**SLIDE-1**

**Text Mining & Clustering**

**SLIDE-2**

**Text Mining – Importance**

We have 20% of data in structured format

And 80% of data in unstructured format

* Avenues of textual unstructured data
* Call transcripts
* Email to customer service
* Social media outreach
* Speech transcripts
* Field agents, salespeople
* Interviews & surveys

**SLIDE-3**

**Bag-of-Words**

ENGLISH Professor!!!

All the world is a stage, and all the men and women merely players:

They have their exits and their entrances;

And one man in his time plays many parts…”

**Statistician**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| word | stage | men | woman | play | exit | entrance | time |
| 1 | 1 | 2 | 1 | 2 | 1 | 1 | 1 |

**SLIDE-4**

**Terminology & Pre-processing**

* + Each row is called as a ‘Document’ & even an empty row is considered as a document
  + Collection of all these documents is called as ‘Corpus’
  + Quirks of languages
    - Terms with typos (e.g., ‘musc’)
    - Terms in lowercase, proper case & uppercase (e.g., usb, Usb, USB)
    - Punctuations & special symbols (‘%’, ‘!’, ‘&’, etc.)
    - Filler words, connectors, pronouns (‘all’, ‘for’, ‘of’, ‘my’, ‘to’, etc.)
  + Stemming – process of considering only stem words (e.g., jumping, jumped; stem-word here is ‘jump’)

