/16/81	Married
Eric Morgenstern	Gotta love it, my 3 year old son gets in the car in the morning and wants to listen	Male	10/8/81	Married
Susan L. Tarson	Passed recertification so I am once again a National Board Certified Teacher! Th	Female	8/5/14	In a relationship
Marcy Tunison	I'm blaming this blustery snowy Halloween on all the little Elsa's out there today	Female	5/28/85	Married
Greg Franczyk	So Dealer day at the show was interesting. I ran into a guy who collects Disney W	Male	NA	Single
April Stoltman	It kind of saddened me to not see kids out today.....we only passed 2 groups of k	Female	4/18/14	Married
April Stoltman	It's my favorite day of the year! Have a safe fabulous candy filled holiday!	Female	4/18/14	Married
Jennifer Murphy	Update on the car - the insurance adjusted told us today that they are going to g	Female	5/3/72	Married

Example

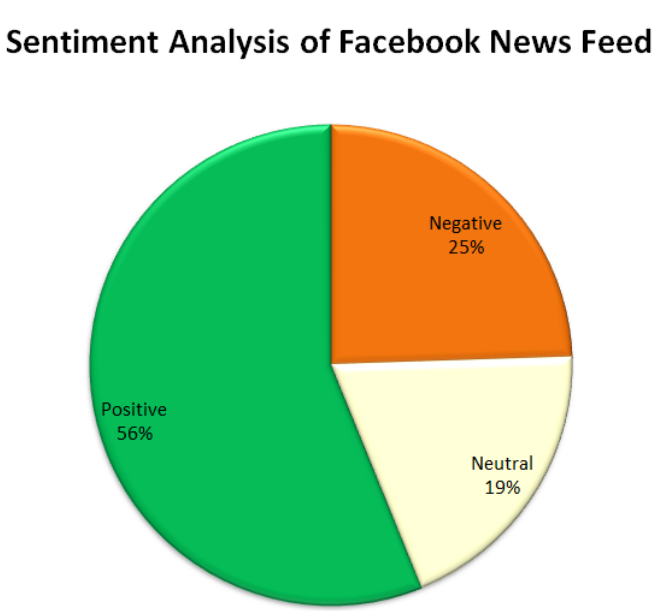
- Wall Post: "Happy Halloween! Great time with family and friends"
  - Natural Language Processing Rules: Happy = [Adj] and Great time = [Adj] + [V]
  - Positive Sentiment: 2, Negative Sentiment = 0
  - Sentiment Score = Positive

