Web Mining

CSE 634 Data Mining

Professor Anita Wasilewska

References

- Web Mining :Accomplishments & Future Directions by Jaideep Srivastava
- https://en.wikipedia.org/wiki/Web_mining#Web_content_mining
- http://www3.cs.stonybrook.edu/~cse634/L8ch5assoc.pdf
- https://www.cs.umd.edu/~samir/498/Amazon-Recommendations.pdf
- https://www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html
- https://www.xenonstack.com/blog/data-science/overview-of-artificial-intelligence-and-role-of-natural-language-processing-in-big-data
- https://twitter.com/

Overview

- Basic idea of Web Mining
- Opportunities and challenges in Web Mining
- Data mining v/s Web Mining
- Classification of Web Mining
- Summary of Paper: You are what you tweet!
- Social media Mining

What is web mining?

- Web is a collection of interrelated files on Web servers.
- In recent years, we have seen an exponential rise in the number of HTML documents, images, multimedia files which are available on the world wide web.
- Considering the heterogeneity of these files, it is very difficult to retrieve interesting information from it.
- Web mining is the application of data mining to extract such interesting information from the internet.

Opportunities

Web offers unprecedented opportunities to data mining

- Abundant and easily accessible data
- Huge variety of data: structured, semi structured, images, multimedia etc.
- Most of the data on web is linked: There are hyperlinks among pages within a site.
- Much of the data is redundant: The same piece of information or its variants appear in number of pages.