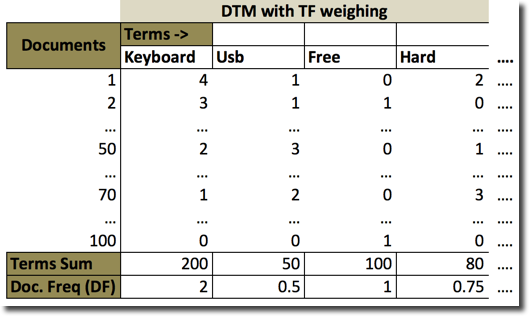
**SLIDE-5**

**DTM & TDM**

Let us understand 100-document corpus of Xbox

**DTM weighing**

* TF - Regular term counts
* TFIDF - Discounts the TF by document frequency



SLIDE-6

Corpus-Level Word Cloud



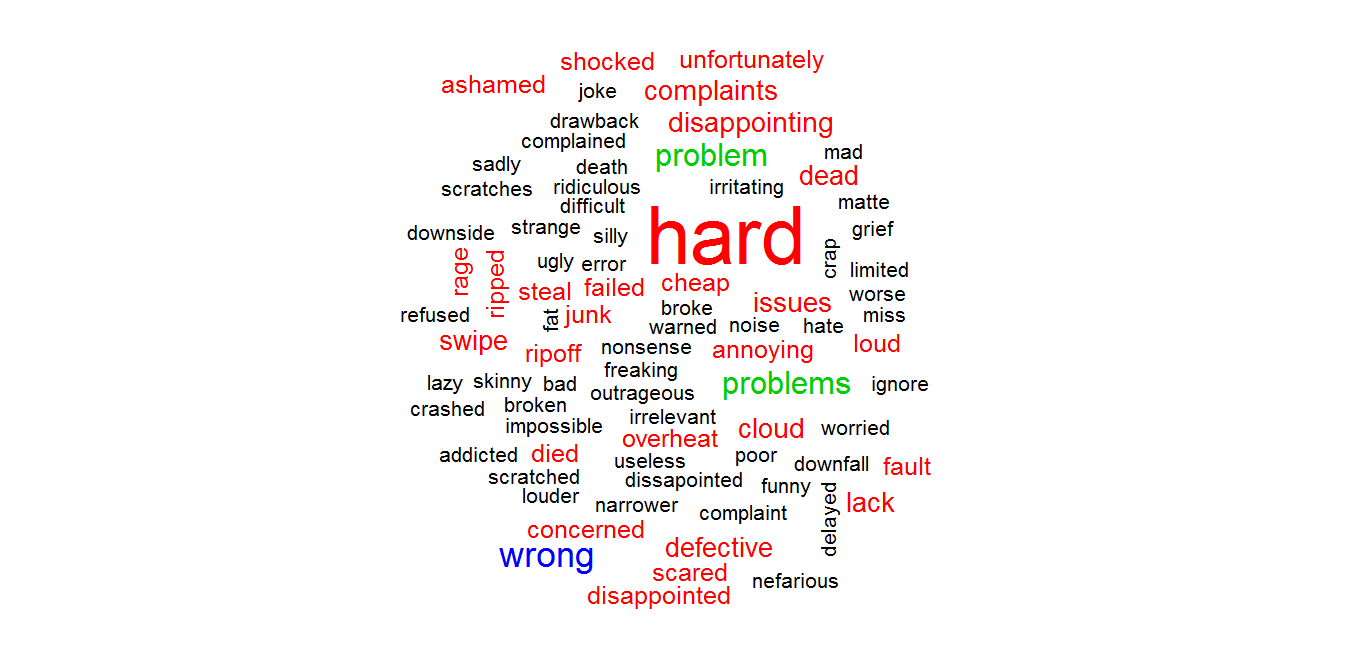
SLIDE-7

Positive Word Cloud



SLIDE-8

Negative Word Cloud



**SLIDE-9**

**Clinical Trials Project**

SLIDE-10

**Clinical Trials – Text Mining**

**Stages**

* + Stage 1: Animals
  + Stage 2: Humans - very few with that specific disease
  + Stage 3: Humans - who have other diseases
  + Stage 4: Humans - larger audience
  + Stage 5: US FDA
  + Stage 6: Adverse events

SLIDE-11

**Clinical Trials – Project in brief**

**Business Objective:** Increase the success rate of the clinical trials

**Project Brief Description:**

Phase 1: Collected the data from open source forums such as “https://clinicaltrials.gov/”

Phase 2: Data Cleansing on XML files by extracting relevant fields from the clinical trials

Phase 3: Segregated the data into Structured & Unstructured data

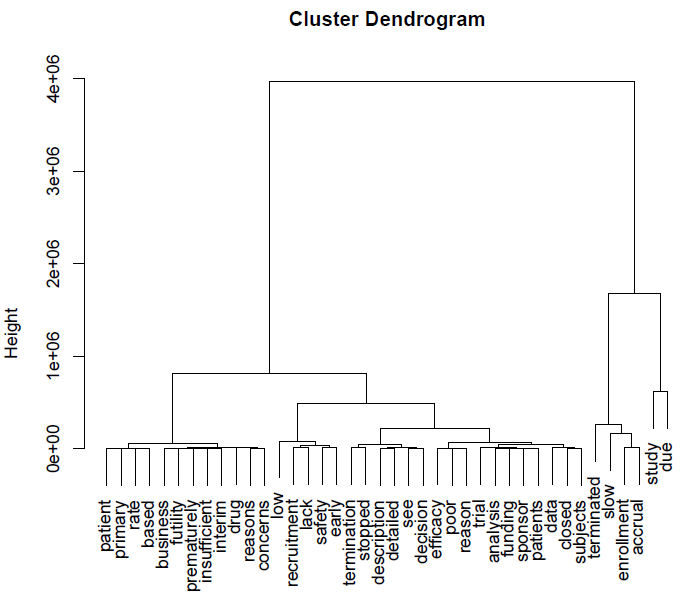
Phase 4: Performed Word Cloud & Sentiment Analysis on unstructured data to identify the reasons for termination of clinical trials

**Techniques used:**

Term Frequency (TF), Term Frequency Inverse Document Frequency (TFIDF), Positive & Negative Word cloud, Dendrogram, Semantic Network, k-Means clustering