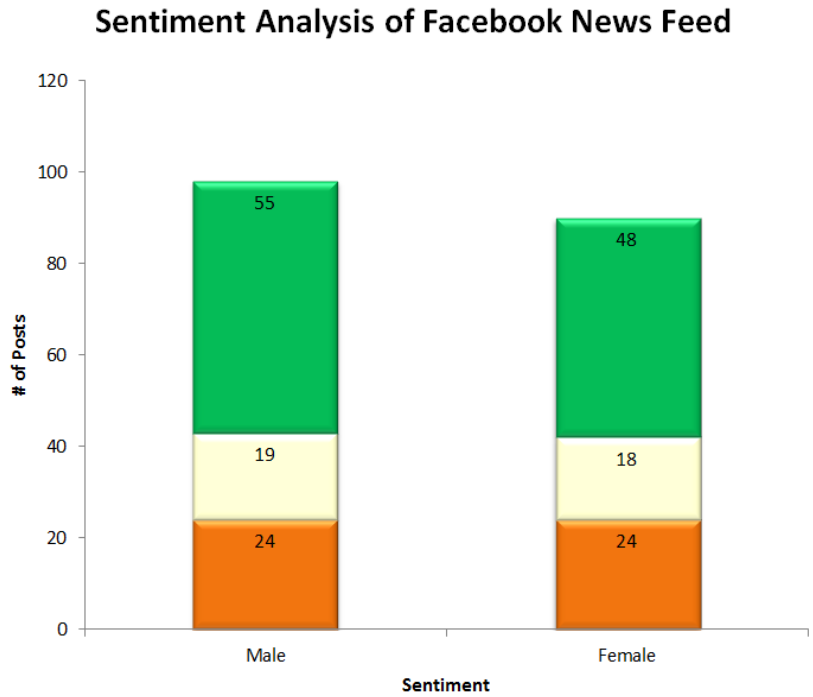
Facebook Sentiment Analysis

Wall Post	Sentiment
Happy Halloween! Great time with family and Friends	Positive
Gonna be a long night preparing for russells 1st bday party...hope to see everyon	Negative
#stupidrandomfacti can do 22 handstand pushups before i collapse upon myself I	Positive
Hawks score and my buzzer remote takes a crap.	Positive
Looking forward to seeing my art buddys and some sweet ass bands at my favori	Positive
Thanks to all for the birthday wishes! The love of my life, Marcey Bender Moorm	Positive
OMG! They are reporting SNOW in McHenry & Woodstock- Ugh- here come the	Negative
The most impressive building I have ever seen	Positive
Happy Halloween, especially to CPD who wouldn't let me surf.	Positive
Happy Birthday!!! Hope you have a great day today!! Good luck on your run!!	Positive
Gotta love it, my 3 year old son gets in the car in the morning and wants to listen	Positive
Passed recertification so I am once again a National Board Certified Teacher! Th	Neutral
I'm blaming this blustery snowy Halloween on all the little Elsa's out there today	Negative
So Dealer day at the show was interesting. I ran into a guy who collects Disney W	Neutral
It kind of saddened me to not see kids out today.....we only passed 2 groups of ki	Negative
It's my favorite day of the year! Have a safe fabulous candy filled holiday!	Positive
Update on the car - the insurance adjusted told us today that they are going to go	Neutral

Facebook Sentiment Analysis

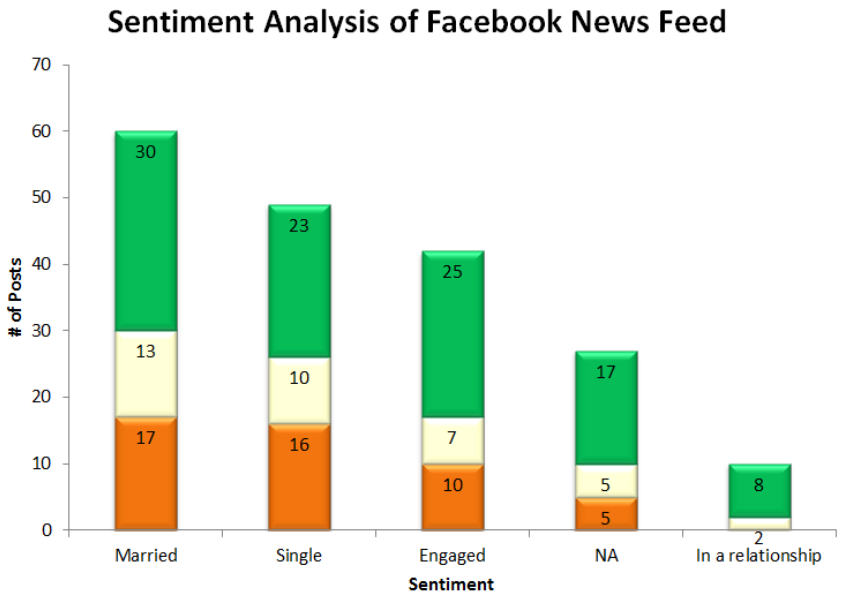
- Here is a breakdown of the sentiment scoring for my Facebook newsfeed that contains 188 posts that occurred on 10/31/2014.

