Text as Data

Business Analytics

Objectives

- Background on quantitative text analysis
- Working with text
    - Cleaning & preprocessing
    - Analysis
        - Frequency & variance
        - ngrams
        - Sentiment
        - Topic modeling

Why is text important

Text is everywhere!: communication between people, not computers, so they're still "coded" as text.

Why is text important

Text is everywhere!: communication between people, not computers, so they're still "coded" as text.

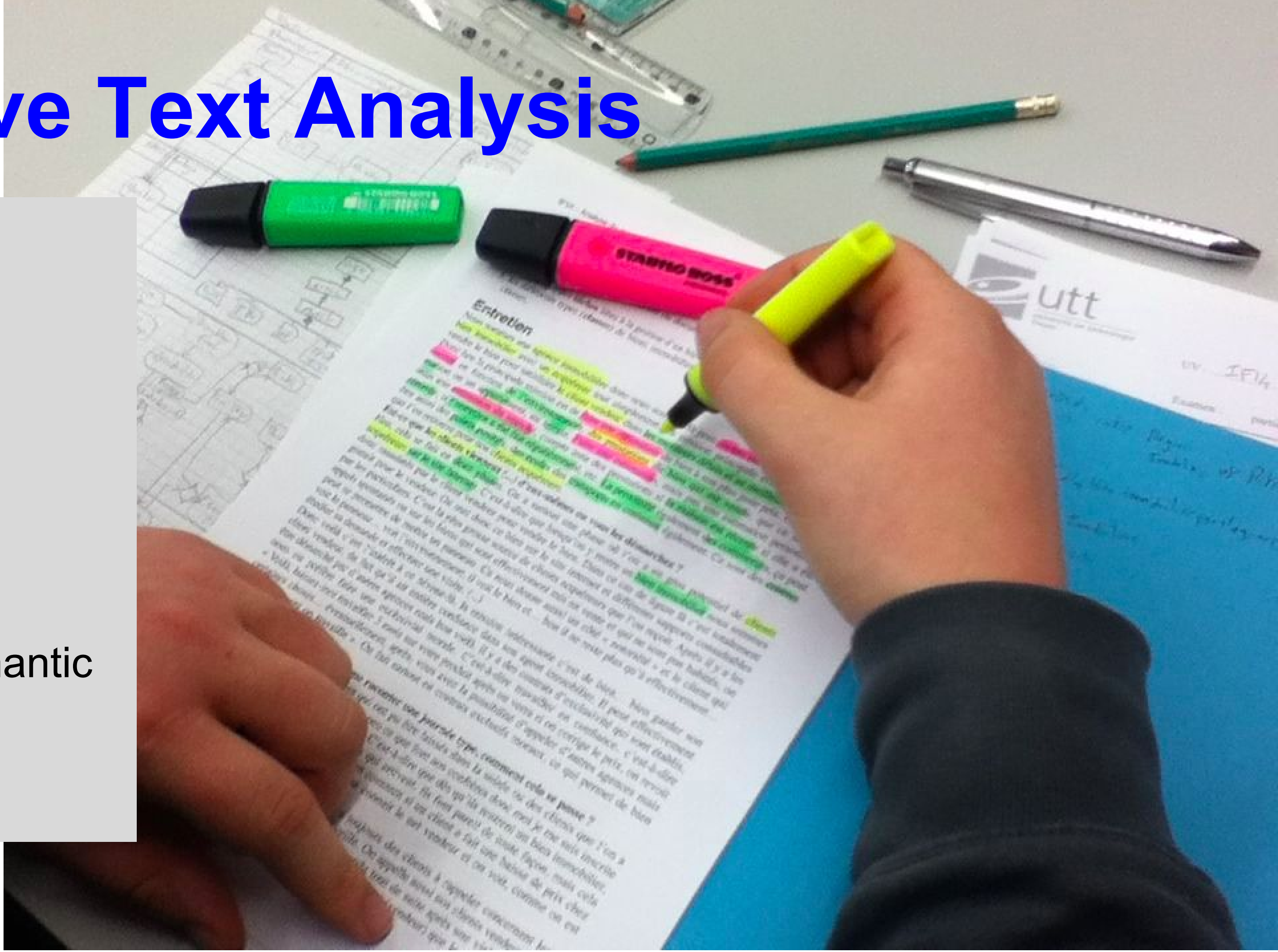
Just think of the following:

- Emails
- Blogs & posts
- Medical records
- Consumer complaint logs
- Product inquiries
- Repair records
- World of user-generated content!!!