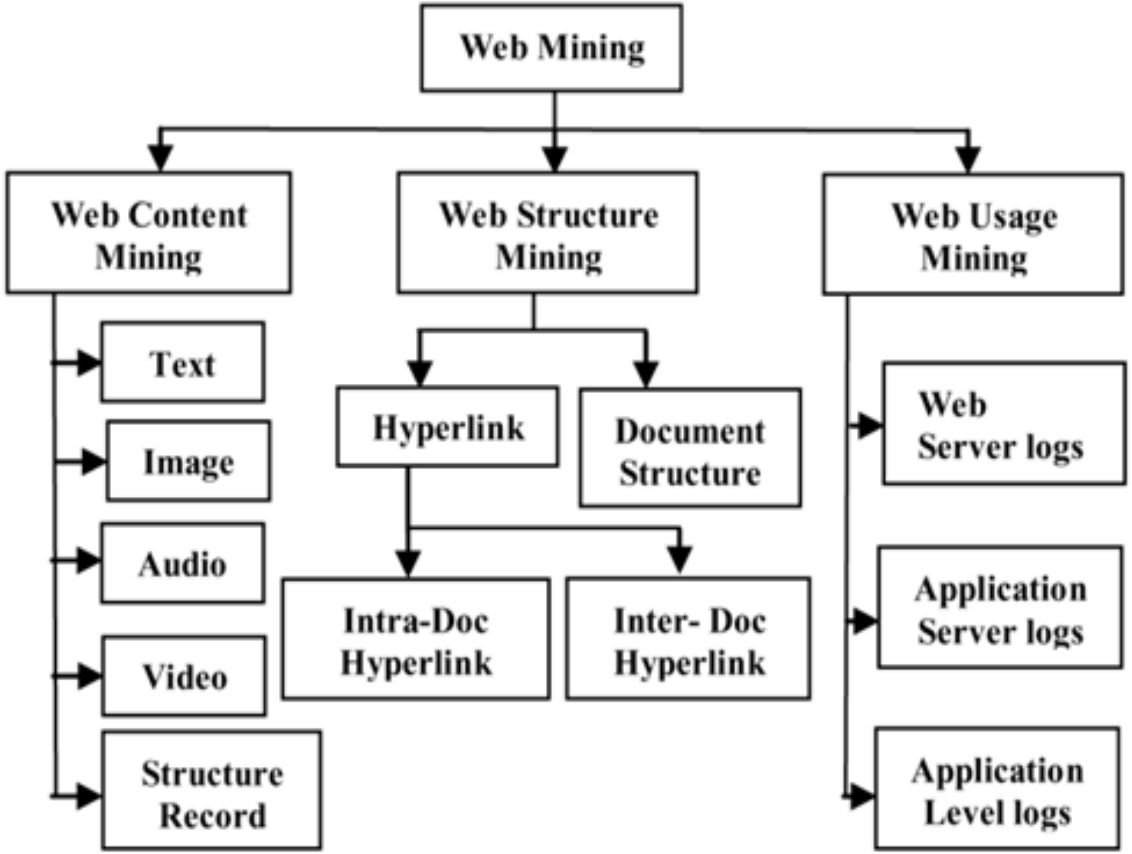
Challenges

Along with opportunities, there are serious challenges in web mining

- **Web is noisy** : A Web page typically contains a mixture of many kinds of information, e.g., main contents, advertisements, navigation panels, copyright notices, etc.
- **Web is dynamic** : Information on the Web changes constantly. Keeping up with the changes and monitoring the changes are important issues.
- **Web is a virtual society** : It is not only about data, information and services, but also about interactions among people, organizations and automatic systems, i.e., communities.
- Many other such restrictions pose a pretty big challenge for mining the web.

Data mining v/s Web mining

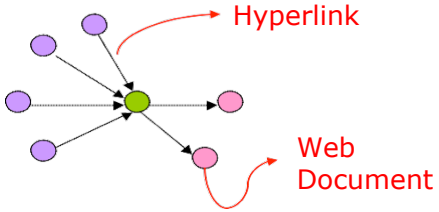