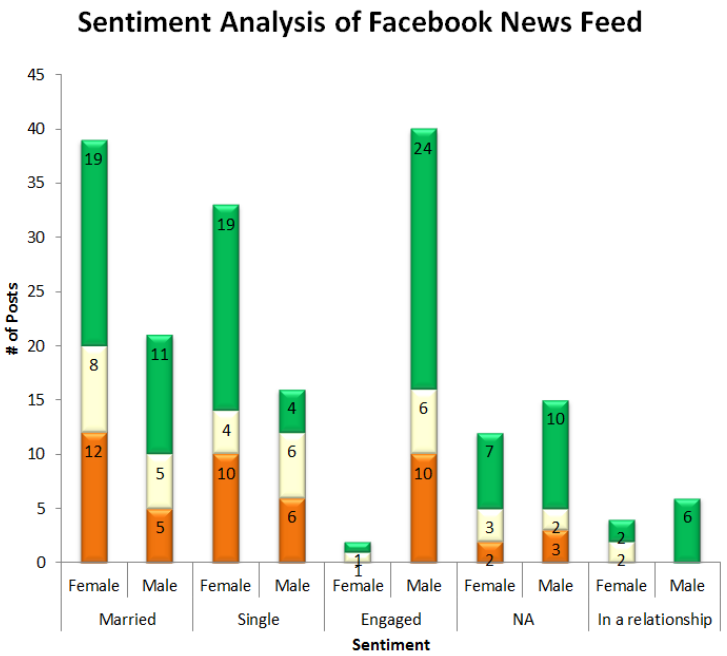




- We can see that there is 98 updates from males and 90 from females. In total, the algorithm had scored 56% posts as positive versus 25% negative.

Facebook Sentiment Analysis





- Married & Engaged friends have the highest degree of positive sentiment within their postings.
- Females who are married or single seem to be the most positive. There are almost no females who are engaged in this cohort.
- Interestingly, engaged males are slightly more positive than married males based off of their posts.

Facebook Sentiment Analysis

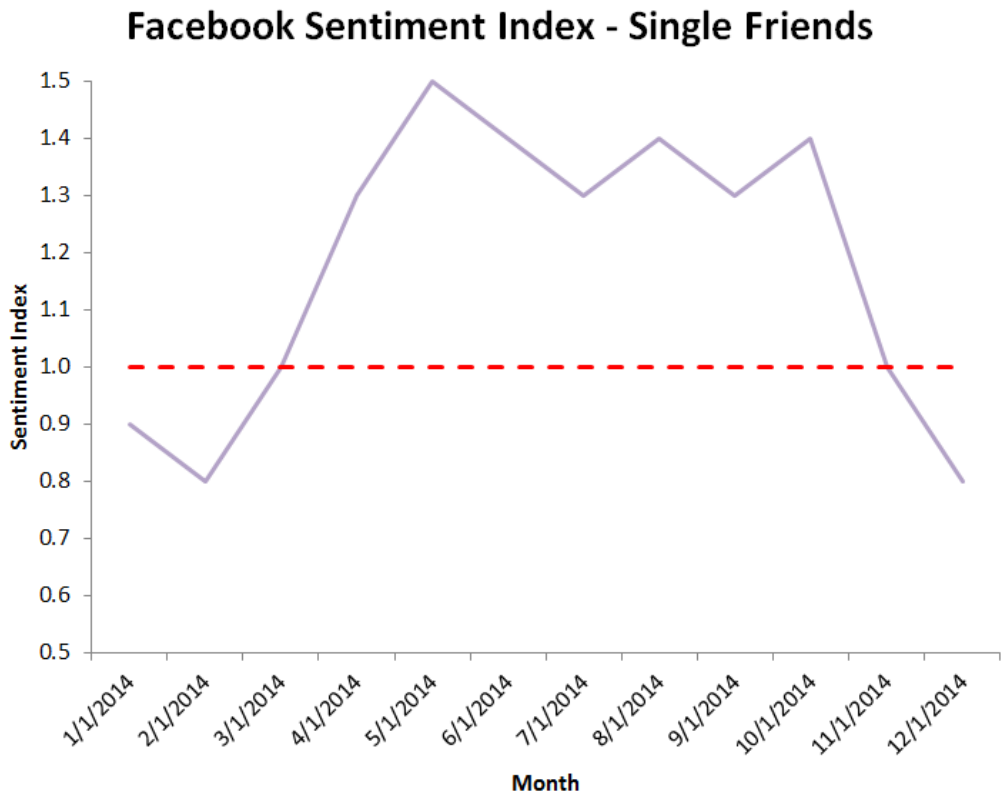
- How can we leverage these sentiment scores to incorporate the knowledge gained into predictive model building?

- Sentiment Index = Positive / Negative

Marital Status	Negative	Positive	Sentiment Index
Married	17	30	1.8
Single	16	23	1.4
Engaged	10	25	2.5
NA	5	17	3.4
In a relationship	1	8	8.0

Date	Marital Status	Sentiment Index
1/1/2014	Single	0.9
2/1/2014	Single	0.8
3/1/2014	Single	1.0
4/1/2014	Single	1.3
5/1/2014	Single	1.5
6/1/2014	Single	1.4
7/1/2014	Single	1.3
8/1/2014	Single	1.4
9/1/2014	Single	1.3
10/1/2014	Single	1.4
11/1/2014	Single	1.0
12/1/2014	Single	0.8
etc...	etc...	etc...