Text as Data is Difficult!

- As data, text is relatively dirty.
- Text is often referred to as “unstructured” data.
- Text does not have the sort of structure that we normally expect for data (i.e. tables)
- Text has linguistic structure— intended for human consumption, not for computers.
    - Words have varying lengths
    - Text fields can have varying numbers of words.
    - Sometimes word order matters, sometimes not.
    - People write ungrammatically
    - Misspell words
    - Sometime words run together
    - Etc, etc, etc

*<u>Qual</u>*itative Text Analysis



- Content analysis approach
- Preserves the strengths of the communication, context, and semantic meaning

Quantitative Text Analysis

Reduces text units as data and analyze them using statistical methods ("text as data"). Known as:

- Text mining
- Statistical text processing
- Natural language processing

Implication of Quantitative Text Mining:

- Involves large-scale analysis of many texts, rather than close readings of few texts
- Requires no interpretation of texts in a non-positivist fashion
- Does not explicitly concern itself with the social or cultural predispositions of the analysts

Its is simplest form

Documents

Far far away, behind the word mountains, far from the countries Vokalia and Consonantia, there live the blind texts. Separated they live in Bookmarksgrove right at the coast of the Semantics ...

The Big Oxmox advised her not to do so, because there were thousands of bad Commas, wild Question Marks and devious Semikoli, but the Little Blind Text didn't listen. She packed her seven versalia

Document-Term Frequency Matrix

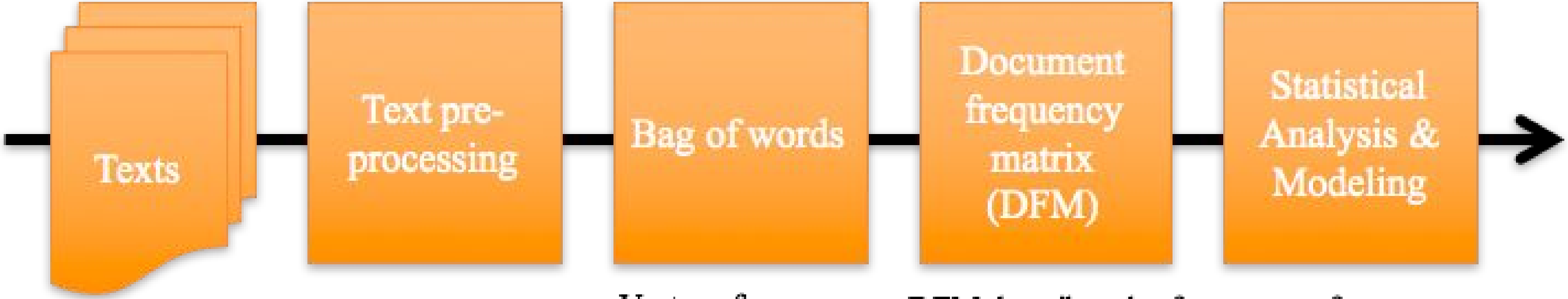
	Far	away	behind	the	word
Document 1	2	1	1	5	3
Document 2	0	0	1	6	1
Document n	1	1	0	4	2

Descriptive Statistics, extraction of topics, sentiment analysis, classification of document, vocabulary analysis

Text Analysis Framework

Probabilistic & statistical exploration
Account for:
Model validation
Semantic reconstruction*

Stop words: syntax vs semantic "words"
Stemming: words in its root form
Collocations: bigrams or trigrams



Sample vs. Population:
Observed text is a stochastic realization

Vector of words/terms for analysis (tokens)

DFM describes the frequency of terms that occur in a collection of documents.
Rows: index for documents
Columns index for terms

R Packages for Text Mining

quanteda: Quantitative Analysis of Textual Data

- A fast, flexible toolset for for the management, processing, and quantitative analysis of textual data in R.
- https://cran.r-project.org/web/packages/quanteda/index.html

stm: Estimation of the Structural Topic Model

- The Structural Topic Model (STM) allows researchers to estimate topic models with document-level covariates. The package also includes tools for model selection, visualization, and estimation of topic-covariate regressions.
- https://cran.r-project.org/web/packages/stm/index.html

