ANALYTIXLABS

Text Mining (Analytics)

Agenda

- ✓What is Text Mining/Text Analytics?
- ✓Major Areas of Text Analytics
- ✓Text Mining Framework & process
- ✓Detailed Steps in Text Mining
- ✓Text Data processing
- ✓Concept of Natural Language Processing
- ✓Text Visualization
- ✓Role of Machine Learning in Text Analytics
- ✓Text mining Use cases

Text Mining (Analytics)

Types of Data

Structured Data

- Loadable into a spreadsheet
  - Rows & columns
  - Each cell filled, or could be filled
  - Data is consistent, uniform
- Data Mining Friendly

Un-Structured Data

- Not Structured into 'Cells'
  - Variable record length; notes, free-form survey answers
  - Text is relatively sparse, inconsistent, and not uniform
  - Also... Images, video, music, etc.

Types of Un-Structured Data

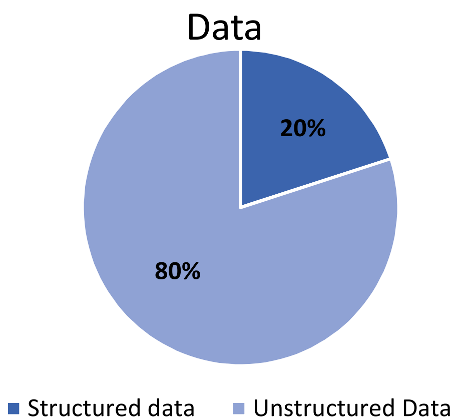
- **Weakly Structured data:** few structural cues to text based on layout or markups
  - Research papers
  - Legal memoranda
  - News stories
- **Semi structured data:** extensive format elements, metadata, field labels
  - EMAIL
  - HTML Web pages
  - Pdf files/ XML web pages /JSON Data
  - Log files

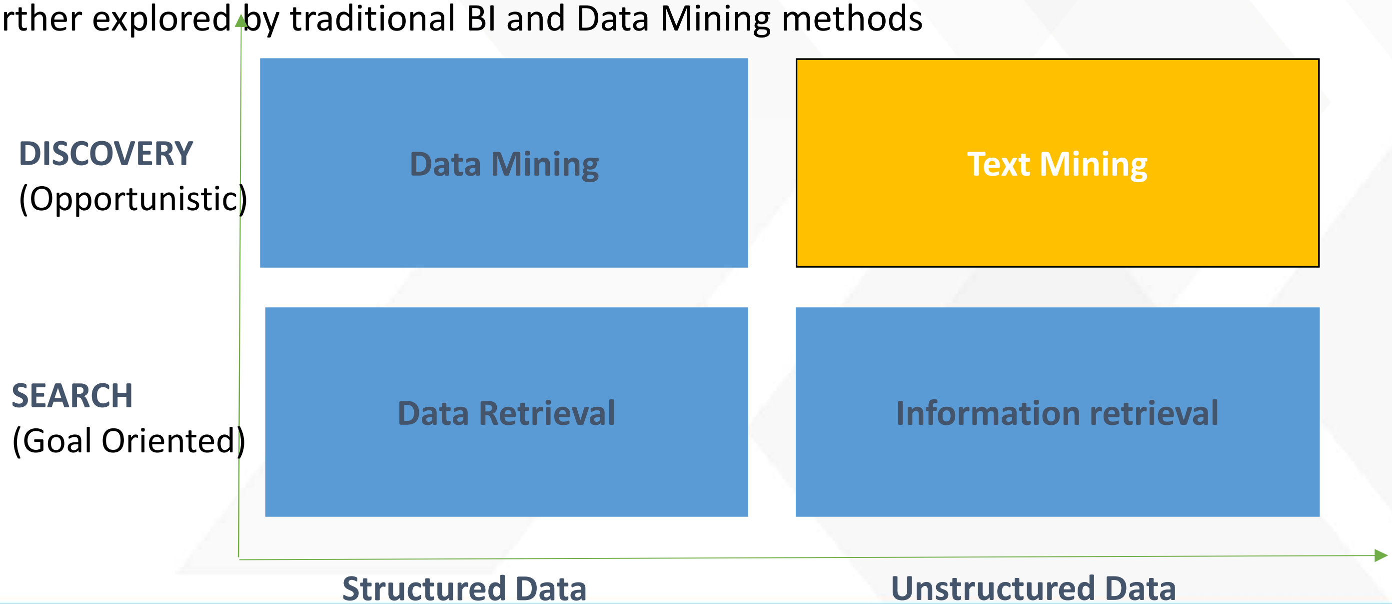
What is text mining?

Text Mining and Text Analytics are broad umbrella terms describing range of technologies for analyzing and processing semi-structured and unstructured data

Text Mining (aka Text Analytics) is the discovery by computer of new, previously unknown information, by automatically extracting information from (large amount of free form) textual data.

Text mining starts by extracting key points, opinions, people, actions, events from textual sources thus enabling forming new hypotheses that are further explored by traditional BI and Data Mining methods

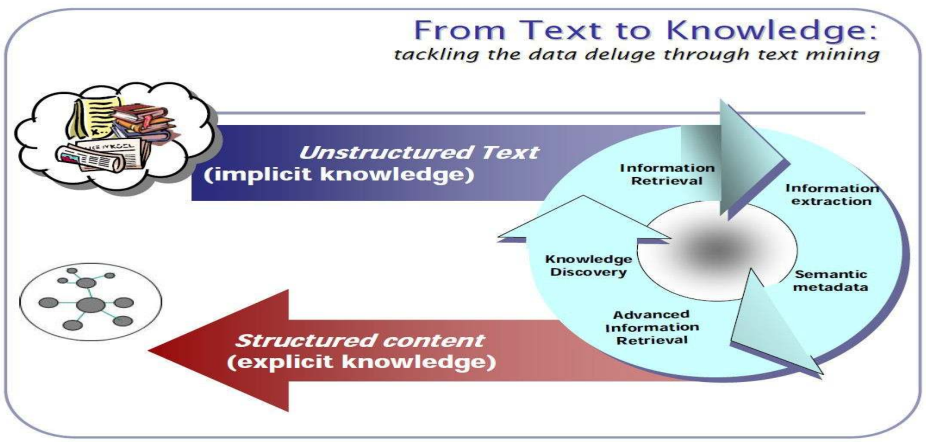


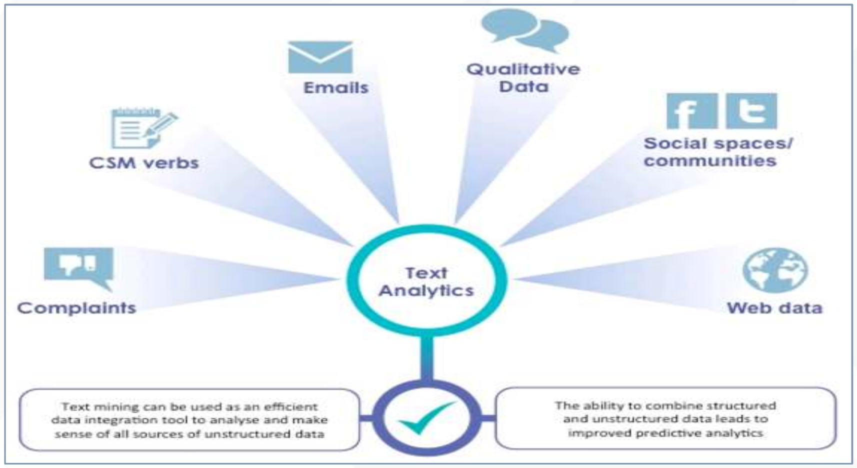


Sources of Data for Text Mining?

Sources are highly varied –

- Web sites, news & journal articles, images, video.
- Blogs, forum postings, and social media.
- E-mail, Contact-center notes and transcripts; recorded conversation.
- Surveys, feedback forms, warranty & insurance claims.
- Office documents, regulatory filings, reports, scientific papers





Data Analytics Vs. Text Analytics

Data Analytics

Customer analytics

- Profiling and segmentation of customers
- Customer retention
- Profitability analysis

Operational analytics

- IT infrastructure optimization
  - Capacity & performance
  - Workload characterization
  - Change detection
- Scheduling and optimization

Financial & Risk analytics

- Financial & sales forecast
- Pricing
- Credit risk - default prediction

Fraud detection and prevention

- Discover anomalous behavior
- Model various types of frauds

Text Analytics

Analysis of free text in customer surveys

- Automate customer clustering/segmentation
- Sentiment Analysis / Feedback Analysis
- Derive insights about customers

Analysis of call center transcripts

- Improve productivity in the call centre
- Call center performance
- Contextual feedback on customer experience

Analysis of forums, blogs and comments

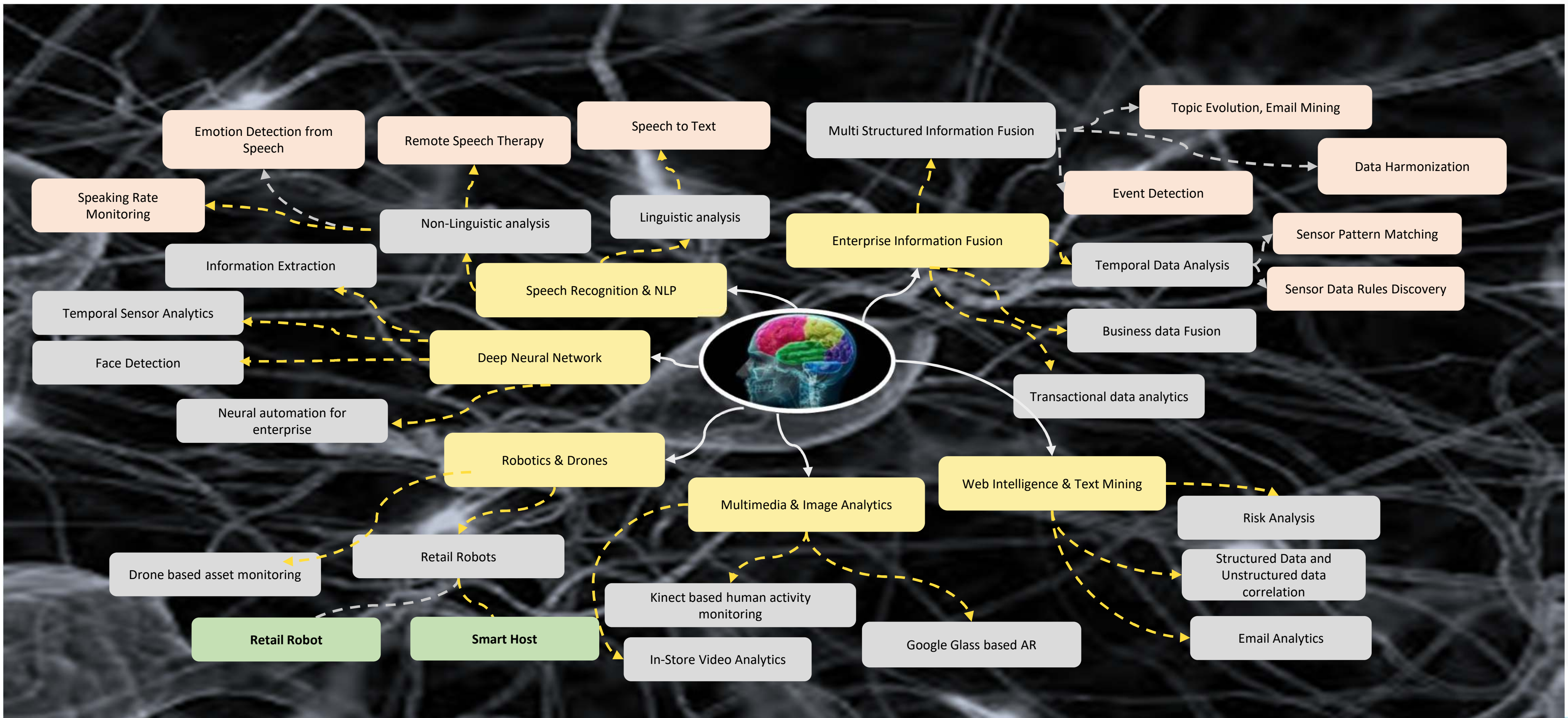
- Understand true voice of customer
- Understand social networks