Facebook Sentiment Analysis

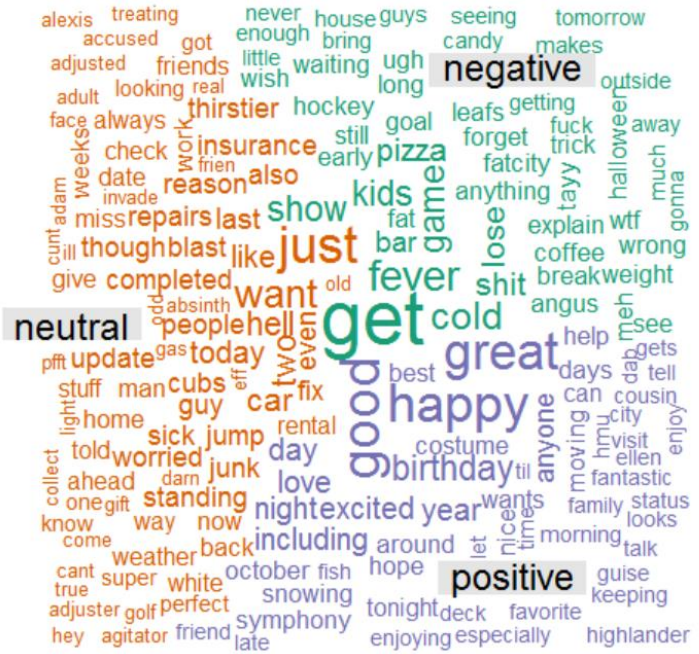


Business Application:
Product mentions instead of single friends

- This example graph shows that there is a negative sentiment during the colder periods of time and around the major family holidays (Thanksgiving, Christmas, St. Valentines Day)
- Once the weather warms up in Chicago, the overall sentiment of these singles improves.

Facebook – Word Cloud

- A word cloud is valuable tool used to present a visual image of the magnitude of the individual terms used throughout the overall postings.
- We will present two separate word clouds for review: total terms and sentiment based.



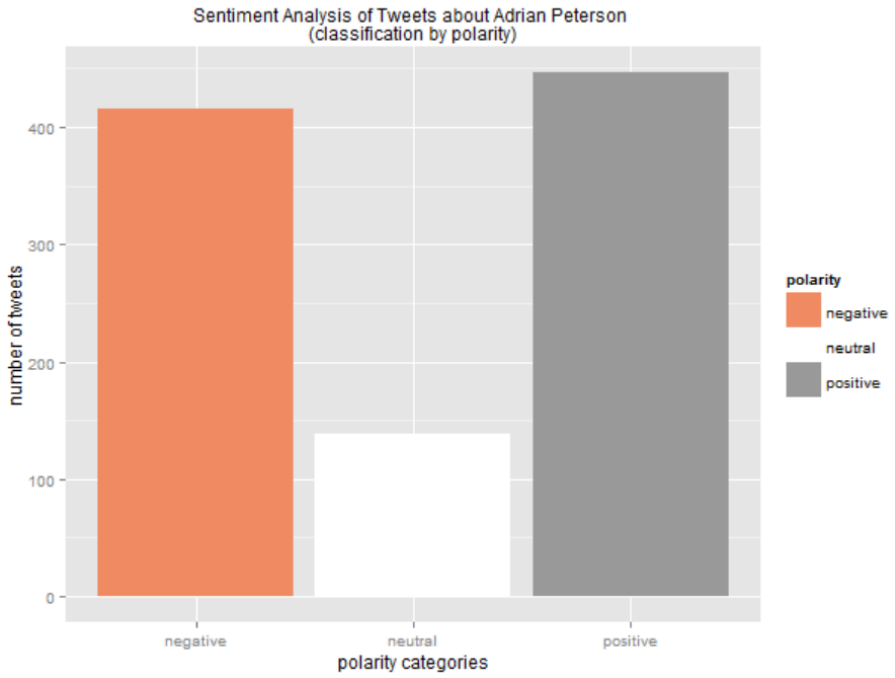
Twitter – Sentiment Exploratory Analysis