R Packages for Text Mining

tm: Text Mining Package

- A framework for text mining applications within R
- https://cran.r-project.org/web/packages/tm/index.html

NLP: Apache OpenNLP Tools Interface

- An interface to the Apache OpenNLP tools (version 1.5.3). The Apache OpenNLP library is a machine learning based toolkit for the processing of natural language text written in Java. It supports the most common NLP tasks, such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and coreference resolution
- https://cran.r-project.org/web/packages/openNLP/index.html

Data & Text Pre-Processing

- **Corpus** - (plural corpora) is a large and structured set of texts used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules within a specific language territory.
- **Tokenization** - Process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens. The list of tokens becomes input for further processing such as parsing or text mining.
- **Stopwords** - words which are filtered out before or after processing of natural language data (text).
- **Stemming** - process for reducing inflected (or sometimes derived) words to their word stem, base or root form

Preprocessing

PR Sample (original)

At Internet Week New York, Yahoo! (NASDAQ:YHOO), the premier digital media company, today announces Genome from Yahoo! (www.genomeplatform.com), an online advertising solution that combines Yahoo! data with interclick's third

provide marketers w

available in July 201

display ad agreeme

in December 2011,

personalization capa

PR Sample (stopwords)

internet week new york yahoo yhoo premier digital media company today announces genome yahoo wwwgenomeplatformcom online advertising solution combines yahoo data interclicks third party data

advertise

industry

microsof

solution

PR Sample (stopwords + stemming)