- Web mining is an application of data mining techniques.
- Web mining is studied as a specific branch of data mining to consider the specific structures of the available web data.
- Web data:
  - Web content : text, images etc.
  - Web usage: http logs, app server logs etc.
  - Web structure: hyperlinks, tags


Web Mining
Web Content Mining
Web Structure Mining
Web Usage Mining
Text
Image
Audio
Video
Structure Record
Hyperlink
Document Structure
Intra-Doc Hyperlink
Inter- Doc Hyperlink
Web Server logs
Application Server logs
Application Level logs

Web Structure Mining

What is Web Structure Mining?

The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages



Web Graph Structure

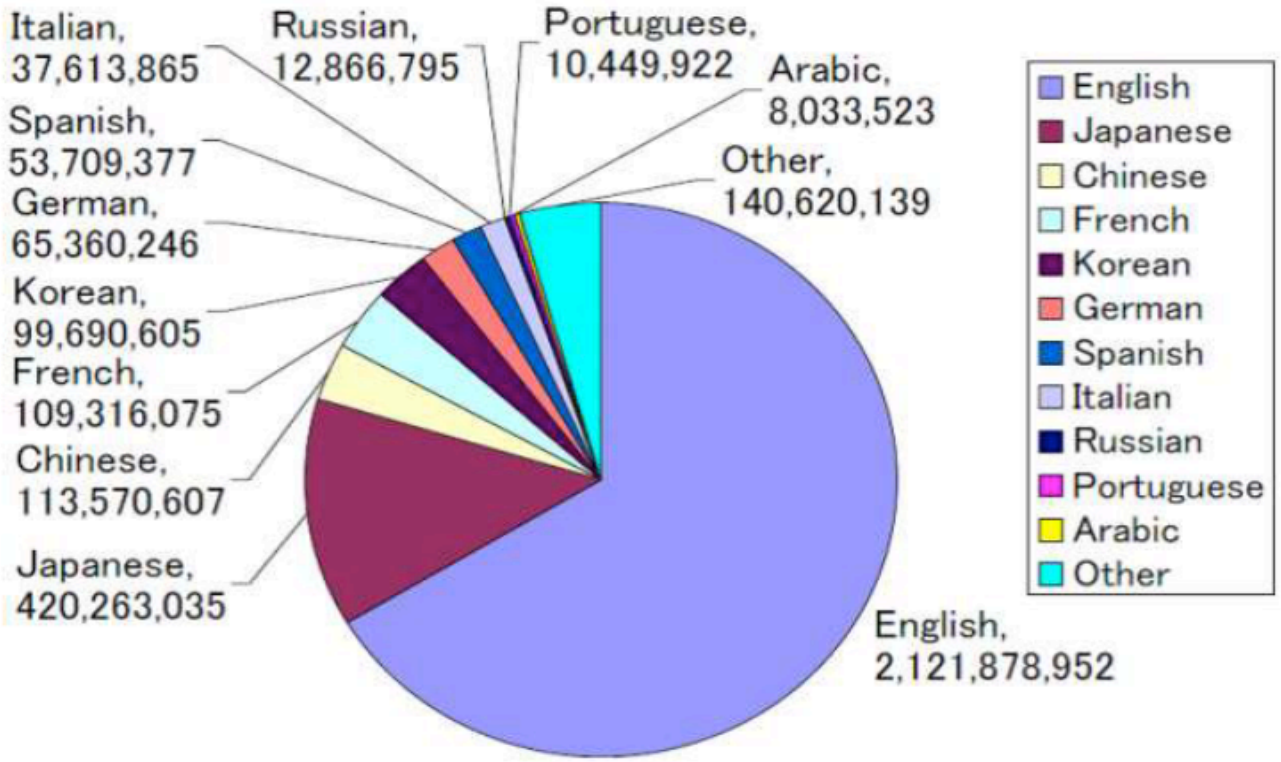
Web Structure Mining is the process of discovering structure information from the Web