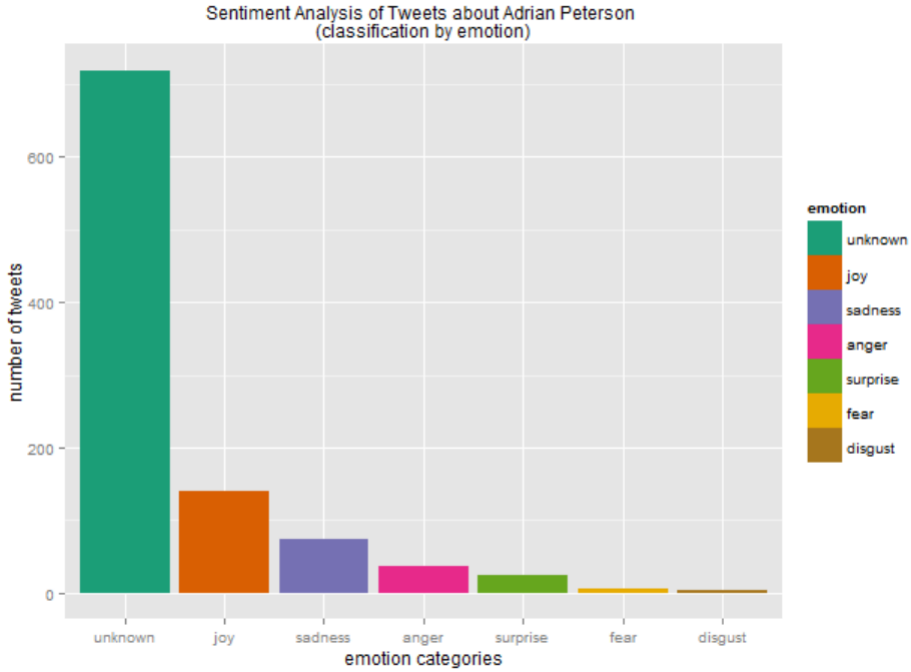
- Adrian Peterson is a professional football player who was suspended by the NFL for overly disciplining his child. This dataset was prepared directly after his hearing with the league to reinstate him. Mr. Peterson was denied this request and we would like to analyze the social media reaction.

Date	Name	Tweet	retweetCount
11/18/14	DariusHoward_PR	#AdrianPeterson is suspended for the remainder of the season w/o pay.	0
11/18/14	Young_Hypocrite	RT @TheRoot: #AdrianPeterson is suspended from the @NFL without pay	11
11/18/14	stroker66ace	@FCC please look into the actions of @1057FMTheFan. I believe they are	0
11/18/14	Adorable_Mikey	Watched my buddy hit his wife and get fired. Hit my kid and got the same	0
11/18/14	LilBill2345	Whether you agree with physical child discipline or not you have to admit	0
11/18/14	DJRobertHorry	homeboy is a dum-dum, but the NFLPA is going to sue the fuck out of the	0
11/18/14	bwolfe23	RT @AceKlubKasanova: My thoughts #NFL #AdrianPeterson http://t.co/S	1
11/18/14	legalspeaks	RT @BringMN: All the developments in the #AdrianPeterson suspension,	2
11/18/14	ItsShanaRenee	NEW! New Rules: #RogerGoodell doesn't fight fair in #AdrianPeterson's s	0
11/18/14	Krismaer53	#AdrianPeterson Not sure having an angry unemployed without pay man	0
11/18/14	Dj_Ango_	Ha "@fishsports: Do we grasp the irony in #AdrianPeterson thinking he's	0
11/18/14	ReddFoxxThePoet	Hm rt"@thetoyman1: #AdrianPeterson has been suspended without pay	0
11/18/14	The_Rob_Wagner	The only people dumber than #AdrianPeterson are the morons defending	0