internet week

wwwgenome

data along pr

genom culmi

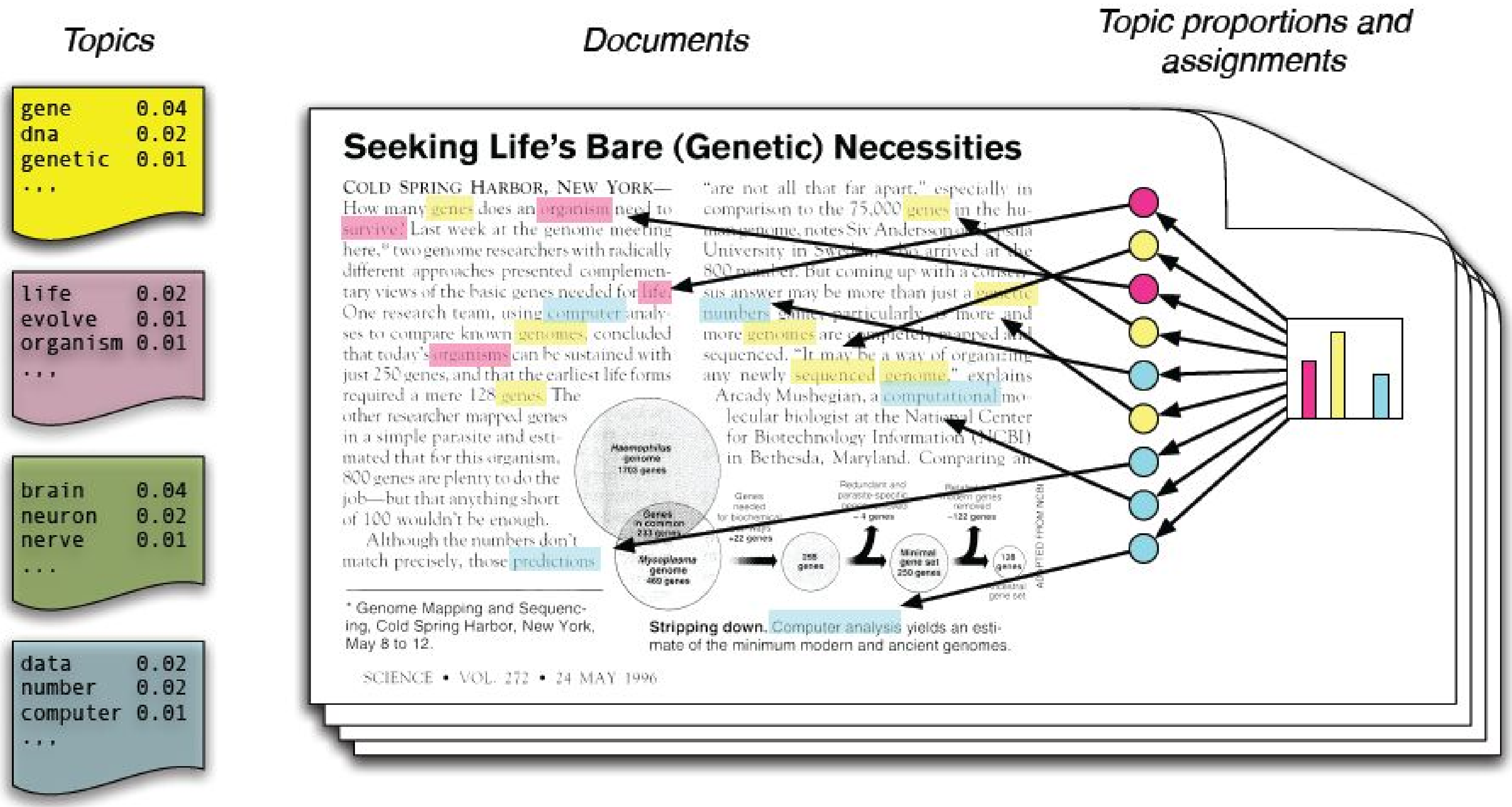
acquir decem

	Far	away	behind	the	word
Document 1	2	1	1	5	3
Document 2	0	0	1	6	1
Document n	1	1	0	4	2

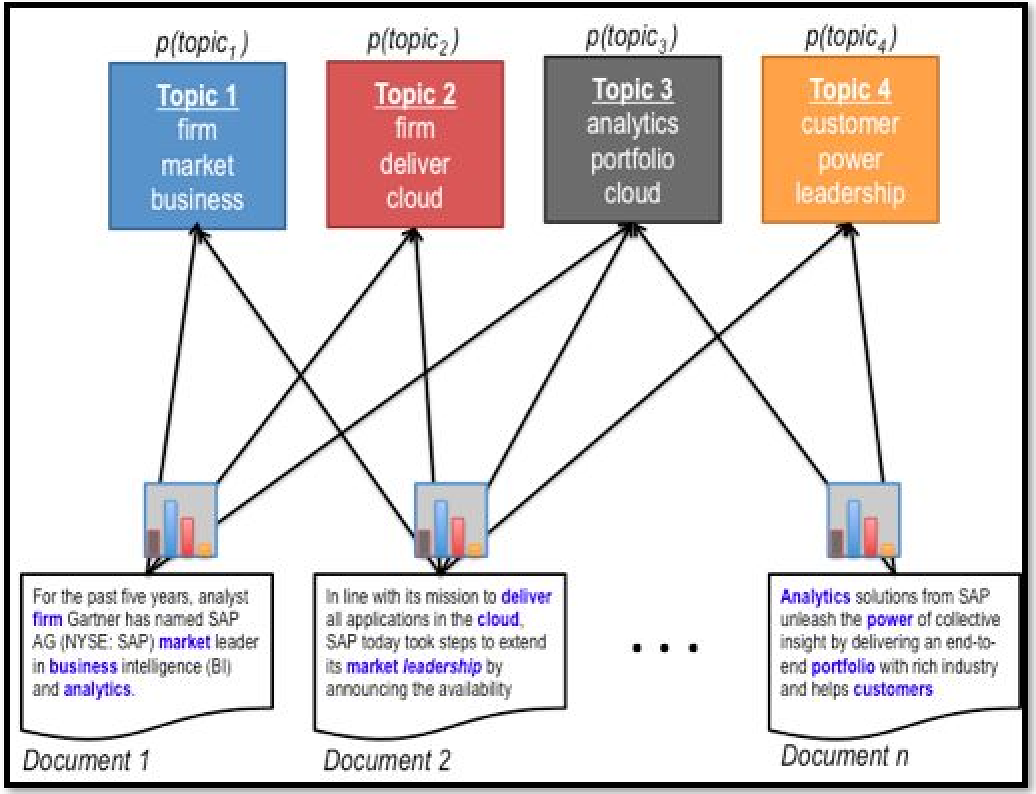
Analysis

- **ngrams** - set of co-occurring words within a given window
- **Frequency & variance analysis**
- **Sentiment analysis**
- **Topic Modeling** - techniques learn underlying features of text without explicitly imposing categories of interest. TM use assumptions and properties of the texts to estimate a set of topics and simultaneously assign documents (or parts of documents) to those topics

