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

	DTM with TF weighing				
Documents	Terms ->				
	Keyboard	Usb	Free	Hard	....
1	4	1	0	2	....
2	3	1	1	0	....
...	...	...	...	...	....
50	2	3	0	1	....
...	...	...	...	...	....
70	1	2	0	3	....
...	...	...	...	...	....
100	0	0	1	0	....
Terms Sum	200	50	100	80	....
Doc. Freq (DF)	2	0.5	1	0.75	....

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

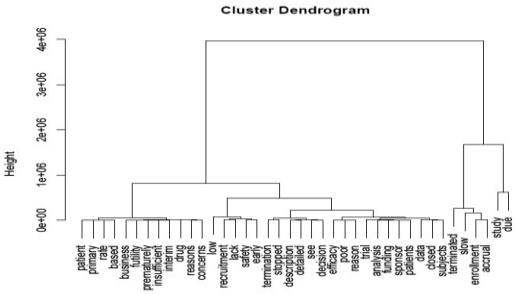
Phase 4: Performed Word Cloud & Sentiment Analysis on unstructured data to identify the reasons for termination of clinical trials

Techniques used:

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

SLIDE-12

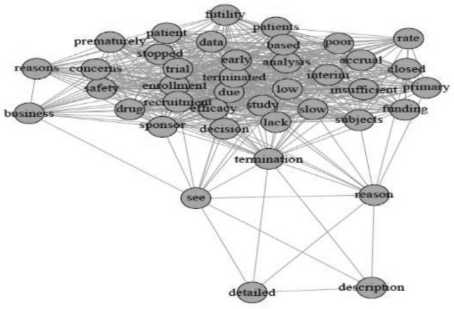
Unigram Word Cloud & Dendrogram



- 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



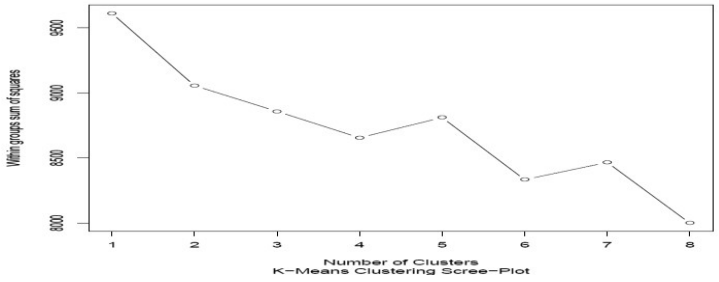
- Semantic network shows that the relationship between the words & the key themes mentioned in previous slide are becoming relevant
- One key thing is safety concerns. At the first sight it sounds as if safety concerns were reason for termination, but when we see it

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

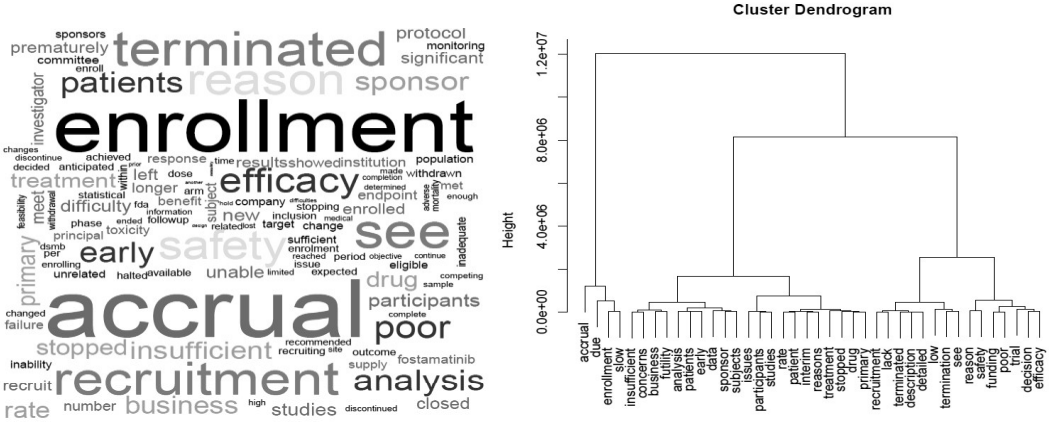
K-Means Clustering Scree Plot



- 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

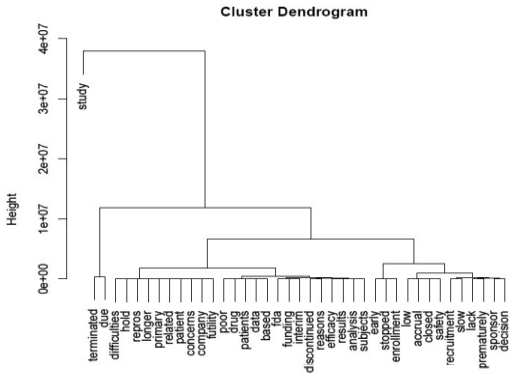
Word Cloud & Dendrogram - First Cluster



- 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