

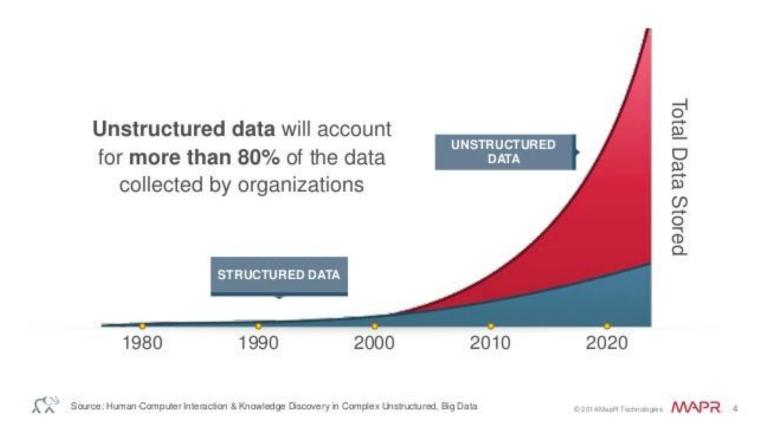
Unstructured Data AnalysisIntroduction

By

Kathirmani Sukumar



Growth of Unstructured Data



Companies started investing in technologies to work with unstructured data like text, images, videos, audios etc.



Types of Data

Data

Structured

- Mostly tabular data
- JSON, web logs

Unstructured

- Text data
- Images,
- Video and audio files

Semi Structured

 Tabular data along with large text column



Unstructured Data Analysis Challenges

By

Kathirmani Sukumar



Sources of Text Data

- Social media platforms
 - ▶ Twitter, Facebook, Whatsapp, etc
- Documents
 - Memos, research papers, articles etc
- ▶ Email conversations
- Web pages
- **SMS**
- ▶ Transcripts





Humans vs Machines

Humans	Machines
Can understand numbers, text, images	Can understand only numbers
Can easily spot spelling mistakes	Require algorithms to detect typos.
Can understand sarcasm	Cannot understand them
Can summarize large document	Complex algorithms are required to get basic summary of text
Polysemy: Can differentiate same word used in different context. Ex: bank (banking, river bank)	Only few algorithms can differentiate different meaning of same word
Once trained, manual effort to repeat a process. Ex: Classify 1000 tweets individually as positive, negative & neutral	Once trained, can easily repeat the process within short span (Can classify even 1L tweets within seconds)



Poor data quality

- ▶ Most text data are generated by humans Prone to data entry
- ▶ Text data contains abbreviations (MoM, IMHO), short forms, poor grammar
- ▶ Mix of language Regional language typed with alphabets
- ▶ Casual language Fillers, Idioms etc
- ▶ Encoding issues when applications load text files



Feed me just numbers...

- ▶ Machines can understand only numbers
- ▶ Text data -> Structured data
- Possibility of missing the context during this process





High Dimension Data

- Conversion to structured data creates high dimension data
- ▶ Requires high end machines to process medium sized data
- ▶ Takes lot of time to train a model
- Performance of models comes down
- ▶ Redundant data People write elaborate sentences



Unstructured Data Analysis Applications

By

Kathirmani Sukumar



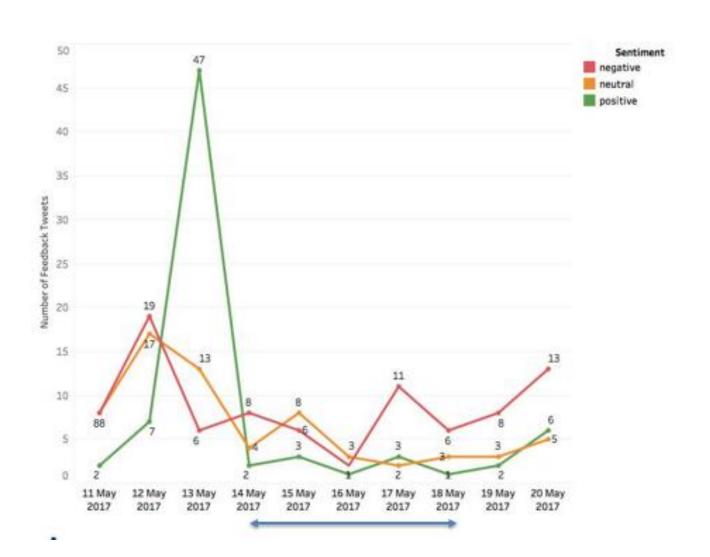
Text mining techniques

- Sentiment Analysis
 - Customer feedback & emotion
- Categorization
 - Redirecting complaints to concerned department
 - Organize documents
- Feature extraction
 - Extract phone, email ids, address etc
 - ▶ Identify themes, Identify relationships
- ▶ Text Summarization
 - One line news description
- Natural Language processing
 - ▶ Chat bots, personal assistant etc