Topic Modeling



Probabilistic Topic Modeling



Probabilistic Topic Modeling

Common models include:

- Probabilistic Latent Semantic Indexing (pLSI)
- Correlated Topic Model (CTM)
- Latent Dirichlet Allocation (LDA)*

LDA (One of simplest topic model techniques

- A generative model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar
- The topic distribution is assumed to have a Dirichlet prior

DOI:10.1145/2133806.2133826

Surveying a suite of algorithms that offer a solution to managing large document archives.

BY DAVID M. BLEI

Probabilistic Topic Models

AS OUR COLLECTIVE knowledge continues to be digitized and stored—in the form of news, blogs, Web pages, scientific articles, books, images, sound, video, and social networks—it becomes more difficult to find and discover what we are looking for. We need new computational tools to help organize, search, and understand these vast amounts of information.

Right now, we work with online information using two main tools—search and links. We type keywords into a search engine and find a set of documents related to them. We look at the documents in that set, possibly navigating to other linked documents. This is a powerful way of interacting with our online archive, but something is missing.

Imagine searching and exploring documents based on the themes that run through them. We might "zoom in" and "zoom out" to find specific or broader themes; we might look at how those themes changed through time or how they are connected to each other. Rather than finding documents through keyword search alone, we might first find the theme that we are interested in, and then examine the documents related to that theme.

For example, consider using themes to explore the complete history of the New York Times. At a broad level, some of the themes might correspond to the sections of the newspaper—foreign policy, national affairs, sports. We could zoom in on a theme of interest, such as foreign policy, to reveal various aspects of it—Chinese foreign policy, the conflict in the Middle East, the U.S.'s relationship with Russia. We could then navigate through time to reveal how these specific themes have changed, tracking, for example, the changes in the conflict in the Middle East over the last 50 years. And, in all of this exploration, we would be pointed to the original articles relevant to the themes. The thematic structure would be a new kind of window through which to explore and digest the collection.

But we do not interact with electronic archives in this way. While more and more texts are available online, we simply do not have the human power to read and study them to provide the kind of browsing experience described above. To this end, machine learning researchers have developed *probabilistic topic modeling*, a suite of algorithms that aim to discover and annotate large archives of documents with thematic information. Topic modeling algorithms are statistical methods that analyze the words of the original texts to discover the themes that run through them, how those themes are connected to each other, and how they change over