- This type of mining can be performed either at the (intra-page) document level or at the (inter-page) hyperlink level
- The research at the hyperlink level is also called *Hyperlink Analysis*

Motivation to study Hyperlink Structure

- Hyperlinks serve two main purposes.
  - Pure Navigation.
  - Point to pages with having relevant information.
- This can be used to retrieve useful information from the web.

Web Structure by Language



Source: https://www.slideshare.net/AmirFahmideh/web-mining-structure-mining

Web Structure Terminology

- ***Web-graph:*** A directed graph that represents the Web.
- ***Node:*** Each Web page is a node of the Web-graph.
- ***Link:*** Each hyperlink on the Web is a directed edge of the Web-graph.
- ***In-degree*****:** The in-degree of a node, *p,* is the number of distinct links that point to *p.*
- ***Out-degree*****:** The out-degree of a node, *p,* is the number of distinct links originating at *p* that point to other nodes.

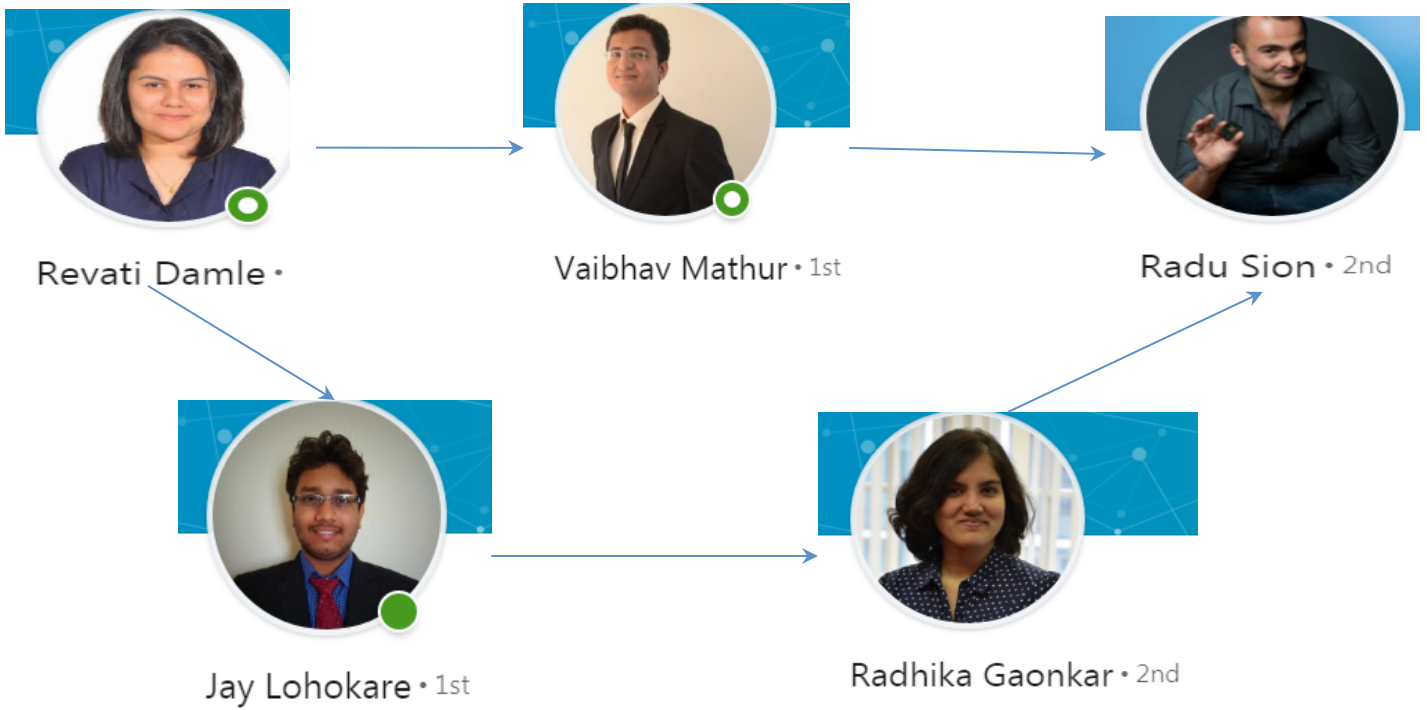
Web Structure Terminology(2)