Analysis of free text within the Enterprise

- Auto-summarization of documents, reports for exec level knowledge sharing
- Reduce costs and enhance speed of knowledge publishing

Text mining techniques: IE, Keyword extraction, Clustering, Summarization using lexical chain

Artificial Intelligence – Text Mining

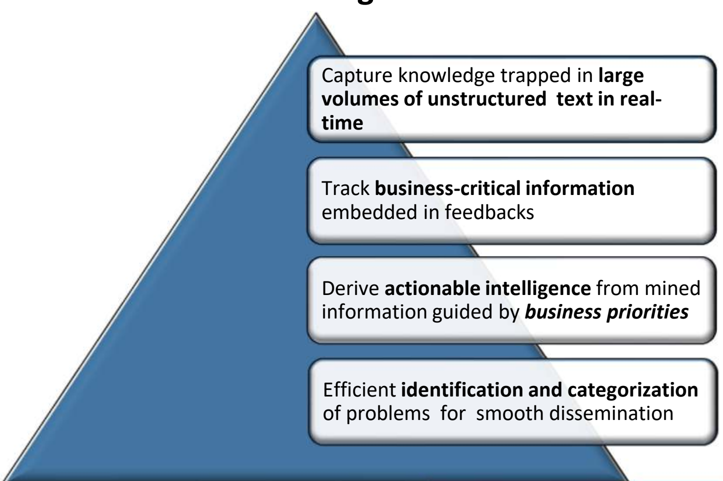


ANALYTIXLABS

What does Text Mining bring to business?

- Link structured and unstructured data to better know consumer
- Ad hoc studies to help marketing decisions
- Understand the consumer reactions
  - Answer a Marketing hypothesis: "What if I change from XX to YY?"
- Dashboard for Top Management
  - Extract all verbatim related to major consumer issues: “browsing speed, cheap broadband, call drop"

Business Challenges



Each of these teams have specific and overlapping needs

CEO	DECISION SUPPORT
CUSTOMER CARE	CUSTOMER OPNION SURVEY
MARKETING	PRESS RELEASE/WATCH
SALES	ENTERPRISE INTERNAL INFORMATION ANALYSIS (SALES VISITS)
HUMAN RESOURCE	SKILL ANALYSIS (CV, INTERNAL KNOW HOW)
NETWORK	CUSTOMER OPNION SURVEY

What does Text Mining bring to business?

Area	More Common Use Cases
Business	Competitive intelligence, document categorization, HR (voice of the employee), records retention, risk analysis, website navigation
Marketing	Voice of the customer, social media analytics, churn analysis, survey analysis, market research
Analytics	Fraud detection, e-discovery, warranty analysis, medical research
Education	Syllabus classification (compliance analysis), GRE, SAT (writing analysis)
Law Enforcement	Crime and terrorism detection, psychological assessments based on written data (by suspects), fraud detection

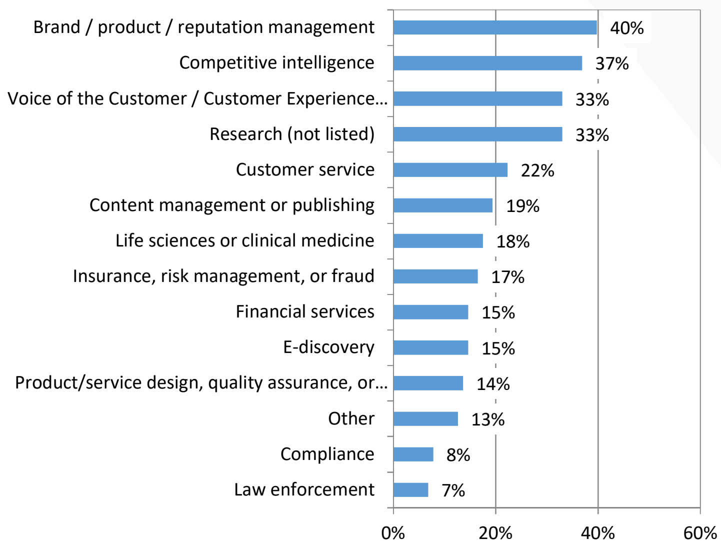
What does Text Mining bring to business?

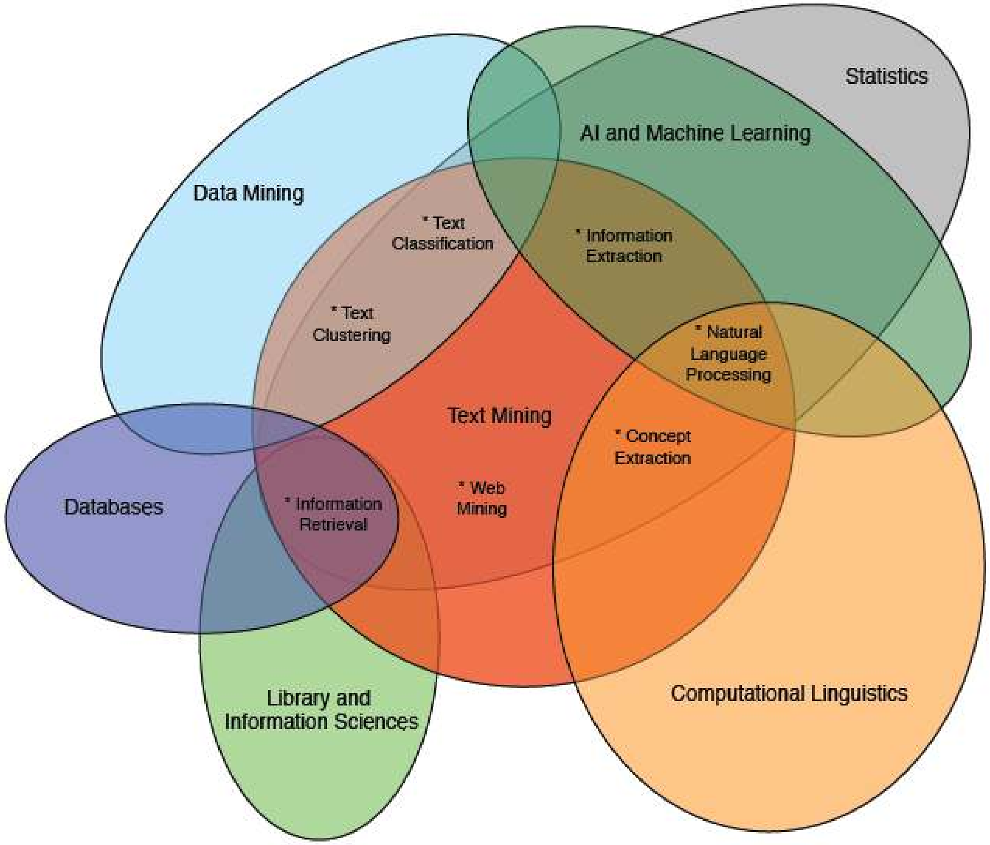
The automatic analysis of text information can be used for several general purposes, such as to:

- Provide an overview of the contents of a large document collection
- Generate summaries and categorize documents in the most efficient way
  - e.g., news, emails, call center/helpdesk inquiries
- Identify hidden patterns between documents or groups of documents
  - e.g., customer complaints, warranty claims, free form survey data
- Increase the efficiency and effectiveness of a search process to find similar information
- Analyze textual information with other structured information to build models
  - e.g., predict customer satisfaction, claim fraud, drug efficacy
- Detect duplicate information or documents in an archive

Primary Applications & Analyzed Textual Information

What are your primary applications where text comes into play?





What textual information are you analyzing or do you plan to analyze?