**SLIDE-12**

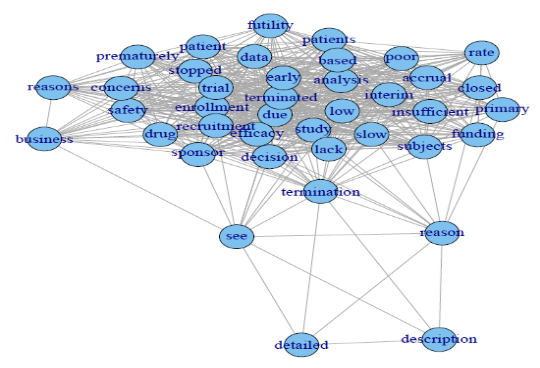
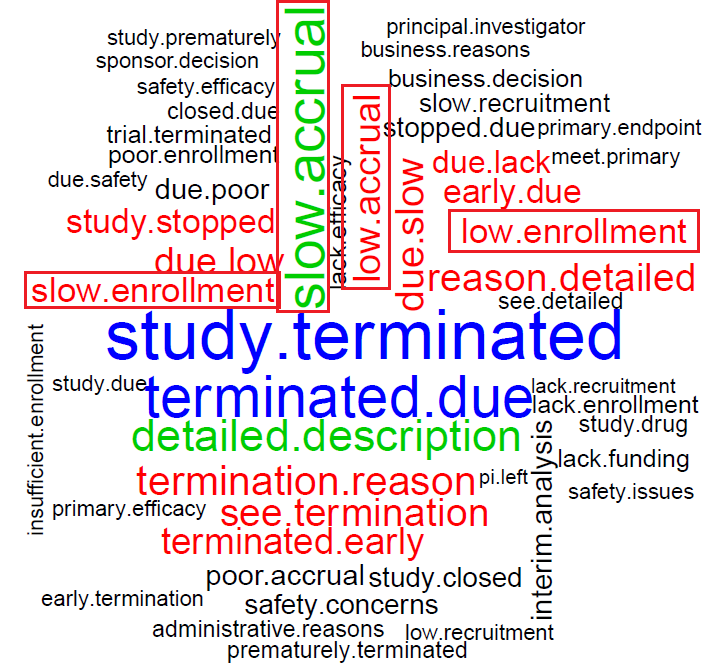
**Unigram Word Cloud & Dendrogram**

**** ****

* Key words standing out of the rest are Accrual, Enrollment, Slow, Safety, Efficacy, Sponsor, Lack, Low etc.
* These words should be seen in the context to gain business value
* When we see this word cloud in conjunction with dendrogram, we notice that slow accrual, slow enrollment, poor efficacy, sponsor funding seem to be the broad themes for termination of clinical trials

SLIDE-13

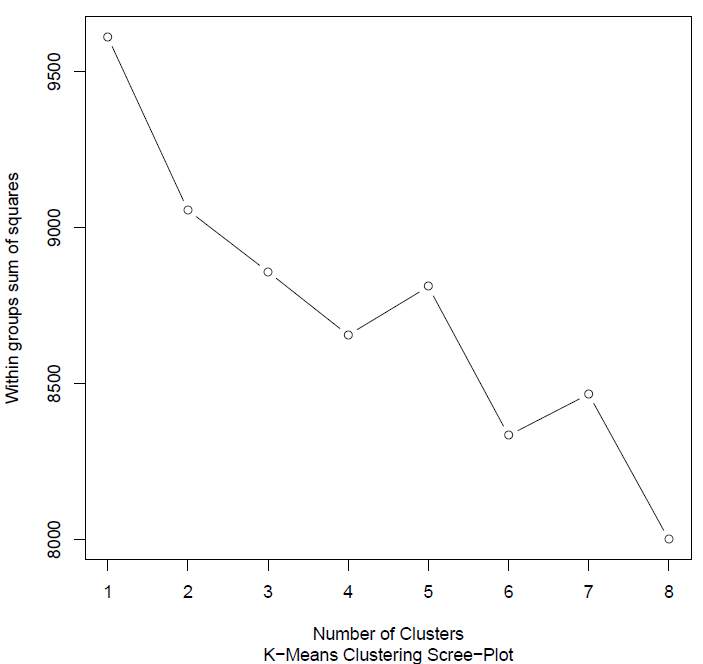
**Bi-gram Word Cloud & Semantic Network**

**** ****

* Semantic network shows that the relationship between the words & the key themes mentioned in previous slide are becoming relevant
* One key thing is safety concerns. At the first sight it sounds as if safety concerns were reason for termination, but when we see it in context, more termination reasons say that there are “No Safety Concerns”
* Bi-gram is used to see 2 words to extract business value & the key themes mentioned earlier are more evident here