» key insights

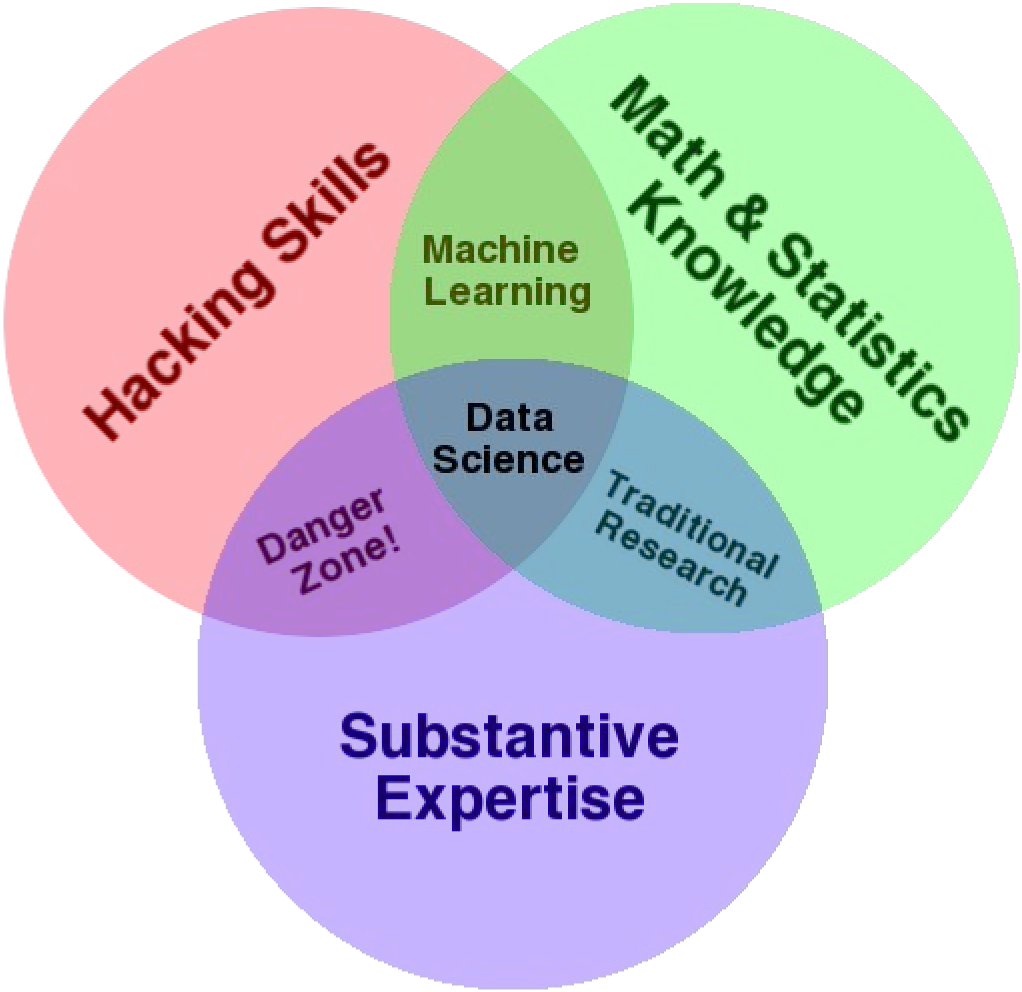
- Topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents. Topic models can organize the collection according to the discovered themes.
- Topic modeling algorithms can be applied to massive collections of documents. Recent advances in this field allow us to analyze streaming collections, like you might find from a Web API.
- Topic modeling algorithms can be adapted to many kinds of data. Among other applications, they have been used to find patterns in genetic data, images, and social networks.