Blogs and other social media (twitter, social-network sites, etc.)	62%
News articles	55%
On-line forums	41%
E-mail and correspondence	38%
Customer/market surveys	35%

Text Mining - Application areas

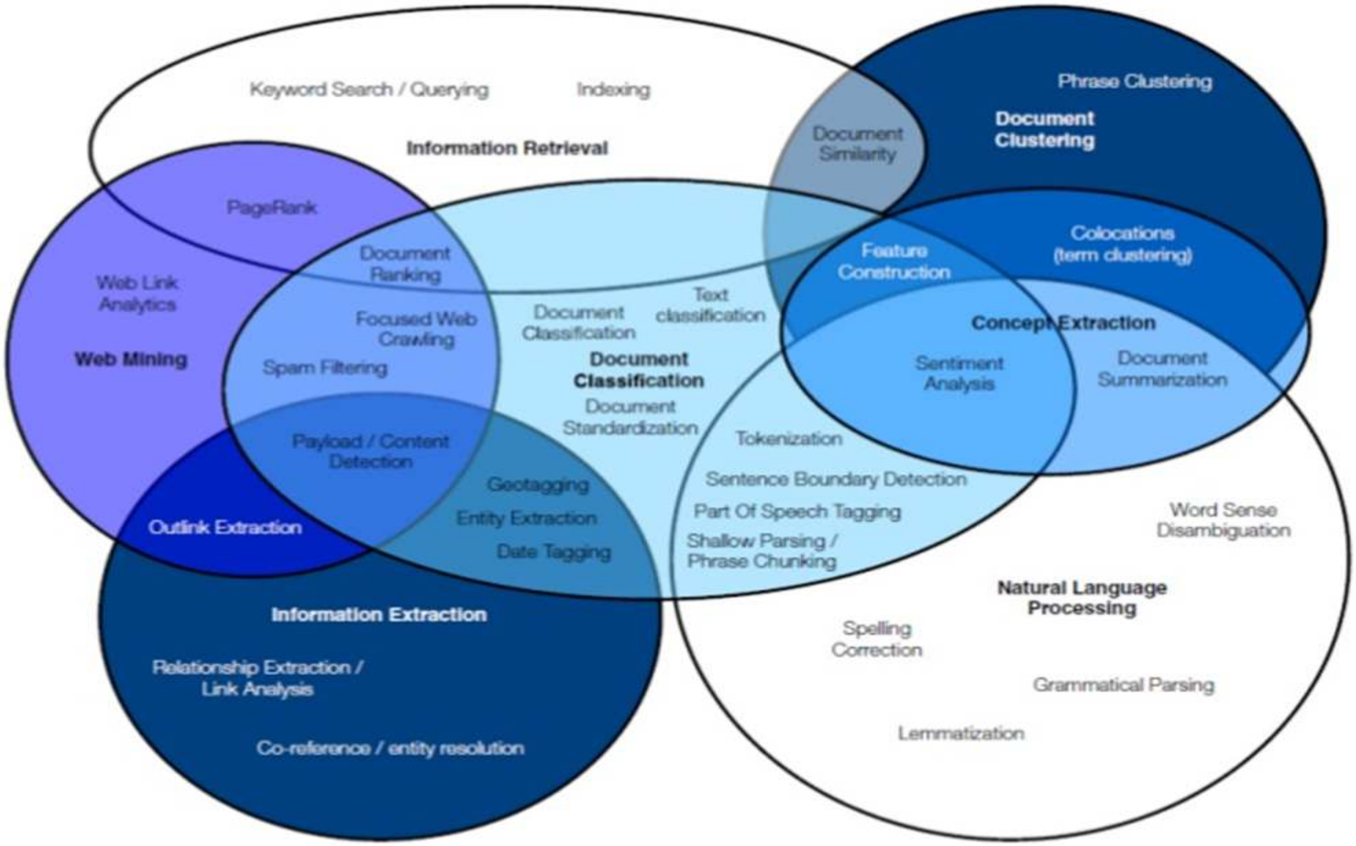
- **Search and Information Retrieval (IR):** Storage and retrieval of text documents, including search engines and keyword search
- **Document Clustering:** Grouping and categorizing terms, snippets, paragraphs or documents using data mining clustering methods
- **Text Categorization/Document classification:** Grouping and categorizing snippets, paragraphs, or documents using data mining classification methods, trained or labelled examples
- **Web Mining:** Data and text mining on the internet, with specific focus on the scale and interconnectedness of the web
- **Information Extraction (IE):** Identification and extraction of relevant facts and relationships from unstructured and semi structured text
- **Natural Language process (NLP):** Low-level language processing and understanding tasks (Ex: Tagging parts of speech); often used synonymously with computations linguistics
- **Concept Extraction:** Grouping of words or phrases into semantically similar groups
- **Association between Terms/Word clustering:** Discovering associations between terms
- **Text Summarization:** Summarizing large amount of textual and factual data.

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 keywc	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 Mining in Telecom-Online Commerce

- Text analytics is applied for marketing, search optimization, competitive intelligence.
  - Analyze social media and enterprise feedback to understand the Voice of the Market:
    - Opportunities
    - Threats
    - Trends
  - Categorize product and service offerings for on-site search and faceted navigation and to enrich content delivery.
  - Annotate pages to enhance Web-search findability, ranking.
  - Scrape competitor sites for offers and pricing.
  - Analyze social and news media for competitive information.

Text Mining in Telecom-Online E-Discovery & Compliance

- Text analytics is applied for compliance, fraud and risk, and e-discovery.
  - Regulatory mandates and corporate practices dictate –
    - Monitoring corporate communications
    - Managing electronic stored information for production in event of litigation
  - Sources include e-mail ,news, social media
  - Risk avoidance and fraud detection are key to effective decision making
    - Text analytics mines critical data from unstructured sources
    - Integrated text-transactional analytics provides rich insights

Text Mining in Telecom-Online Voice of the Customer

- Text analytics is applied to enhance customer service and satisfaction.
  - Analyze customer interactions and opinions –
    - E-mail, contact-center notes, survey responses
    - Forum & blog posting and other social media

      … to ..
    - Address customer product & service issues
    - Improve quality
    - Manage brand & reputation
  - If qualitative information from text can be linked , the following become possible–
    - Link feedback to transactions
    - Assess customer value
    - Understand root causes
    - Mine data for measures such as churn likelihood

Application in Insurance Domain – Improve CRM, Product

- Customers call into call centers / send e-mails / express opinions in surveys or blogs that can indicate their likes / dislikes or sentiments. Customer contact staff can recognize when a customer is at risk of leaving and take appropriate action to **reduce churn** and to **increase satisfaction level**
- Customer calling in to ask about their insurance policy and rep types in or records what the person is saying, and it could prompt a call center person to take an action, such as offering the person a certain price special at that time thus presenting an opportunity for **cross-sell/up-sell**
- Insurers can monitor the quality and effectiveness of the reps taking the phone calls by analyzing hand-written / typed notes, voice files converted to text and **improve** their **productivity**
- Analyzing unstructured data from data inside and outside an enterprise can give insights for strategic planning, **developing new products**
- Can help in achieving a clearer view of the **competitive landscape**

Application in Insurance Domain – Streamline Claims Process

- Conversion of text data combined with structured data can create a complete claim record providing a **360° view of all relevant claim data** for analysis
- Discover **fraud patterns** hidden in Claim Adjuster Notes, Emails, Service notes, Claimant Interviews such as "stopped for no reason", "high usage of technical terms by insured", "felt like a set up", "gap in bills"
- Detect **Fraud rings** by analyzing linkages between different entities, keywords occurring together
- Text mining logs/ notes/ recorded statements can help in finding patterns in missed **Subrogation opportunities**
- A focus on text can uncover inconsistent use of red flags across claim examiners and identify training opportunities for **productivity improvement**

Features of Text Data & Challenges

- ✓ High dimensionality Large number of features
- ✓ Multiple ways to represent the same concept.
- ✓ Highly redundant data.
- ✓ Unstructured data.
- ✓ Easy for humans, hard for machine. Abstract ideas hard to represent
- ✓ Huge amount of data to be processed.

Typical Text Mining Steps

Text Mining Model

In descriptive mining, the textual comments includes providing detailed information about the terms, phrases, and other entities in the textual collection, clustering the documents into meaningful groups, and reporting the concepts that are discovered in the clusters.

In Predictive mining, involves classifying the documents into categories and using the information that is implicit in the text for decision making.

Where does it works?

- Identify and respond of telecom customer experiences into valuable business driven strategies
- Gain a competitive advantage by monitoring the online reputation and that of competitors
- Automatically cluster and categorize call center logs to identify high-volume issues
- Monitor and forecast sentiment prior to and during a product launch
- Identify emerging issues / problematic areas before they become costly problems

Text mining Process

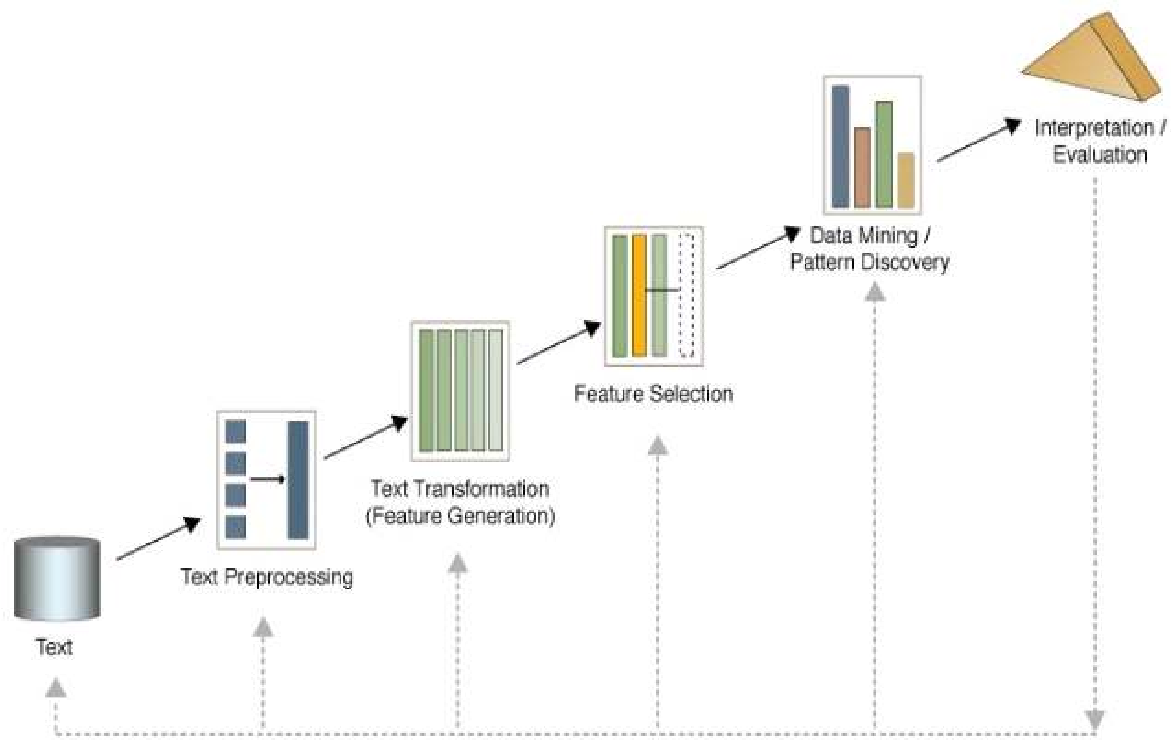
Text Pre Processing
Syntactic/Semantic Text Analytics

Feature generation
Bag of words

Feature selection
Simple Counting
Statistics

Text Data mining
Classification-Supervised learning
Clustering-Unsupervised learning

Analyzing Results
Mapping/Visualization
Result interpretation



Iterative and interactive process

How is text mining done?