Twitter – Sentiment Analysis

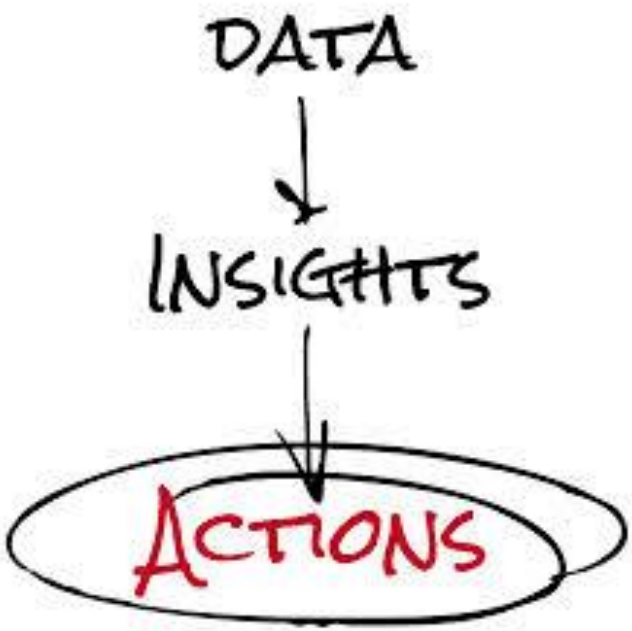
- An analysis of twitter posts reveal that a slim majority of posts have a positive sentiment for the #AdrianPeterson hash tag.





Twitter – Sentiment Analysis

Business Applications of Sentiment Analysis



- One of the most important areas for sentiment analysis, and social media monitoring in general, is bridging the gap between insight and action.
- It's one thing to retrieve a sentiment pie chart. It’s another to masterfully place it within the context of your brand’s social media performance.

Business Applications of Sentiment Analysis

- The key to successful engagement is sentiment prioritization:

- Influence: Because social media mentions are plentiful, prioritization tools must continue evolving. Of the 10,000 tweets and blog posts about your brand, how do you pick the top 50 to focus on?

- Example: If you need to neutralize the mentions that hurt your brand the most, you should drill down into negative mentions, identify the content coming from the most influential people in your industry, understand how far each tweet traveled, and how many people were impacted by this content.

Business Applications of Sentiment Analysis

- **Reputation:** Each notable user should have a social media reputation profile. If someone's negative sentiment indexes higher than average (i.e. that person hates everything equally), then that person's negative sentiment should be somewhat discounted.

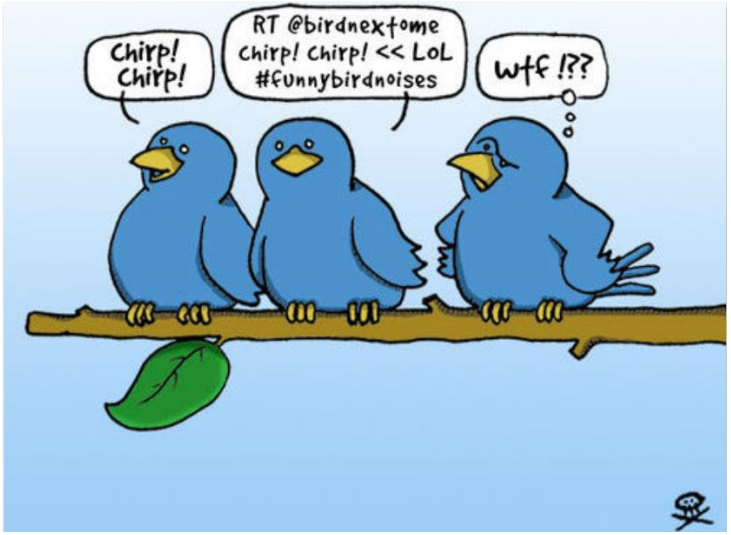
- **Intensity:** As far as sentiment algorithms are concerned, part of a successful prioritization process is going to be identifying the intensity of each mention. "I really hate product X and will never buy it" is quite different from "Product X is running a little slow today." Ability to cross-reference intensity, influence, trajectory, velocity and sentiment of each social media mention will drive us towards a reliable priority system.

- **Isolating content types:** A lot of social media mentions are neutral in nature and some social media sources tend to skew higher on the neutral scale. For example, a higher percentage of updates are neutral on Twitter than any other medium (consider these common examples: "I just had a cup of coffee," or "craving tacos for lunch – who's in?" or "Apple launches the iPad tablet"). Depending on the source you are looking at, your sentiment results will differ — this should be expected. Make sure your sentiment platform allows you to isolate results by content type.