Adv Text Mining

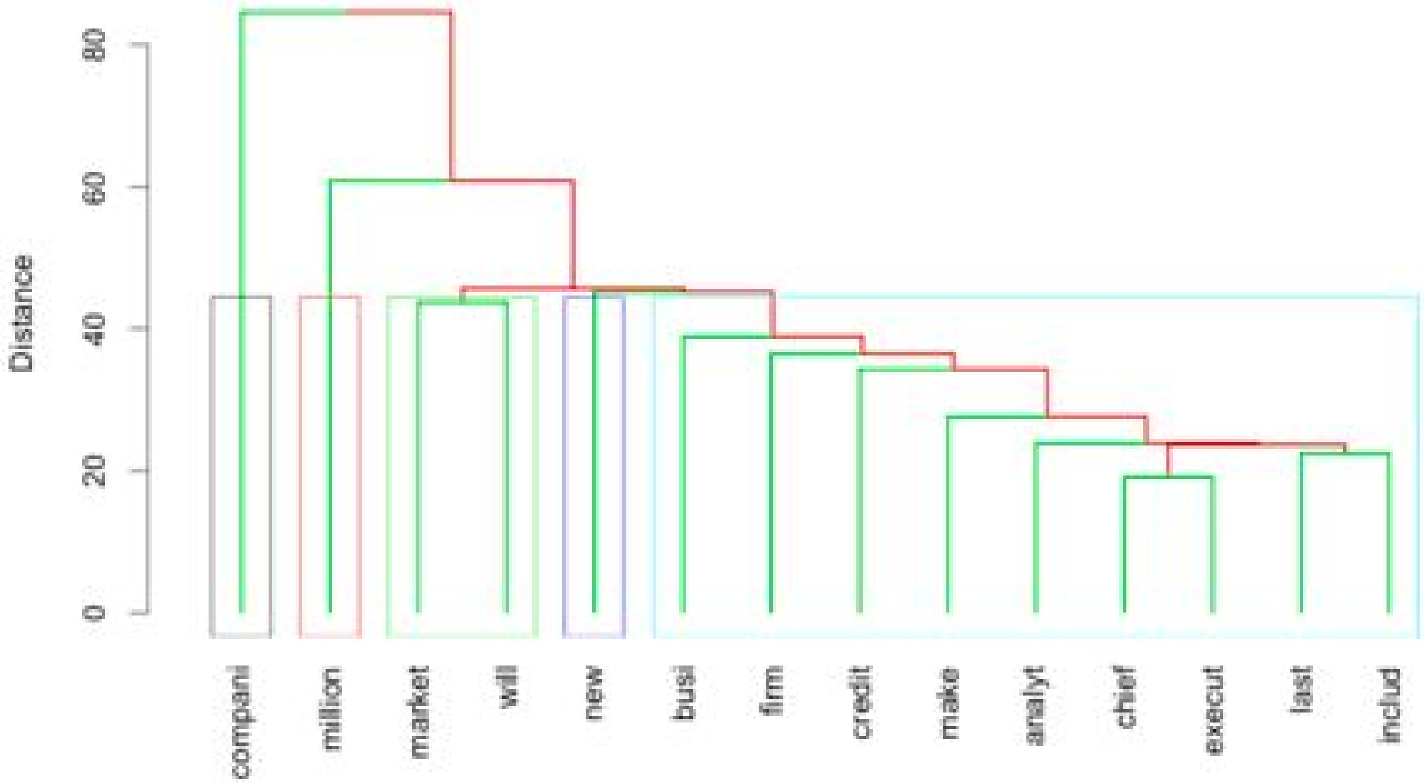
Packages to install

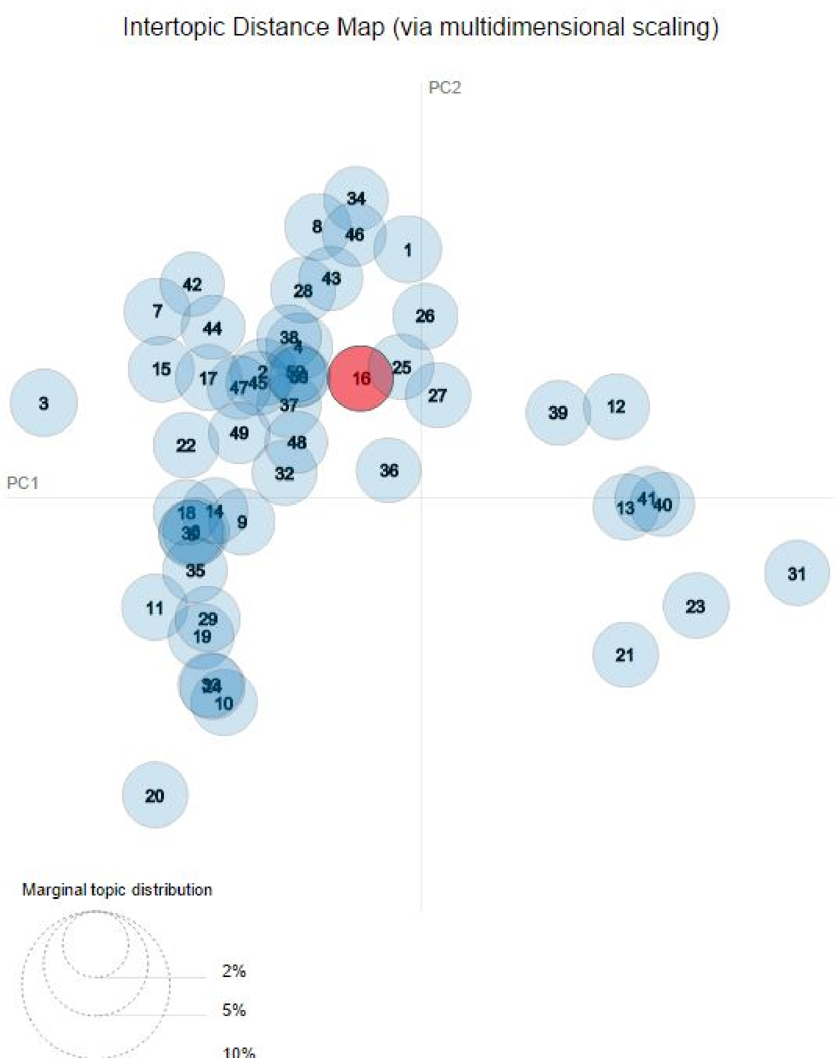
- library(quanteda)
- library(stm)
- library(tm)
- library(NLP)
- library(openNLP)
- library(ggplot2)
- library(ggdendro)
- library(cluster)
- library(fpc)
- library(dplyr)
- require(magrittr)