Collection	Pre-processing	Transformation (Feature generation/selection)	Using algorithms (Text/Data Mining)	Evaluation (Analyzing results)

Collection

Unstructured data for our analyses can be found in these various forms :

- Text files/CSV
- JSON files
- HTML files
- Excel files
- Websites

The objective is to be able to capture and store our data without compatibility issues. In most cases, the data is readily available (except for data stored on the Web, which needs to be crawled)

Pre-processing

Once the data has been loaded, it is vital to make it ready for suitable analyses. A few generic pre-processing steps are :

- Tokenizing text
- POS tagging
- Disambiguation
- Cleanup of non-textual data (white space, numbers, symbols)

Transformation (Feature generation/selection)

Transformation would ensure that the text corpus is enriched and homogenous. These processes include :

- Stop word removal
- Stemming
- Transform cases

A few of these processes have been briefed in the pages ahead.

Using algorithms (Text/Data Mining)

Depending on the business problem at hand, we can use several of the basic machine learning algorithms to crunch our text to discover patterns. These include :

- Word frequencies and word clouds
- K-Means cluster analysis
- Association rule generators
- K nearest neighbors
- Neural networks
- Naïve Bayes classifiers
- N-gram generators

Evaluation (Analyzing results)

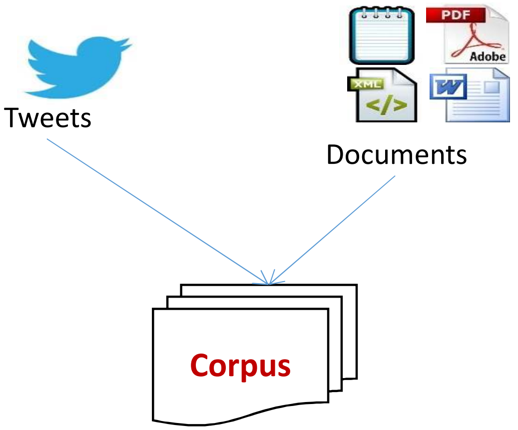
This is an inferential step. The analytics expert has the option to decide if

- another refined iteration of the algorithm is justified, or
- is the current result good enough to discover unseen patterns among the text

POS Tagging: Parts of speech Tagging

Typical Text Mining Steps

Step 1 – Data Assemble

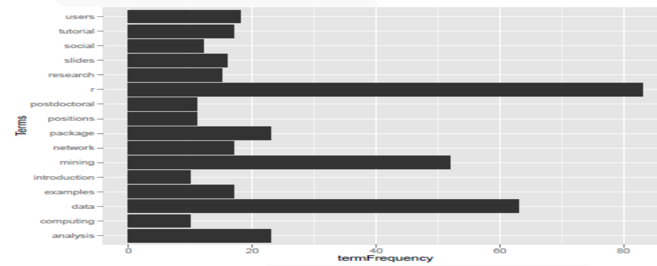


Step 2 – Data Processing

2A - Explore Corpus
2B - Convert text to lowercase
2C - Remove
a) Numbers(if required)
b) Punctuations
c) English stop words
d) Own stop words(if required)
e) Strip whitespace
f) Lemmatization/Stemming
g) Sparse terms
2D - Create document term matrix

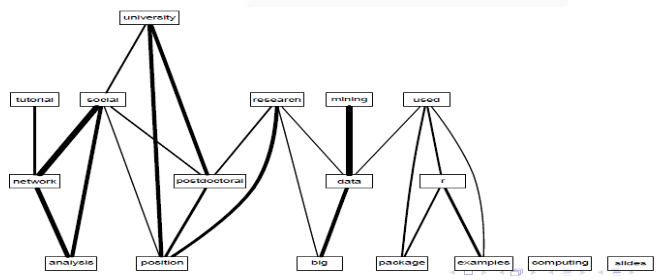
Step 3 - Visualization

Frequency Items



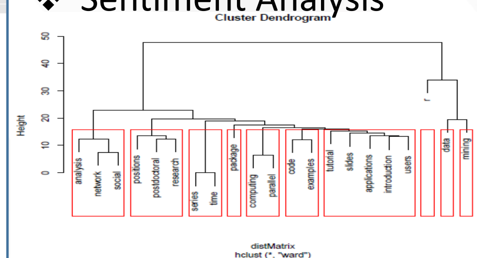
Word Cloud

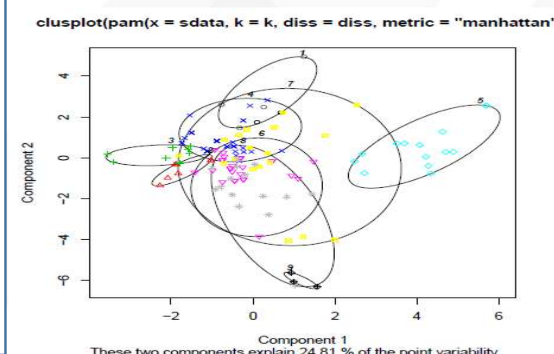
Correlation Plot



Step 4 – Run Model(s)

- Clustering
- Classification
- Sentiment Analysis

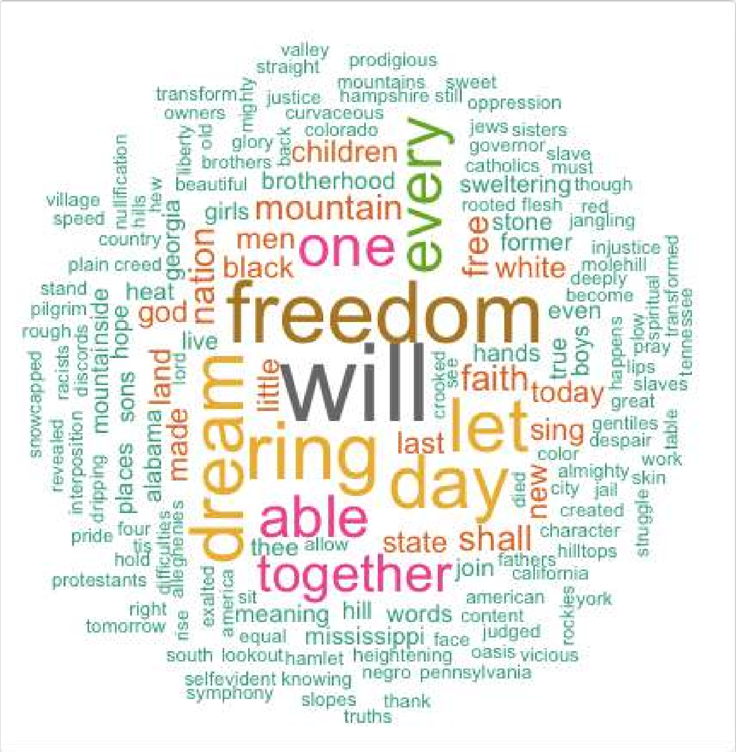




(Description about each step is on next slide)

Word clouds

- A **word cloud** is a **text mining** method that allows us to highlight the most frequently used keywords in a paragraph of texts.
- It is also referred to as a **text cloud** or **tag cloud**.
- A **text mining** package (tm) and **word cloud generator** package (**wordcloud**) are available in R for helping us to analyze texts and to quickly visualize the keywords words as a **word cloud**.



Word clouds

3 reasons you should use word clouds to present your text data

- **Tag cloud** is a powerful method for **text mining** and, it add simplicity and clarity. The most used keywords stand out better in a word cloud
- **Word clouds** are a potent communication tool. They are easy to understand, to be shared and are impactful
- **Word clouds** are visually engaging than a table data

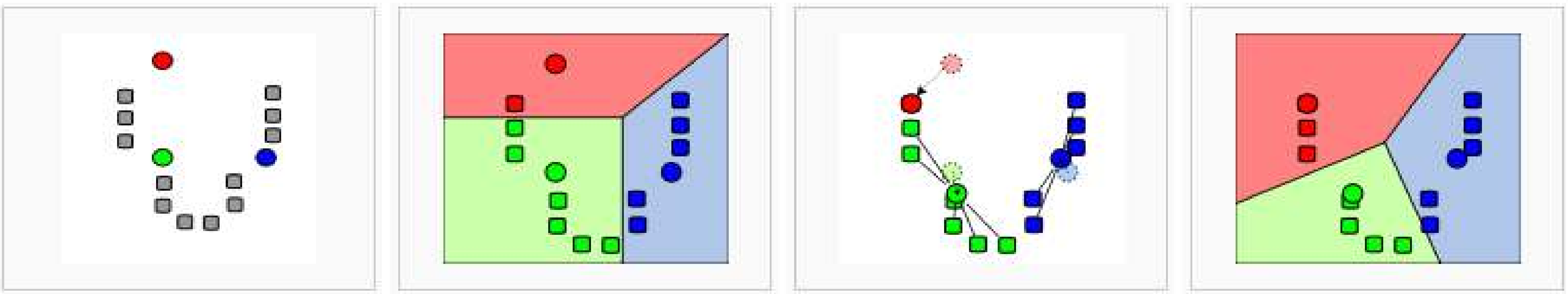
Who is using word clouds ?

- Researchers : for reporting qualitative data
- Marketers : for highlighting the needs and pain points of customers
- Educators : to support essential issues
- Politicians and journalists
- social media sites : To collect, analyze and share user sentiments

K-Means clustering

- unsupervised learning
- group **n** documents into **k** clusters

Demonstration of the standard algorithm



1) *k* initial "means" (in this case *k*=3) are randomly selected from the data set (shown in color).