SLIDE-14

**K-Means Clustering Scree Plot**

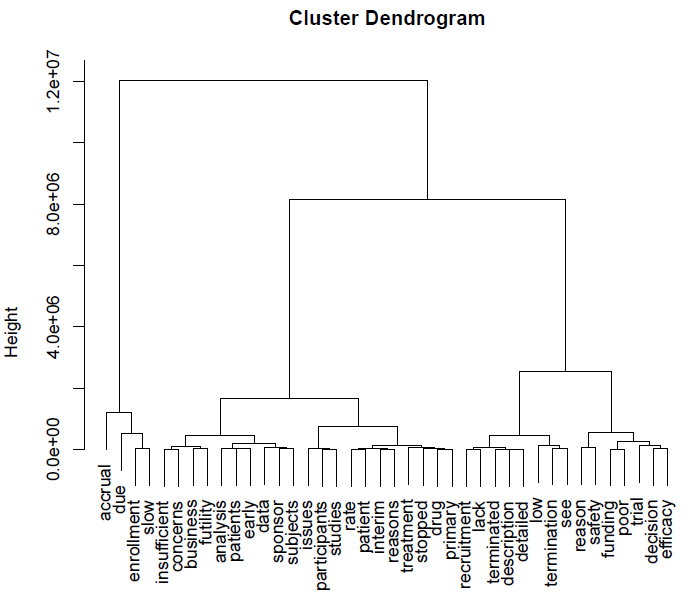
****

* Scree-plot or elbow plot shows that there is a clear bend at 2 clusters, hence we are considering that there are 2 clusters (categories) that the data can be segregated into

Note: Analysis is done considering slight bend at 2nd cluster and considering steep bend at 4th cluster, however, it did not provide any meaningful insights

SLIDE-15

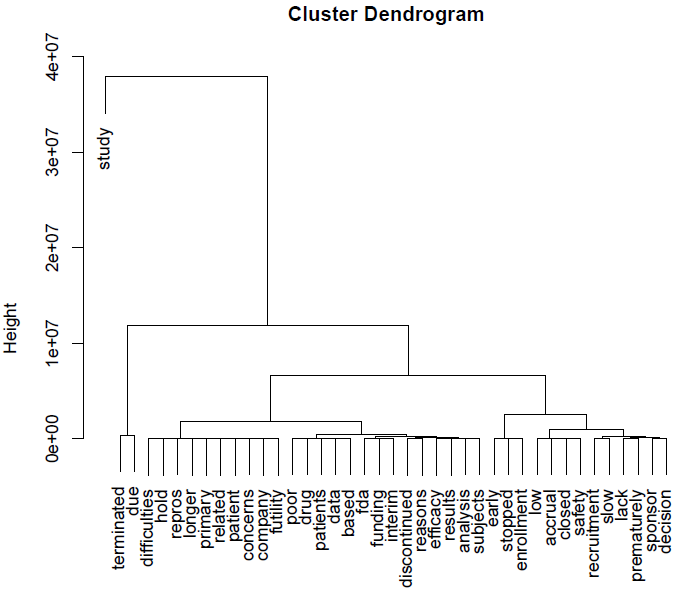
**Word Cloud & Dendrogram - First Cluster**

* Word cloud is clearly highlighting that this cluster is speaking majorly about Accrual: Term referring to the number of patients in a study or clinical trial
* Even the dendrogram clearly shows Accrual, Enrollment, Slow as a major cluster

SLIDE-16

**Word Cloud & Dendrogram - Second Cluster**

* Few key highlights from this word cloud are early & premature termination
* Dendrogram mentions majority of things related to premature closure of clinical trials

SLIDE-17

**Web & Social Media Extraction**

**SLIDE-18**

**NLP Agenda**

* LDA in Text Mining
* Topic extraction using LDA
* Structured information extraction
* Sentiment extraction in a narrative
* Lexicons & Emotion Mining

NLP can be done various languages including Kannada, Telugu, Hindi, Chinese. Use Python for this. Computationally expensive, 8 GB will not cut it for all analysis, so use server or AWS.

Python is very good for web scraping.

LDA Pronounce as Dee rick ley Dikret

Sentiment analysis apart from positivity & negativity

Lexicons are parsers which are open source mostly which are used as dictionaries using these dictionaries we will extract

SLIDE-19

**NLP**

**NLP is used in**

* Search engine
* Spam detection
* Recommendation systems
* Anomaly detection
* Trend mapping
* Social media analysis

**Steps to be followed to make necessary inferences from the data**

* Data collection/  
  Information Retrieval
* Cleaning/
* Normalization
* Feature extraction
* Lexical analysis/ Entity analysis
* Extraction of insight

Search engines use NLP, Google switched from query based search to AI NN. Text analytics is start on AI, teaching engine a pattern

Text mining on scale is search engine because it does this on population other than sample.