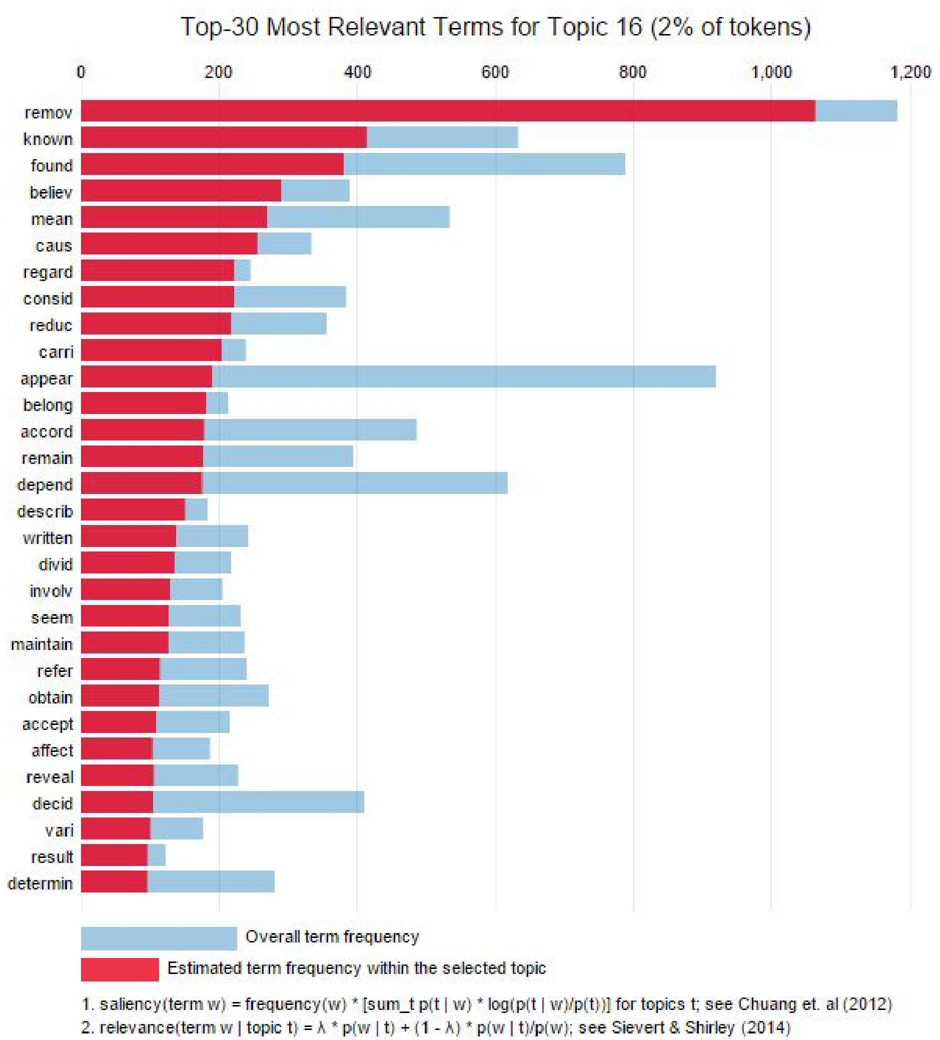
- library(stringr)
- library(lda)
- library(LDAvis)
- library(servr)


Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise

2012 Five Cluster Dendrogram




Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
34
8
46
1
42
43
28
7
26
44
38
4
15
17
2
47
45
50
16
25
27
3
37
39
12
22
49
48
32
36
13
41
40
18
14
9
35
31
11
23
29
19
21
10
20
Marginal topic distribution
2%
5%
10%


Top-30 Most Relevant Terms for Topic 16 (2% of tokens)
0
200
400
600
800
1,000
1,200
remov
known
found
believ
mean
caus
regard
consid
reduc
carri
appear
belong
accord
remain
depend
describ
written
divid
involv
seem
maintain
refer
obtain
accept
affect
reveal
decid
vari
result
determin
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)