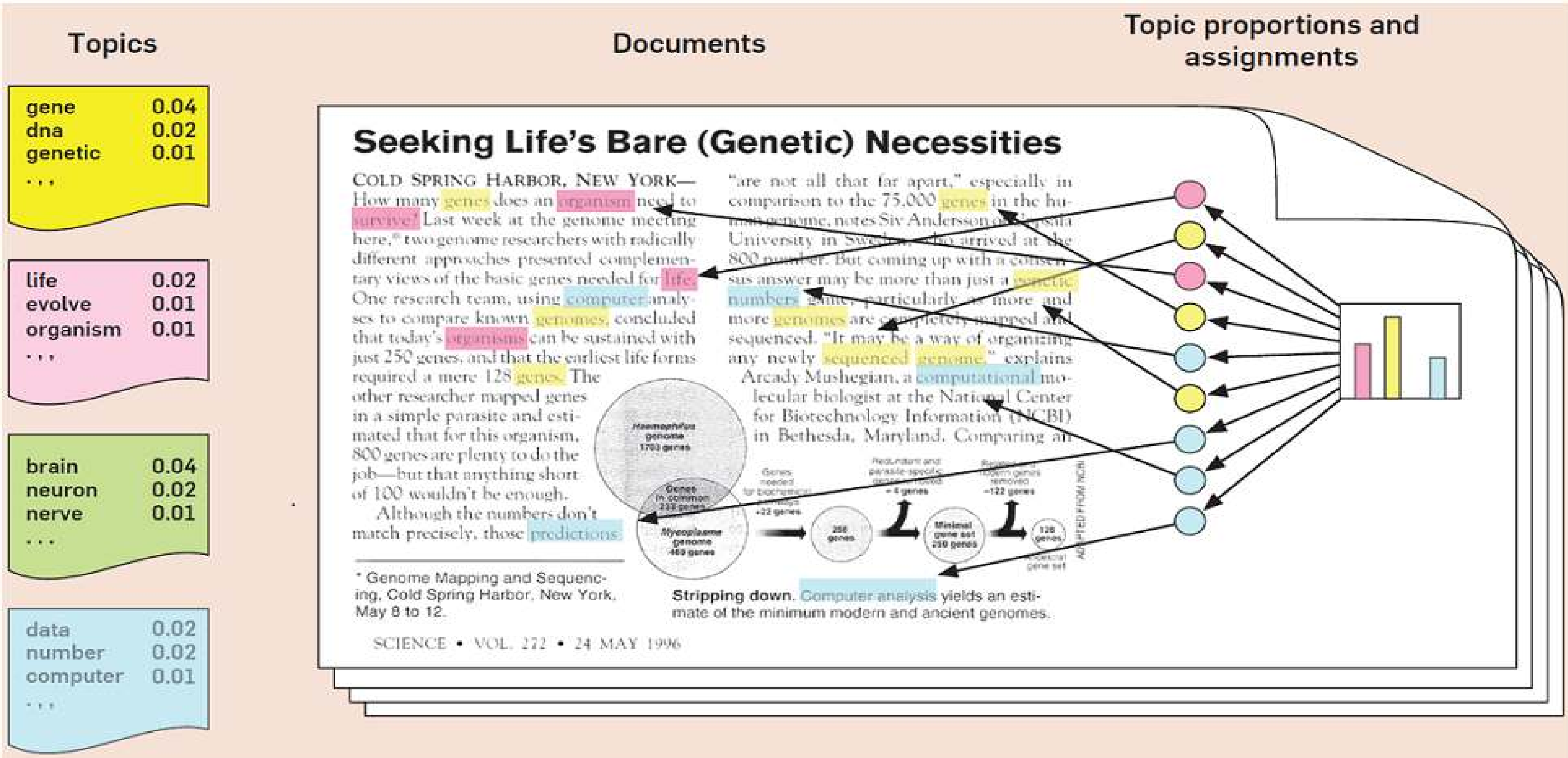
2) *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

3) The centroid of each of the *k* clusters becomes the new means.

4) Steps 2 and 3 are repeated until convergence has been reached.

ANALYTIXLABS

Topic modeling with topic models



Blei, 2012, Communications of the ACM

Brief Description about Data Processing Steps

Data Processing Step	Brief Description
Explore Corpus	Understand the types of variables, their functions, permissible values, and so on. Some formats including html and xml contain tags and other data structures that provide more metadata.
Convert text to lowercase	This is to avoid distinguish between words simply on case.
Remove Number(if required)	Numbers may or may not be relevant to our analyses.
Remove Punctuations	Punctuation can provide grammatical context which supports understanding. Often for initial analyses we ignore the punctuation. Later we will use punctuation to support the extraction of meaning.
Remove English stop words	Stop words are common words found in a language. Words like for, very, and, of, are, etc, are common stop words.
Remove Own stop words(if required)	Along with English stop words, we could instead or in addition remove our own stop words. The choice of own stop word might depend on the domain of discourse, and might not become apparent until we've done some analysis.
Strip whitespace	Eliminate extra white-spaces.
Stemming	Stemming uses an algorithm that removes common word endings for English words, such as “es”, “ed” and “'s”.
Sparse terms	We are often not interested in infrequent terms in our documents. Such “sparse" terms should be removed from the document term matrix.
Document term matrix	A document term matrix is simply a matrix with documents as the rows and terms as the columns and a count of the frequency of words as the cells of the matrix.

Calculate Term Weight – TF-IDF

How frequently term appears?

Term Frequency: TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document)

How **important a term is?**

DF: Document Frequency = d (number of documents containing a given term) / D (the size of the collection of documents)

*To normalize take log(d/D), but often D > d and log(d/D) will give negative value. So invert the ratio inside log expression. Essentially we are **compressing the scale of values so that very large or very small quantities are smoothly compared***

IDF: Inverse Document Frequency IDF(t) = log(Total number of documents / Number of documents with term t in it)

Example:

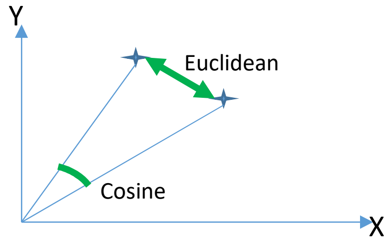
Consider a document containing 100 words wherein the word **CAR** appears 3 times

TF(CAR) = 3 / 100 = 0.03

Now, assume we have 10 million documents and the word **CAR** appears in one thousand of these

IDF(CAR) = log(10,000,000 / 1,000) = 4
TF-IDF weight is product of these quantities: 0.03 * 4 = 0.12

Similarity Distance Measure



$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

- cosine value will be a number between 0 and 1
- Smaller the angel bigger the cosine value/similarity

Example:

Text 1: statistics skills and programming skills are equally important for analytics
Text 2: statistics skills and domain knowledge are important for analytics
Text 3: I like reading books and travelling

	statistics	skills	and	programming	knowledge	are	equally	important	for	analytics	domain	I	like	reading	books	travelling
Text 1	1	2	1	1	0	1	1	1	1	1	0	0	0	0	0	0
Text 2	1	1	1	0	1	1	0	1	1	1	1	0	0	0	0	0
Text 3	0	0	1	0	0	0	0	0	0	0	0	1	1	1	1	1

The three vectors are:
T1 = (1,2,1,1,0,1,1,1,1,1,0,0,0,0,0,0)
T2 = (1,1,1,0,1,1,0,1,1,1,1,0,0,0,0,0)
T3 = (0,0,1,0,0,0,0,0,0,0,0,1,1,1,1,1)

Microsoft Excel
97-2003 Worksheet

Degree of Similarity (T1 & T2) = (T1 %*% T2) / (sqrt(sum(T1^2)) * sqrt(sum(T2^2))) = 77%
Degree of Similarity (T1 & T3) = (T1 %*% T3) / (sqrt(sum(T1^2)) * sqrt(sum(T3^2))) = 12%

Additional Reading: *Here is a detailed paper on comparing the efficiency of different distance measures for text documents.*
URL - http://www.ijert.org/view-pdf/2373/space-and-cosine-similarity-measures-for-text-document-clustering

n-gram

Example: "defense attorney for liberty and montecito"

1-gram:
defense
attorney
for
liberty
and
montecito

2-gram:
defense attorney
for liberty
and montecito
attorney for
liberty and
attorney for

3-gram:
defense attorney for
liberty and montecito
attorney for liberty
for liberty and
liberty and montecito

4-gram:
defense attorney for liberty
attorney for liberty and
for liberty and montecito

5-gram:
defense attorney for liberty and montecito
attorney for liberty and montecito

Definition:

- n-gram is a contiguous sequence of n items from a given sequence of text
- The items can be syllables, letters, words or base pairs according to the application

Application:

- Probabilistic language model for predicting the next item in a sequence in the form of a (n − 1)
- Widely used in probability, communication theory, computational linguistics, biological sequence analysis

Advantage:

- Relatively simple
- Simply increasing n, model can be used to store more context

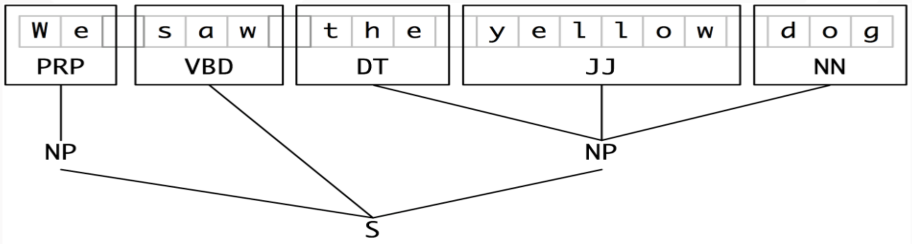
Disadvantage:

- Semantic value of the item is not considered

Shallow NLP Technique

Definition:

- Assign a syntactic label (noun, verb etc.) to a chunk
- Knowledge extraction from text through semantic/syntactic analysis approach



Application:

- Taxonomy extraction (predefined terms and entities)
    - Entities: People, organizations, locations, times, dates, prices, genes, proteins, diseases, medicines
- Concept extraction (main idea or a theme)

Advantage:

- Less noisy than n-grams

Disadvantage:

- Does not specify role of items in the main sentence

Shallow NLP Technique

Sentence - "The driver from Europe crashed the car with the white bumper"

Concept Extraction:

1-gram	Part of Speech
the	DT – Determiner
driver	NN - Noun, singular or mass
from	IN - Preposition or subordinating conjunction
europe	NNP - Proper Noun, singular
crashed	VBD - Verb, past tense
the	DT – Determiner
car	NN - Noun, singular or mass
with	IN - Preposition or subordinating conjunction
the	DT – Determiner
white	JJ – Adjective
bumper	NN - Noun, singular or mass

- Convert to lowercase & PoS tag
- Remove Stop words
- Retain only Noun's & Verb's
- Bi-gram with Noun's & Verb's retained

Bi-gram	PoS
car white	NN JJ
crashed car	VBD NN
driver europe	NN NNP
europe crashed	NNP VBD
white bumper	JJ NN

- 3-gram with Noun's & Verb's retained

3-gram	PoS
car white bumper	NN JJ NN
crashed car white	VBD NN JJ
driver europe crashed	NN NNP VBD
europe crashed car	NNP VBD NN

Conclusion:

1-gram: Reduced noise, however no clear context

Bi-gram & 3-gram: Increased context, however there is a information loss

PoS Tagging: http://nlp.stanford.edu:8080/corenlp/process

NER Demo: http://nlp.stanford.edu:8080/ner/process

Deep NLP technique

Definition:
- Extension to the shallow NLP
- Detected relationships are expressed as complex construction to retain the context
- Example relationships: Located in, employed by, part of, married to

Applications:
- Develop features and representations appropriate for complex interpretation tasks
  - Fraud detection
  - Life science: prediction activities based on complex RNA-Sequence

Example:

The above sentence can be represented using **triples (Subject: Predicate [Modifier]: Object)** without loosing the context.