Building off of Sentiment Analysis

- **Sentiment override:** Because automated sentiment is not going to be 100% accurate, you, the user, need to have some kind of override control. When picking a tool, ensure that it allows you to override sentiment, and toss irrelevant results.

- **Entity level vs. article level sentiment:** Until recently, the industry default has been able to measure sentiment at the level of the article. Over time, some platforms have developed ways to measure sentiment on the level of the entity (entity level analysis can measure the sentiment of an entity or multiple entities within an article even if the overall sentiment of the article is different).



Imagine this Scenario

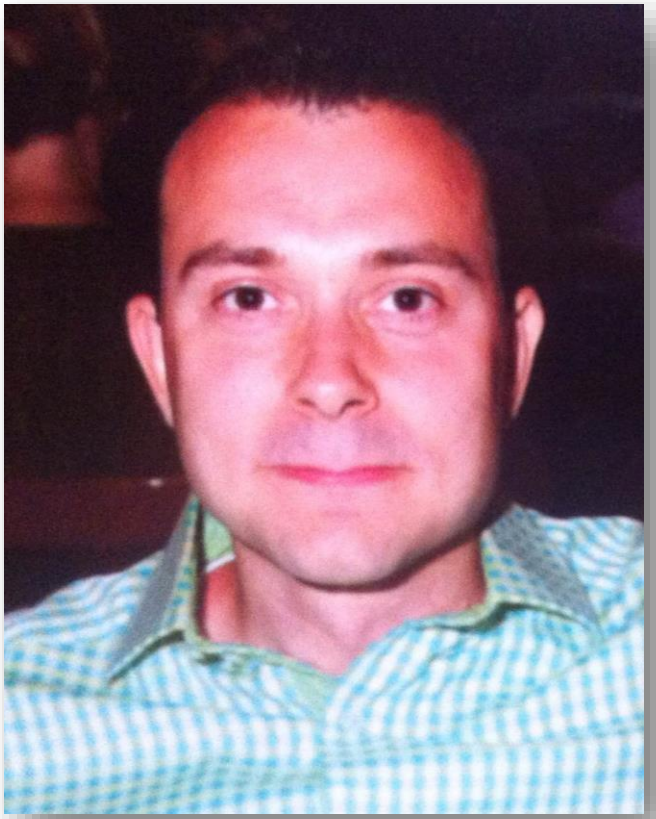
- You wake up to a large spike in social media mentions.
- You drill down to understand sentiment and discover that it's mostly negative.
- Prioritize the negative mentions and portion off 50 most important pieces of content to cover.
- Discover that a flurry of negative tweets originated from an influential social media personality and spread like wildfire.
- You read this person's blog, discover a performance issue that this person had with your product.

Imagine this Scenario

- Leave an honest and humble comment on the blog, committing in public to fixing the issue.
- Reach out to the blogger privately, gain an understanding of what would fix the situation, and then do so.
- All of the above is automatically logged into the system under the blogger's name, which allows you to track the progress of that relationship over time.
- Because you now have a history on this person, and your system cross-references his/her social media profiles across several platforms, you are able to track this person's affinity and sentiment towards your brand over time, monitor their purchase intent, and note their influence on others' purchase decisions.

About Me

- Reside in Wayne, Illinois
- Active Semi-Professional Classical Musician (Bassoon).
- Married my wife on 10/10/10 and been together for 10 years.
- Pet Yorkshire Terrier / Toy Poodle named Brunzie.
- Pet Maine Coons' named Maximus Power and Nemesis Gul du Cat.
- Enjoy Cooking, Hiking, Cycling, Kayaking, and Astronomy.
- Self proclaimed Data Nerd and Technology Lover.



Fine

Acknowledgements

- This presentation consisted of many examples from a variety of subject matter experts. Thank you for your contributions. Please check them out.

- http://www.slideshare.net/mcjenkins/how-sentiment-analysis-works?related=1
- http://www.slideshare.net/gagan1667/opennlp-demo
- http://research.microsoft.com/en-us/groups/nlp/
- http://www.textanalyticsworld.com/wp-content/uploads/2012/03/PracticalTextMining_Excerpt.pdf

Excerpt from: *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications*
G. Miner, D. Delen, J. Elder, A. Fast, T. Hill, and R. Nisbet, Elsevier, January 2012