Binge uses NN

Spam detection – Google is able to achieve 100% accuracy, LinkedIn is not so good. Raw text is read and understands whether it is spam or not. Bayesian classifier, mixture model

Recommendation system – Amazon

Anomaly detection – Narratives, twitter, hash tag, product sentiment in Twitter – Positive sentiment & suddenly it is negative now. Corrective action. R ngram package– sequence of words and help you understand n+1 the word sentiment using Hidden Markov. Subtle disturbance in a sequence of words trigger an alarm. Hidden Markov is useful to understand the aberrations. Stock market index going on and suddenly something goes wrong

Trend mapping – Form of Anomaly detection. Trend pattern

Social media analysis – Scrape web data & sentiment

Alchemy algorithm - sentiment mining, scrape web & positivity & negativity & provide output in JSON format

Feature extraction = DTM/TDM

Lexical analysis – Bag of words to mine sentiment; Entity analysis = treating with labels to do predictions

Extraction of insight – Text analytics, Hypothesis testing

SLIDE-20

**Latent Dirichlet Allocation (LDA)**

* LDA is a generative model (a model for randomly generating observable data values)
* Observations are words collected into documents
* It assumes that each document is a mixture of a small number of topics
* Each word’s presence is attributable to one of the document’s topics
* LDA can be viewed as a Bayesian model, where each item is modeled as a result of a mixture of underlying set of topics
* Each document may be viewed as a mixture of various topics

LDA – developed by Andrew, Assistant professor at Stanford, Chief scientist in Baidu China’s search engine, Develop new algorithms.

David Blei, Andrew Ng, and Michael I. Jordan in 2003

Raw text -> generate data value -> assign labels; If I say 10 sentences then each sentence is a document & it has text features in it.

Assume I am giving a speech; If I speak then the statement might mean multiple things in different contexts, LDA will allow for duplication of that context space; Bayesian model; prior probability, calculate posterior probability.

SLIDE-21

**Latent Dirichlet allocation (LDA) Vs. Clustering**

|  |  |
| --- | --- |
| **LDA** | **Clustering (K-means)** |
| * + Unsupervised learning algorithms   + Mixture model where a document can be assigned to one or more topics   + Each topic is a culmination of multiple documents   + A popular example using term usage   + A man sees a boy with a telescope   + Who has the telescope? | * + Unsupervised learning algorithms   + Specify an optimal ‘k’ that allows us to extract topics or segments from the data   + Does a raw partition of the data   + Resultant clusters are disjoint from each other   + In this example a term’s usage leads to confusion owing to its placement   + In the same way a sentence from a corpus could infer a different meaning in conjunction with another sentence |

LDA is mixture /Gaussian model; LDA will help you identify or allocate a sentence to different context

Let us do R now. NLTK is most important package in Python for text mining & tm is the most important package for R in text mining

The initial cluster distances in Ward's minimum variance method are therefore defined to be the squared Euclidean distance between points:

{\display style d\_{ij}=d(\{X\_{i}\},\{X\_{j}\})={\|X\_{i}-X\_{j}\|^{2}}.} d\_{{ij}}=d(\{X\_{i}\},\{X\_{j}\})={\|X\_{i}-X\_{j}\|^{2}}.

The difference between ward.D and ward.D2 is the difference between the two clustering criteria that in the manuscript are called Ward1 and Ward2.

It basically boils down to the fact that the Ward algorithm is directly correctly implemented in just Ward2 (ward.D2), but Ward1 (ward.D) can also be used, if the Euclidean distances (from dist()) are squared before imputing them to the hclust() using the ward.D as the method.

For example, SPSS also implements Ward1, but warn the users that distances should be squared to obtain the Ward criterion. In such sense, implementation of ward.D is not deprecated, and nonetheless it might be a good idea to retain it for backward compatibility.