Triples:
driver : crash : car
driver : crash with : bumper
driver : be from : Europe

Techniques - Summary

Technique	General Steps	Pros	Cons
N-Gram	- Convert to lowercase - Remove punctuations - Remove special characters	Simple technique	Extremely noisy
Shallow NLP technique	- **POS tagging** - **Lemmatization** i.e., transform to dictionary base form i.e., "produce" & "produced" become "produce" - **Stemming** i.e., transform to root word i.e., 1) "computer" & "computers" become "comput" 2) "product", "produce" & "produced" become "produc" - **Chunking** i.e., identify the phrasal constituents in a sentence , including noun/verb phrase etc., and splits the sentence into chunks of semantically related words	Less noisy than N-Grams	Provide a relatively less, computationally expensive solution for analyzing the structure of texts. Does not specify the internal structure or the role of words in the sentence
Deep NLP technique	- Generate syntactic relationship between each pair of words - Extract subject, predicate, nagation, objecct and named entity to form triples.	Context of the sentence is retained.	Sentence level analysis is too structured

ANALYTIXLABS

Customer Complaint Classification

Demonstration of Use case implementation –Customer Complaints Prioritization

Current Project Objective:
Using historical unstructured data of Customer complaints, classify the complaints in different categories (NAC, LIC, AER) and provide matched case resolutions through Text mining and Machine Learning Algorithms

Proposed Project Benefits:
- Reduce Manual intervention
- Minimized Time, Effort & Cost (quantification can be done once the date is available)
- Focused resolution efforts by accurate classification of customer complaints

Our understanding of requirement:

1. Huge volumes –Thousands of customer complaints received each day
2. Unstructured text format of data
3. Lot of time is being spent to identify NAC/LIC - Not A Complaint or Low Impact Complaint
4. Prioritization of customer complaints to meet SLAs is a challenge

Solution offered:

1. Use of historical data
2. Text mining functions (Corpus, Data cleaning, DTM, Cosine similarity, NLP etc.)
3. Develop machine learning algorithm on the customer complaints text data
4. Train the ML model to get the highest level of accuracy

CUSTOMER COMPLAINTS PRIORITIZATION (PROJECT EXECUTED):

✓ Use of historical data
✓ Text mining functions (Corpus, Data cleaning, DTM, Cosine similarity, NLP etc)
✓ Developed a completely different and cutting edge complaint prioritization framework
✓ Applied sophisticated machine learning techniques such as Random Forest, Support Vector Machine, Neural Networks and Generalized Linear Model

Benefits Delivered:
- **5% reduction in complaints redressal cost within 2 months**
- **66% reduction in number of customer complaints; improved customer satisfaction**

ANALYTIXLABS

Business context

- Large installed base of customer's medical equipment across several healthcare centers
- Thousands of customer complaints received each day; shortage of Field Engineers to address them on time
- Ineffective framework to prioritize customer complaints and focus Field Engineers' efforts
- Additional challenge in analysis due to unstructured text format of complaints records

What we observed

- Large volume of varied un-structured customer feedback

- Current sentiment analysis methodology ineffective in dealing with variety and scale of data

- Poor customer satisfaction
- High cost to resolve complaints

What we did

- Developed a completely different and cutting edge complaint prioritization framework

- Applied sophisticated machine learning techniques such as Random Forest, Support Vector Machine, Neural Networks and Generalized Linear Model

Data Preparation and Exploration

Load & Clean: Create corpus, stemming, transformation e.g. remove whitespaces, numbers etc.

Document Term Matrix: Sparse matrix, inspect DTM, word cloud etc.

Analyse DTM: Frequency stats, correlation b/w features, create word cloud etc.

Data Size :
~1GB unstructured data
Tools used:
R & Excel

Model Development & Performance Improvement

Model development

- Split data - training and validation set
- Apply machine learning techniques e.g. RF, SVM etc.

Model Validation:

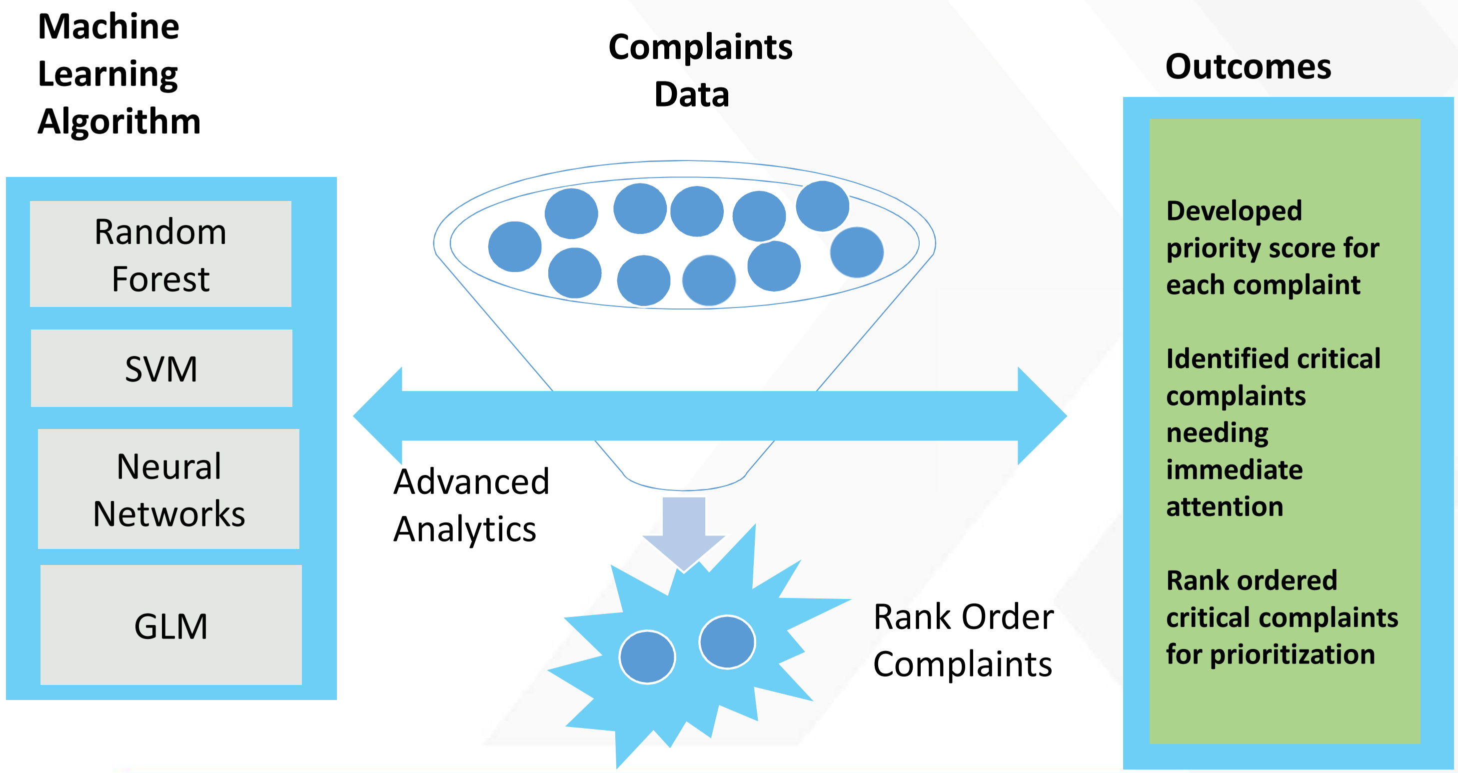
- Test sample validation or unseen data
- Cross validation (k-fold validation)
- Analyse key features in the model

Accuracy Improvement

- Optimize bias and variance (SVM, RF etc.)- bias-variance trade-off
- Model boosting techniques

The method

- Machine learning algorithm with multiple techniques deployed



Benefits Delivered

- Focused resolution efforts by prioritizing attention on the most critical complaints
- Minimized time & effort on mundane/ less critical concerns
- Thereby reduced the number of Field Engineers employed

- **5% reduction in complaints redressal cost within 2 months**
- **66% reduction in number of customer complaints; improved customer satisfaction**

We also prescribed...

- Current model developed based on complaints data for two equipment modalities only
- Replicate the model across other modalities
    - separate the signal from noise given the nature of text complaints
    - focus and gain additional benefit by customizing the model for each equipment modality

Use Case of Airlines

Objective

Text Analytics helps to identify potential business opportunities for one of leading Manufacture Client