- ***Directed Path*****:** A sequence of links, starting from *p* that can be followed to reach *q.*

- ***Shortest Path:*** Of all the paths between nodes *p* and *q,* which has the shortest length, i.e. number of links on it.

- ***Diameter*****:** The maximum of all the shortest paths between a pair of nodes *p* and *q,* for all pairs of nodes *p* and *q* in the Web-graph.

LinkedIn : Shortest Path Example



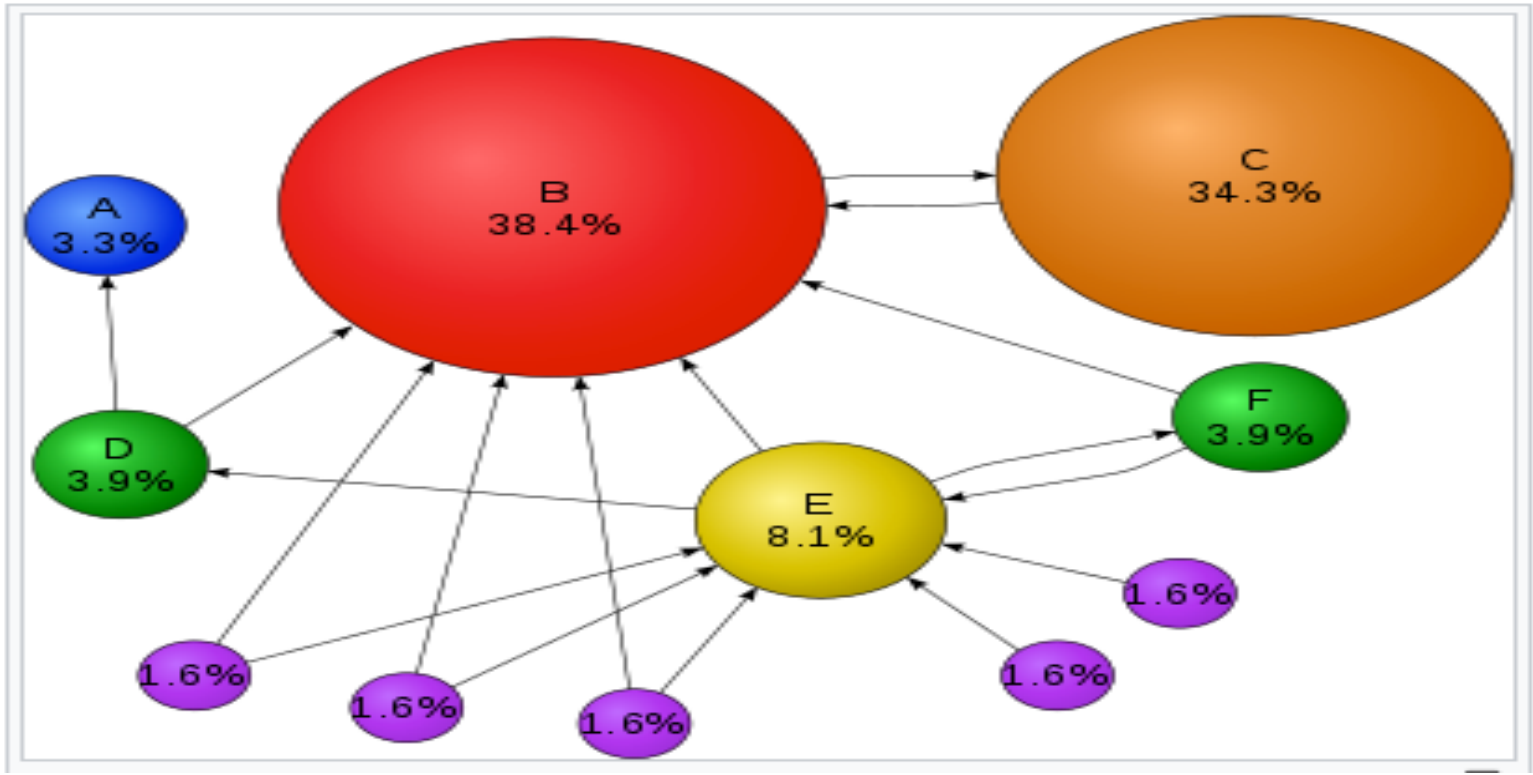
Google Page Rank Algorithm

- **Page Rank** (PR) is an algorithm used by Google Search to rank websites in their search engine results.

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)}$$

i.e. the PageRank value for a page **u** is dependent on the PageRank values for each page **v** contained in the set $\mathbf{B_u}$ (the set containing all pages linking to page **u**), divided by the number *L*(*v*) of links from page **v**.

Page Rank (Cont.)



Source: https://en.wikipedia.org/wiki/PageRank

Web Structure Applications

Web Structure is a useful source for extracting information such as

1. Quality of Web Page - *Ranking of web pages*
2. Interesting Web Structures - *Graph patterns like Co-citation, Social choice, etc.*
3. Web Page Classification - *Classifying web pages according to various topics*
4. Finding Related Pages - *Given one relevant page, find all related pages*
5. Detection of duplicate pages - *Detection of neared-mirror sites to eliminate duplication*