Use LDA on resume also

SLIDE-22

**Structured data extraction**

* Many sources of data contain large amount of artifacts that lend a lot of information
* Text data can be subjected to methods that can help mine structured information
* This is information retrieval using previously generated labeled data
* Raw Text
* Parser
* Names Entities
* Data from healthcare – Huge data, clinical trials are verbose;
* Parser is a group of lexicon or dictionary and we do a Boolean. Computationally expensive so do it on server
* openNLP used to be 80MB file so cran decided not to made modules of NLP. If you want dependencies it is hosted on data cube
* openNLPmodels.en - German, France, etc.
* Maxent\_Word\_Token\_Annotator () maximum entropy -> We have sentence & certain repetition of words & sentences in term frequency, probability is max times a word repeats in a corpus. Based on this repetition it will do a function & give the total frequency; TF with more focus

SLIDE-23

**Lexicons**

* Lexicons serve as dictionaries for extracting sentiment from raw unlabeled data
* These are useful in estimating semantic orientation (polarity)
* They are applied to polarity prediction tasks and serve as a bag of words that help assign a score/label to terms in text

**Three Lexicons used in this session are:**

* + [Bing](https://www.cs.uic.edu/~liub/) (Developed by Professor Bing Liu)
  + [AFINN](http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010) (Informatics and Mathematical Modelling, Technical University of Denmark)
  + [NRC](http://www.saifmohammad.com/WebPages/ResearchInterests.html) (Dr. Saif M. Mohammad)
* English linguists have come up with lexicons and made it open source for us.
* Sentiment polarity -5 to +5. How a company is faring in social media market
* Bing Liu maintains and freely distributes a sentiment lexicon consisting of lists of strings.
* Distribution page (direct link to rar archive)
* Positive words: 2006
* Negative words: 4783
* Useful properties: includes mis-spellings, morphological variants, slang, and social-media mark-up
* The NRC Emotion Lexicon is a list of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). The annotations were manually done by crowdsourcing.
* AFINN is a list of English words rated for valence with an integer between minus five (negative) and plus five (positive). Finn Arup Nielsen has manually labeled the words in 2009-2011. The file is tab-separated. There are two versions:
* AFINN-111: Newest version with 2477 words and phrases.
* AFINN-96: 1468 unique words and phrases on 1480 lines. Note that there are 1480 lines, as some words are listed twice. The word list in not entirely in alphabetic ordering.