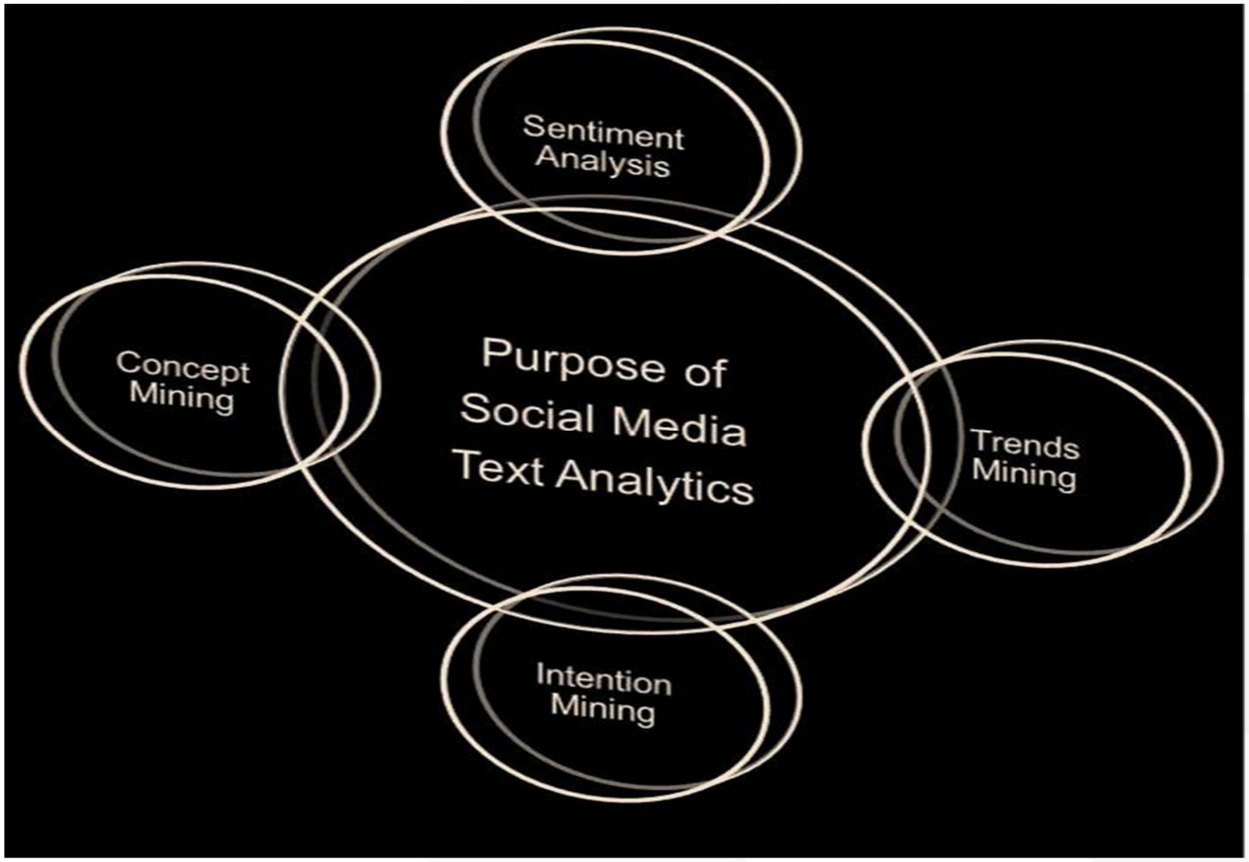
Airlines
- American
- Lufthansa

Data Sources
- Press releases
- News articles
- Facebook
- Twitter

Keywords
- Technology
- Digital
- Responsibilities
- Vision
- Innovation
- Goal

- Extracted ~ 100 news articles for American & Lufthansa airlines from various sources.
- Official websites of respective airlines has been used during data collection.
  - http://hub.aa.com/en/nr/pressrelease/fleet
  - https://www.facebook.com/AmericanAirlines
  - https://twitter.com/americanair
  - https://www.lufthansagroup.com/en/press/news-releases/press-releases.html
  - https://www.facebook.com/lufthansa/
  - http://airwaysnews.com/
  - http://aviationblog.dallasnews.com/
  - http://edition.cnn.com/

Purpose of Social Media Text Analytics



Why do we need to analyze consumer language?

To help discover the true value of information!

Consumer language = Facts +

1 Product experience

2 Emotions

3 Attitudes / Perception

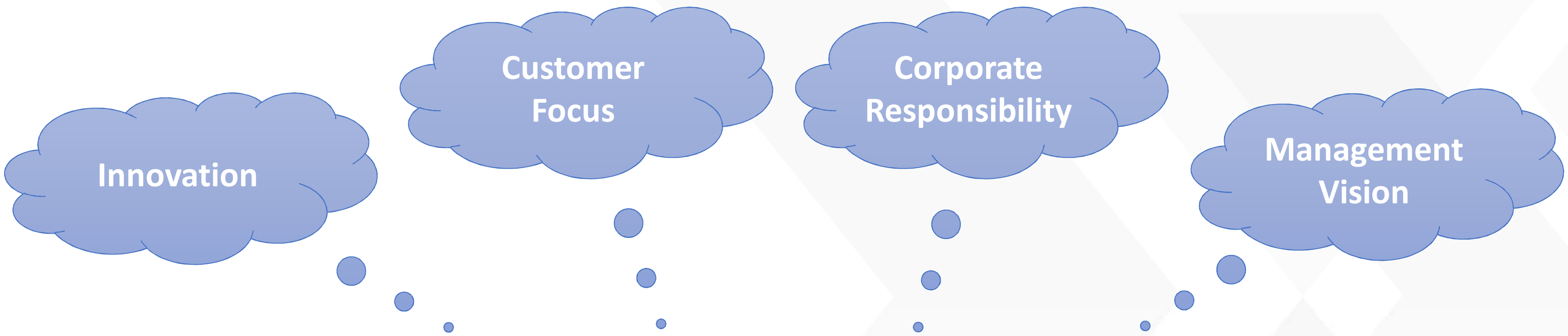
4 Brand image

5 Competition

Word Cloud

Most frequently used words were "Technology, Customers, Services, Digital, Strategic, opportunities, Security, Quality, Growth, etc.." **by American & Lufthansa** in multiple forums

Topic Modeling



Topic-1	Topic-2	Topic-3	Topic-4
technology	services	strategic	operations
security	digital	sustainability	financial
climate	customer	corporate	maintenance
safety	entertainment	opportunities	vision
research	baggage	responsibility	management
noise	economy	environment	organization

Topic - 1

American Airlines:

- Aircraft upgrades: Modern, more efficient aircraft. (600+)
- Safety and Security: new aircraft technology - **Rockwell Collins' MultiScanThreatTrack weather radar.**
- Reducing **greenhouse emissions, Flying smarter**

Lufthansa Airlines:

- "Innovation hub" established in Berlin, closer to the start-up and digital technology scene.
  - 500 million euros to be invested in innovations by 2020
  - **Fuel efficiency, Noise protection and Climate research project IAGOS, Flying Lab, Zero-G arm**
- Invests massively in ecological sustainability of flight operations:
  - Biggest fleet renewal program and invests in **highly efficient and quiet aircraft.** by 2025.
  - New engine technology, the 85 decibel noise contour of an A320neo at take-off. **Ordered a total of 116 aircraft** of this type.
  - **"vortex generators"** under the wings.

Innovation
technology
security
climate
safety
research
noise

Topic - 2

American Airlines:

- **Premium Economy**
  - Boeing 787-9, which is expected to enter service in late 2016.
  - In Airbus A350, which arrives in 2017.
  - Boeing 777-300ERs, 777-200ERs, 787-8s and Airbus A330s over the next three years.
- **T-link software** – Improve Baggage handling
- **Baggage Reroute Tool** help's to manage baggage when customers are rerouted.

Lufthansa Airlines:

- The **Innovation Hub**, to ensure identifying future customer needs and trends at an early stage and participates in shaping them.
- BoardConnect - wireless service for customers to use their own devices to access entertainment.
- **Big Data and Analytics Technology:**
  - Location-based services
  - Electronic baggage receipt(**RIMOWA** Electronic Tag & Lufthansa app)
  - **SMILE program** – Surpass My Individual Lufthansa Experience
- **Flight planning app,** enables passengers to plan travel (provide information about airport traffic and security) **Coming Soon**

Customer Focus
services
digital
customer
entertainment
baggage
economy

Topic - 3

American Airlines:

- Environmental Performance:
  - Join the EPA's Climate Leaders program, committed to a **30% reduction in greenhouse gas** intensity ration by 2025 and will work with Climate Leaders to set a mid-range goal to help meet this long-range target.
  - Employee-led **Fuel Smart fuel** conservation program
- Corporate Citizenship
  - Donations towards Education, Kids, Partnership & programs
  - The Police Athletic League of Philadelphia (PAL) announced today a $180,000 gift from American Airlines

Lufthansa Airlines:

- Comprehensive sustainability agenda with following files of entrepreneurial responsibility:
  - Economic sustainability
  - Corporate Governance and compliance
  - **Climate and environmental responsibility**
  - Social responsibility
  - Product responsibility
  - Corporate citizenship

Corporate Responsibility
strategic
sustainability
corporate
opportunities
responsibility
environment

Topic - 4

American Airlines:

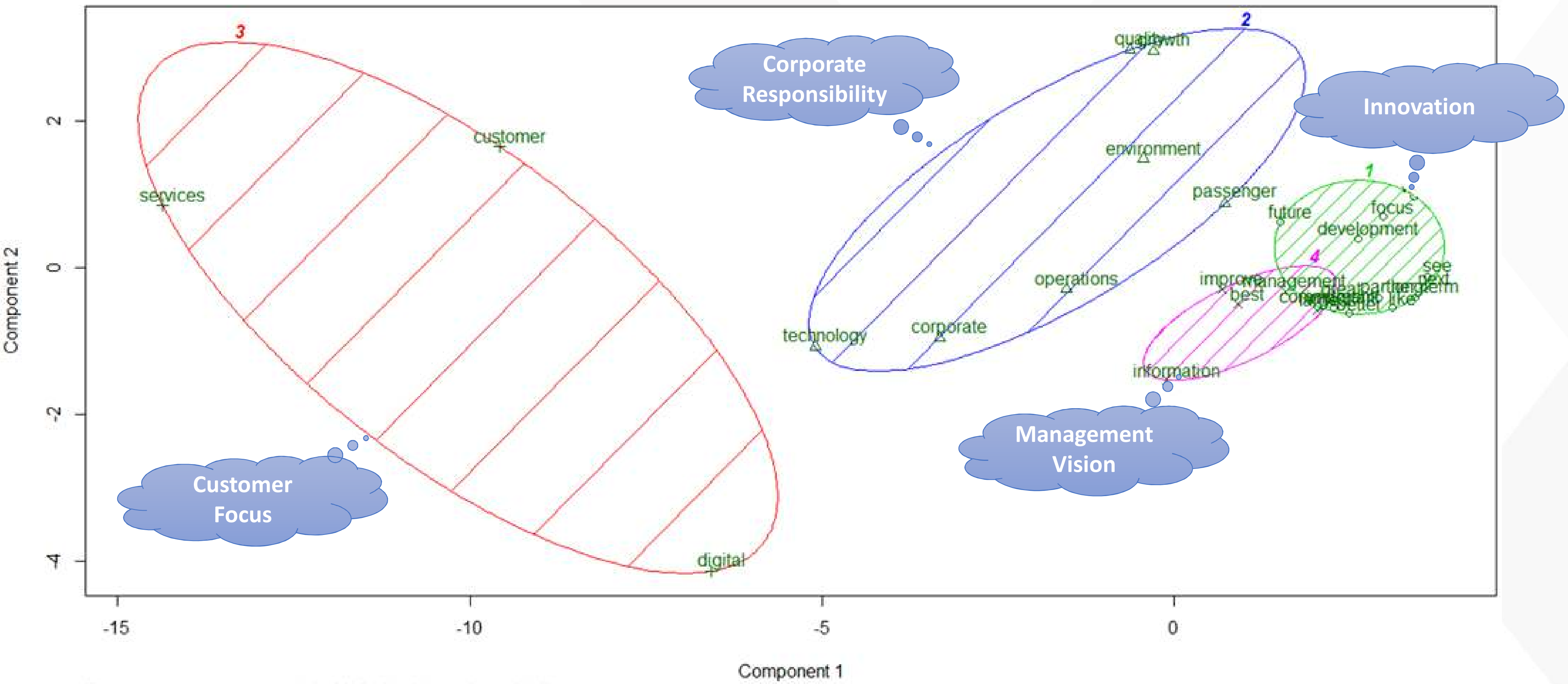
- Premium Economy - New Planes and Retrofit Plans
- **Adding up to 2 new aircraft to the fleet each week**. The new build planes are replacing older retiring aircraft.
- AA Tulsa (Maintenance & Engineering Base)
  - American's entire program of **fuel efficiency** and range increasing winglets conversion to its 737, 757, and 767 fleet all occurred in house.
  - They accomplished this with a **45% reduction** in the OSHA injury rate from 2009.
- CEO's Vision of Labor Peace.