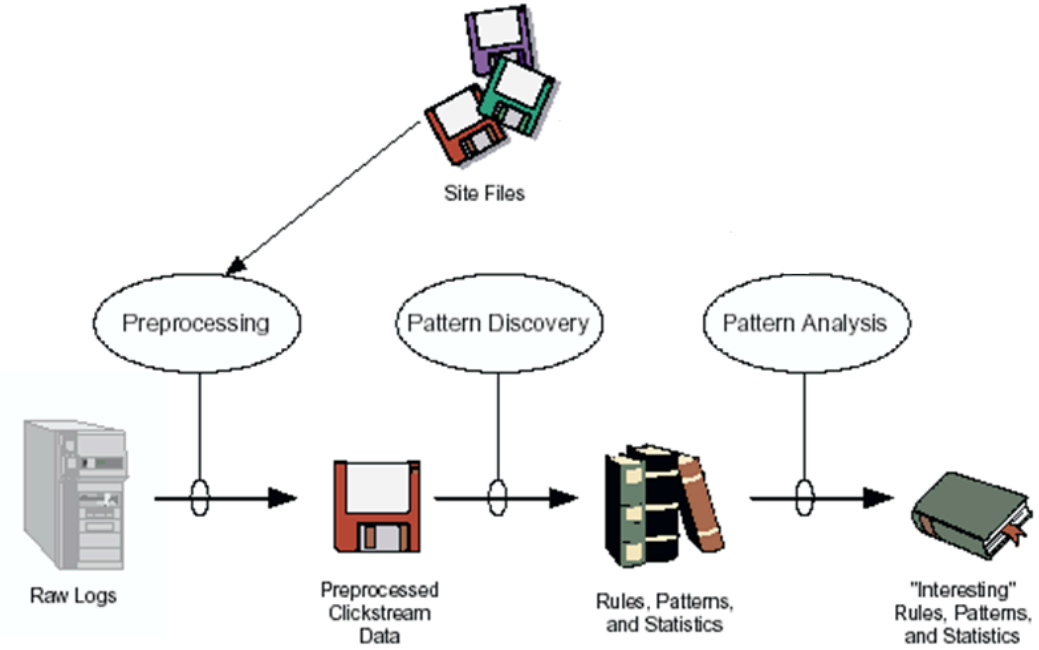
Web Usage Mining

Web Usage Mining

- **Web Usage Mining** is the process of applying data mining techniques to the discovery of usage patterns from Web data, in order to understand and better serve the needs of Web-based applications.

- Web Usage Mining utilizes data from user interactions with a Web sites, including Web logs, click streams and database transaction at a Web site or a group of related sites.

Web Usage Mining Processes



Web Usage Mining - Pattern Discovery Tasks

- Statistical Analysis
- Clustering
- Classification
- Association Rules

Web Usage Mining - Pattern Discovery Tasks (Cont.)

Statistical Analysis:

- Different kinds of descriptive statistical analyses (frequency, mean, median, etc.) on variables such as page views, viewing time and length of a navigational path gives useful knowledge.

Clustering:

- Clustering is a technique to group together a set of items having similar characteristics.
- In the Web Usage domain, there are two kinds of interesting clusters to be discovered :
- Clustering of **users** : discover groups of users with similar navigation patterns. => Perform market segmentation in E-commerce.
- Clustering of **pages:** discover groups of pages having related content => Useful for search engines

Web Usage Mining - Pattern Discovery Tasks (Cont.)

Classification: Classification is the task of mapping a data item into one of several predefined classes.

- In the Web domain, one is interested in developing a profile of users belonging to a particular class or category. Uses Decision Tree classifiers, Naive Bayesian classifiers, Neural Networks, SVM etc.

Association Rules:

- **Given:** A database of transactions, where each transaction is a list of items. **Find:** all rules that associate the presence of one set of items with that of another set of items.
- For web mining, it refer to sets of pages accessed together with a support value exceeding some specified threshold.
- Are applicable for business and marketing applications, and can help Web designers to restructure their Website.

Source: http://www3.cs.stonybrook.edu/~cse634/L8ch5assoc.pdf

Web Usage Mining - Pattern Analysis

- Last step in the overall Web Usage mining process.
- Motivation : Filter out uninteresting rules or patterns from the set found in the pattern discovery phase.
- The exact analysis methodology - governed by the application for which Web mining is done.
- The most common form of pattern analysis consists of:

– A knowledge query mechanism (Like SQL).

– Visualization techniques (Like graphing patterns or assigning colors to different values) - highlight overall patterns or trends.

Web Usage Mining Application: User Profiles

- The Web has taken user profiling to completely new levels.
- In a ’brick and-mortar’ store, data collection happens only at the checkout counter (’point-of-sale’).
- In an online store, the complete click-stream is recorded:
  - Provides a detailed record of every single action taken by the user.
  - Allows creating a detailed user profile
- Most organizations build profiles using user behavior limited to their own sites (IMDB, Netflix).
- Web-wide profiling also exists (Facebook, Google)

Web Usage Mining Application: User Profiles

Amazon's Recommendation Systems:

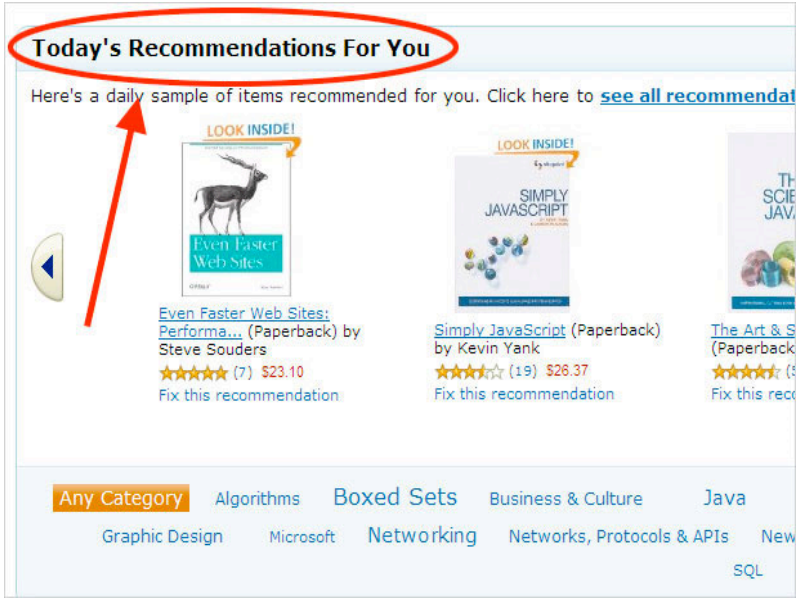
The data Amazon mines:

- Purchased shopping carts = real money from real people spent on real items.
- Wishlists - what's on Amazon specifically for you.
- Demographic information they know what is popular in your general area
  for your kids, yourself, your spouse, etc.
- User segmentation = did you buy 3 books in separate months for a toddler?
  likely you have a kid.

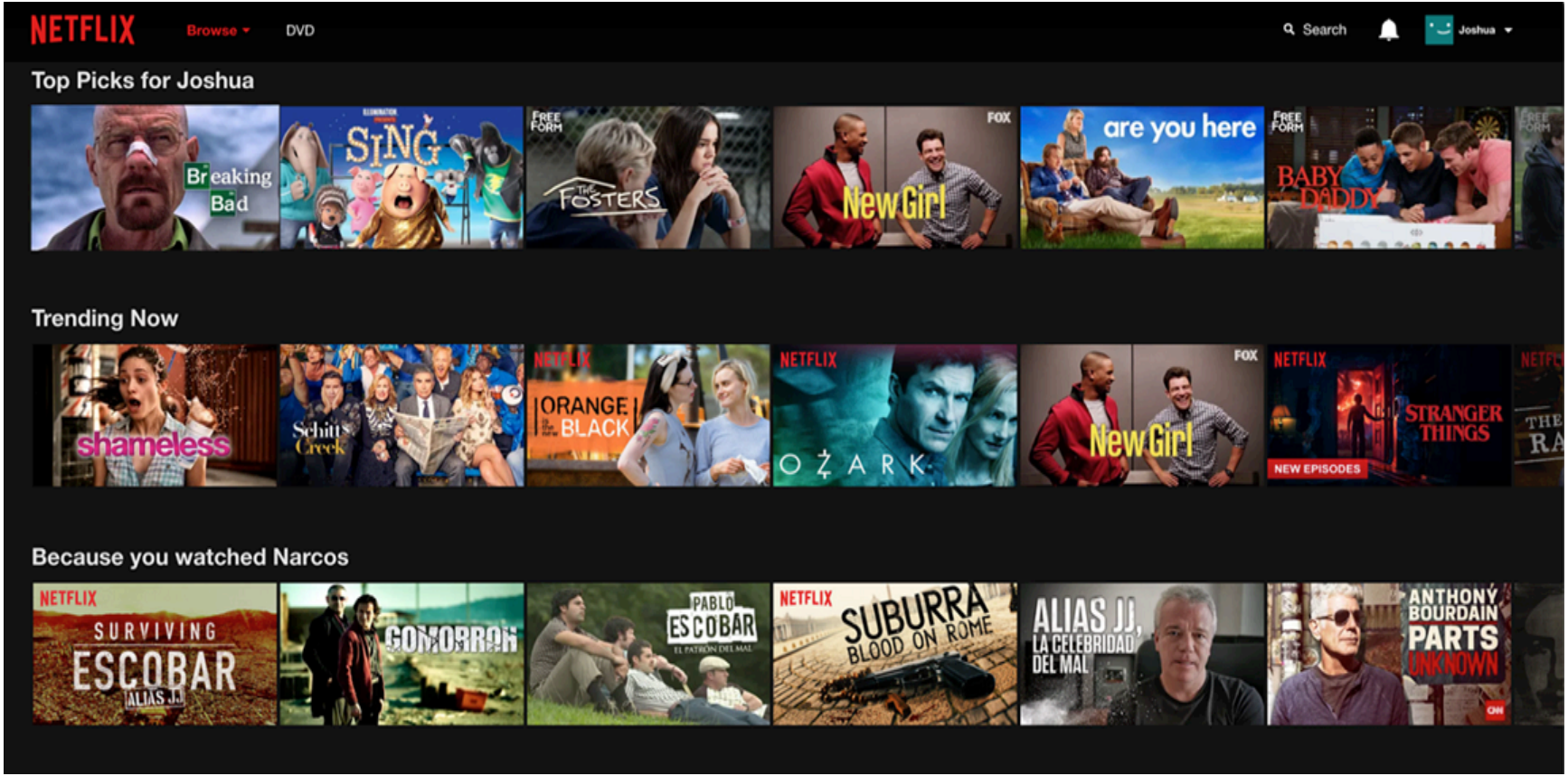
And lots more!