**SLIDE-24**

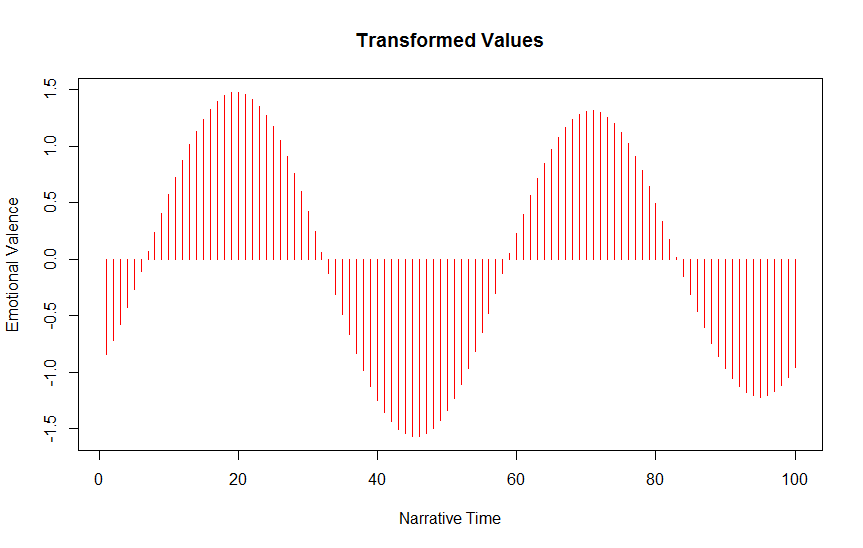
**Emotion Mining**

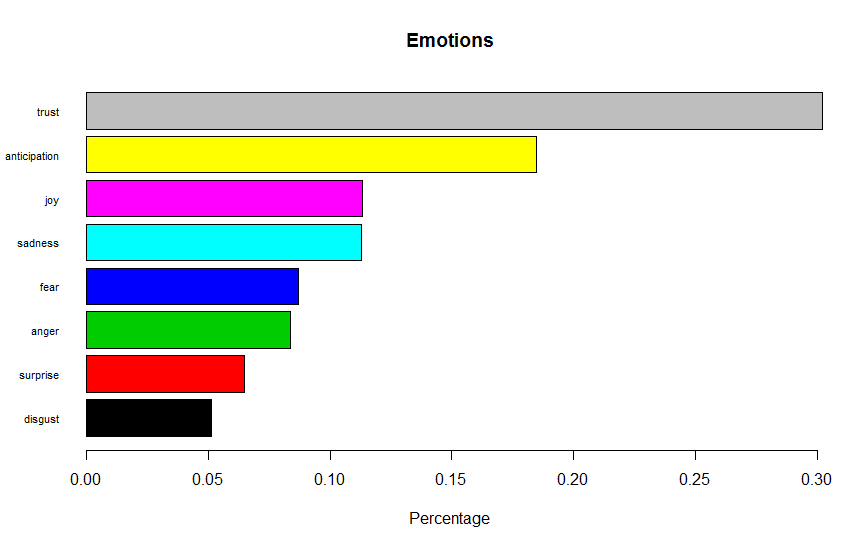
This methodology allows the extraction of the most negative to positive sentiment bearing documents from text:

* negative <- s\_v[which.min(afinn\_s\_v)]
* > negative
* [1] "I fully agree with you. This is the worst card ever and they are running a late charge fee scam here. This is what they do - on my first statement, I made full payment. On the second statement, they charged me a late payment fee and interest. I wrote back to them to tell them they made mistake but never heard back from them. I made full payment on the second statement, which was around $26, and when the third statement came, I was charged with another late payment fee and compounded interests. This time, I got on the phone and spoke to some guy in India who starts off each statement with \"So how do you want to make payment sir\" After something like hearing that damn statement for the 10th time, I got so pissed with trying to get an explanation on the late fees, I was transferred to a supervisor who then said that my first payment got rejected. I then asked why was it rejected (for which they had no clue and said that I should check with my bank) but got nowhere but what I was really pissed about was that the unpaid amount due in the first statement was never reflected on the second statement like other typical credit cards from REPUTABLE banks. Had this been done, I would have known and paid in full with the second payment! In the end, they said that they would waive the interests, which was only a few dollars, but gad to charge me the late fee of $35 per month ($70 in total). Can you imagine if they run this \"sweat shop\" practice from India and suckered in 10,000 people with this practice? That would have been $700,000 into their pockets without breaking a SWEAT! Worst still, they filed my 2 months delinquencies with the credit unions and I lost 50-70 credit score points! Bastards! ”
* > positive <- s\_v[which.max(afinn\_s\_v)]
* > positive
* [1] "\"Well, you got me beat. I have 11 currently. However, I do have several Chase cards (CSP, Marriott Rewards and the Freedom) and to be honest I rarely use my Amazon.com Rewards due to better reward options elsewhere; Barclay's SallieMae is the best card for Amazon.com purchases. 5% back on up to 750 in purchases each month. It also has 5% on gas/groceries on up to 250 on purchases, in each category per month. I really do not like the small cap of 250 on groceries and use Amex BCE for that purpose, but it’s still an extra benefit to have."

SLIDE-25

**Arcs & Emotion**

****

****

* This method allows for generating narrative time for the text and generates how the positive and negative emotional valence has been in the corpus
* The nrc framework allows to mine text that match with eight emotions in the writing:
  + - Trust
    - Anticipation
    - Joy
    - Sadness
    - Fear
    - Anger
    - Surprise
    - Disgust
* The same can be implemented for the different sources from which content is generated for a client, and the outcomes can be compared and evaluated for understating which source provides what form of understanding

Valence, as used in psychology, especially in discussing emotions, means the intrinsic attractiveness (positive valence) or aversiveness (negative valence) of an event, object, or situation. However, the term is also used to characterize and categorize specific emotions.

Stanford NLP parser can be used to mine sentiments

J. R. R. Tolkien bases them on the novel The Lord of the Rings. The films are subtitled The Fellowship of the Ring (2001), The Two Towers (2002) and The Return of the King (2003).

Whole Lord of Rings is towards negative. He started with Hobbit, which is more for children; the he extends that to adults also.

Positive - Starts

Negative Bilbo Baggins leaves,

Positive why-did-frodo-start-his-adventure-17-years-after-he-inherited-the-one-ring

Negative - Gandalf dies