Lufthansa Airlines:

- **" 7 to 1 – Our Way Forward"** strategic agenda was initiated in 2014
- **Airline Business technology award** - 2013 in MRO (maintenance, repair and overhaul) services.
  - The **"Taxibot"** - pilot-controlled tow-tractor
  - Improve **fuel efficiency,** testing new aircraft **paint** with a sharkskin-inspired riblet texture

Management Vision
operations
financial
maintenance
vision
management
organization

Clustering



Corporate Responsibility
Innovation
Customer Focus
Management Vision
customer
services
digital
quality
growth
environment
passenger
operations
corporate
technology
future
focus
development
see
next
improve
management
best
like
information
Component 1
Component 2
-15
-10
-5
0
2
0
-2
-4
3
2
1
4

Next Steps

Business Insights

- Recommendations
- Insights and Patterns

Text Mining

- Topic Modeling
- Clustering
- Machine Learning algorithms

Data collection

- Social media
- Blogs / Forums
- Interviews
- Annual report

Sentiment Analysis

Sentiment Analysis

✓Sentiment analysis refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials.

✓Generally speaking, sentiment analysis allows companies the ability to measure how positive or negative a person feels about their product and service.

✓Companies look at:

Reviews/Surveys
Complaints/Fees/Prices
Password recovery
Technical issues

Sentiment Analysis

Methods:

✓Scaling Systems (-10/+10)
✓Subjectivity/Objectivity Identification
✓Feature/Aspect Based Analysis

Risk:

✓Losing shifting and subjective human dynamics
✓Computers can not tell the context of a statement.
✓*Ex: sarcasm, slang, double négatives*

Example: Survey sentiment analysis

Example: Survey sentiment analysis and clustering

Objective	A major US retail bank conducted a diagnostic around the workplace technologies that they are using through a collection of surveys from an internal employee satisfaction survey. The task was to find out the themes on workplace technologies (e.g. lotus notes, video conferencing, Wi-Fi, OS etc.) and the effect of the technology on productivity.
Analysis Input	An excel file showing the results of the survey along with the comments
Analysis Process	▪ The input of the text file was given directly to the tropes software. Tropes has the option of customizing and adding our own scenarios, but for this particular case, it is not required. ▪ The process included - sentence and proposition Hashing, ambiguity solving (with respect to the words of the text), detection of episodes, detection of the most characteristic parts of text, layout and display of the result. ▪ During the process, the software will: – assign all the significant words to the above categories – analyze their distribution into subcategories (Word categories, Equivalent classes, see below) – examine their occurrence order, both within the propositions (Relations, Actant and Acted) throughout the text
Analysis Output	Through the use of word counts for various relations and scenarios, a table was compiled showing the themes in the technology environment

Text Analytics Information Retrieval – Search Engine

Search Engine

- Web based search engines are the drivers of the 21st century information infrastructure. The tech giants of our life time 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 incorporate into the results, etc…

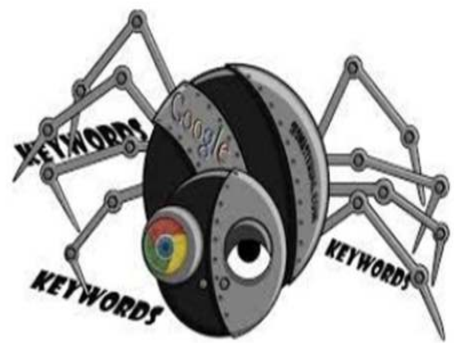
Search Engine

- 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.

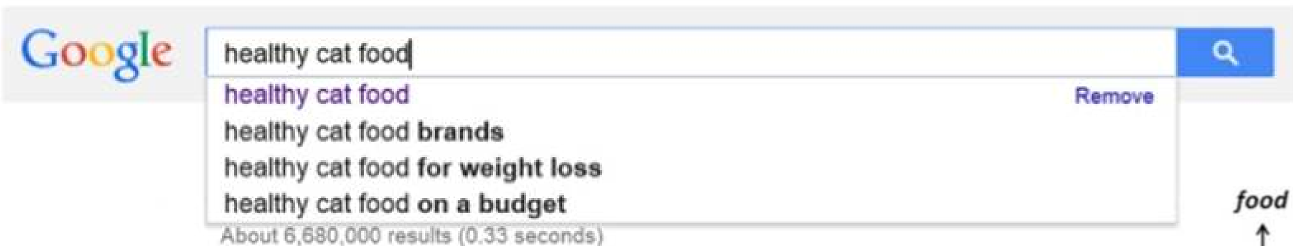
- For our example, we're going to use a tried and true method for our search algorithm, which origins are from the 1960's. We're going to implement the vector space model of information retrieval in R.



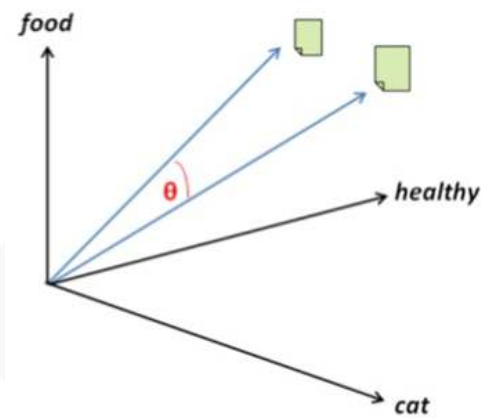
Search Query

- 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 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:

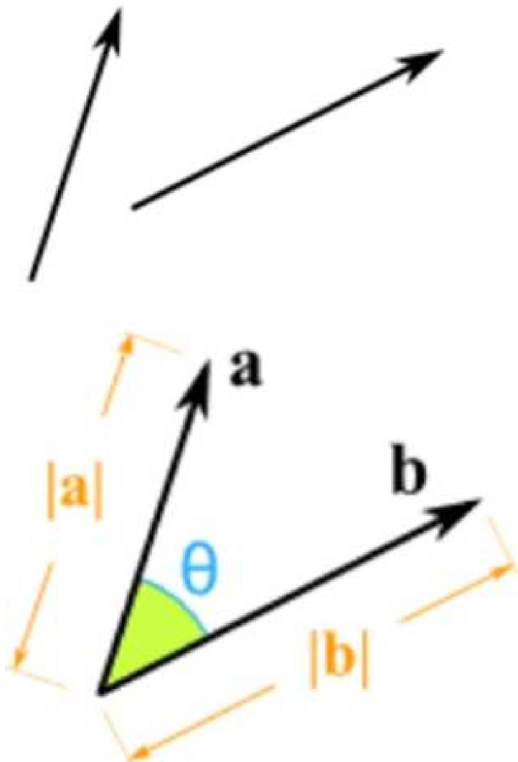
  If tf = 0, then 0, otherwise (1 + Log2(tf)) * Log2(N / df))

- 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:

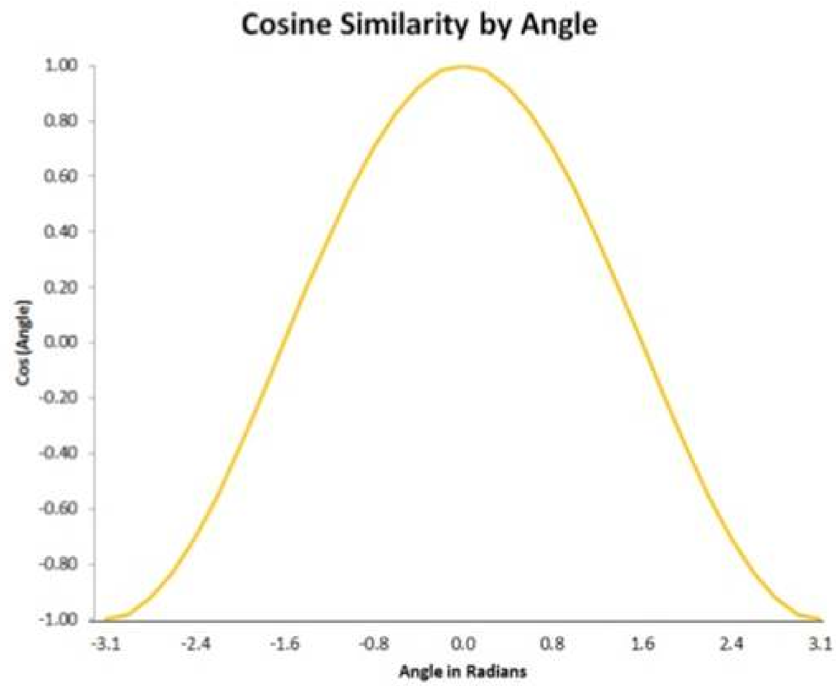
$$a \cdot b = ||a||\ ||b||cos\Theta$$

- where · 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
```

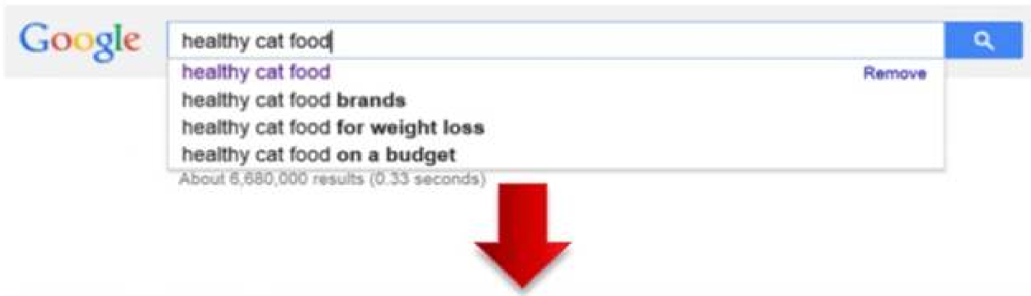
Matrix Multiplication

- 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.

Contact Us

Visit us on: http://www.analytixlabs.in/

For more information, please contact us: http://www.analytixlabs.co.in/contact-us/

Or email: info@analytixlabs.co.in

Call us we would love to speak with you: +91 95-55-219007

Join us on:

Twitter - http://twitter.com/#!/AnalytixLabs

Facebook - http://www.facebook.com/analytixlabs

LinkedIn - http://www.linkedin.com/in/analytixlabs

Blog - http://www.analytixlabs.co.in/category/blog/