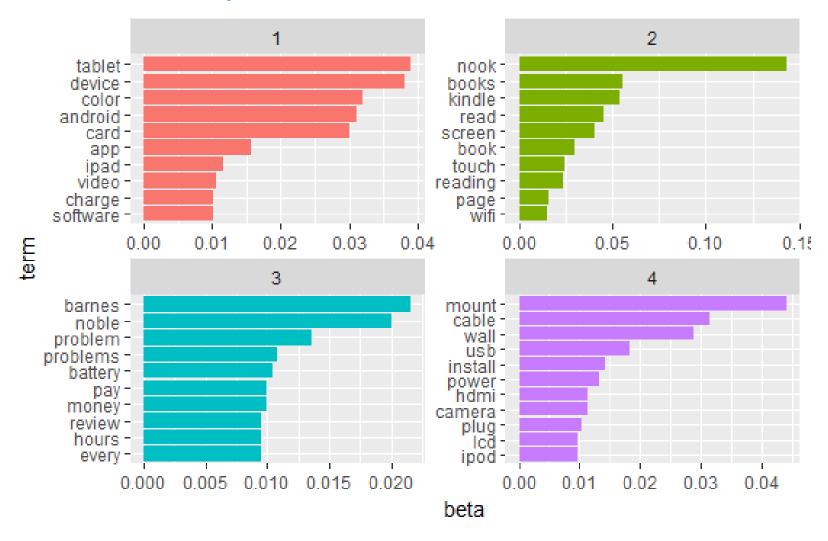
High hopes before big sales day







Extract Topics



- The first group contains words related to tablet (ipad, android, app, device etc). Hence we can label it as *Tablets*
- The second group contains words related to e-reader (nook, kindly, books, touch, page etc). Hence we can label it as *E-Reader*
- The third group contains words related to battery problems. We can name it as "Batteries"
- The fourth group contain words related to wall mounting. We can name it as "Wall mounting"



Transcripts Analysis

Company: Goldman Sachs

Asia business mix trading business CDOs compensation expense mortgage market mortgage business hedge fund tax rate client franchise IP China New York Euro investment banking Global customer activitu asset management credit risk Europe GAAP United States IPO Credit cash flow

operating environment Security capital market investment banking investment bank Basel VAR Service Asia risk management hedge fund US Europe market share Basel II

challenging environment asset management Credit IP tax rate client franchise money market

liquidity position United States Asset Management

management business market environment operating environment

tax rate risk management capital market TP Asia asset management ROE Basel II Euro trading business hedge fund Basel IPO investment bank Credit Service

balance sheet

investment banking market share Securities

customer activity

business segment operating environment

investment banking equity investment investment bank Asia Euro Lending client activity

New York FICC United States balance sheet Basel II European

China Europe asset management

market share Security money market client franchise

Insight: Discussion over BASEL increased post 2008 Crisis (Mouse-over)

risk management investment bank operating environment European prior year market share balance sheet equity investment Lendina Lending Securities tax rate risk profile FICC tax rate ROE Credit Securities Basel II _{IPO} Europe Basel Stock price Global China Basel II Service ROE European
Service client activity GAAP client activity Asia client franchise client franchise operating environment share repurchase

asset management

United States macro environment

market risk asset management investment bank balance sheet

client activity Basel II₁₁

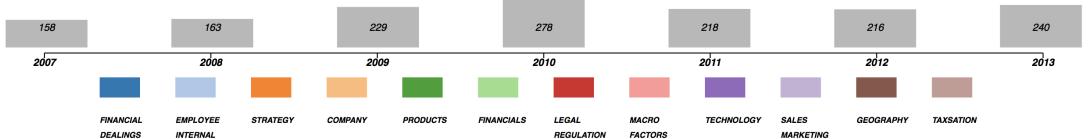
European Basel Europe

PO ROE Euro Global Asia tax rate investment banking market share Japan

investment management compensation expense

United States

QUESTIONS ASKED



Source: Gramener.com





Document Similarity

- ▶ Fake news detection
- ▶ Identify plagiarism
- ▶ Recommend books
- Organize articles and books