Source: https://www.cs.umd.edu/~samir/498/Amazon-Recommendations.pdf

Web usage Mining Application: User Profiles



AMAZON's Recommendation System


NETFLIX
Browse
DVD
Search
Joshua
Top Picks for Joshua
Breaking Bad
SING
THE FOSTERS
New Girl
are you here
BABY DADDY
Trending Now
shameless
Schitt's Creek
ORANGE IS THE NEW BLACK
OZARK
New Girl
STRANGER THINGS
NEW EPISODES
Because you watched Narcos
SURVIVING ESCOBAR ALIAS JJ
GOMORRAH
PABLO ESCOBAR EL PATRÓN DEL MAL
SUBURRA BLOOD ON ROME
ALIAS JJ, LA CELEBRIDAD DEL MAL
ANTHONY BOURDAIN PARTS UNKNOWN

On a lighter note

Web Content Mining

Web Content Mining

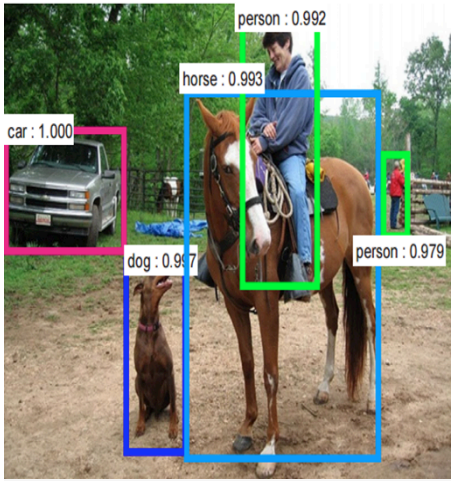
Extraction of useful information from the contents of Web documents (structured and unstructured data)

Text

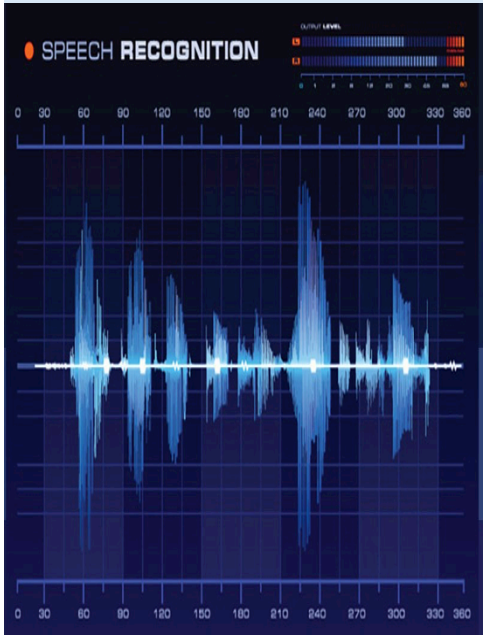
The New York Times

WIKIPEDIA
The Free Encyclopedia

Image



Audio



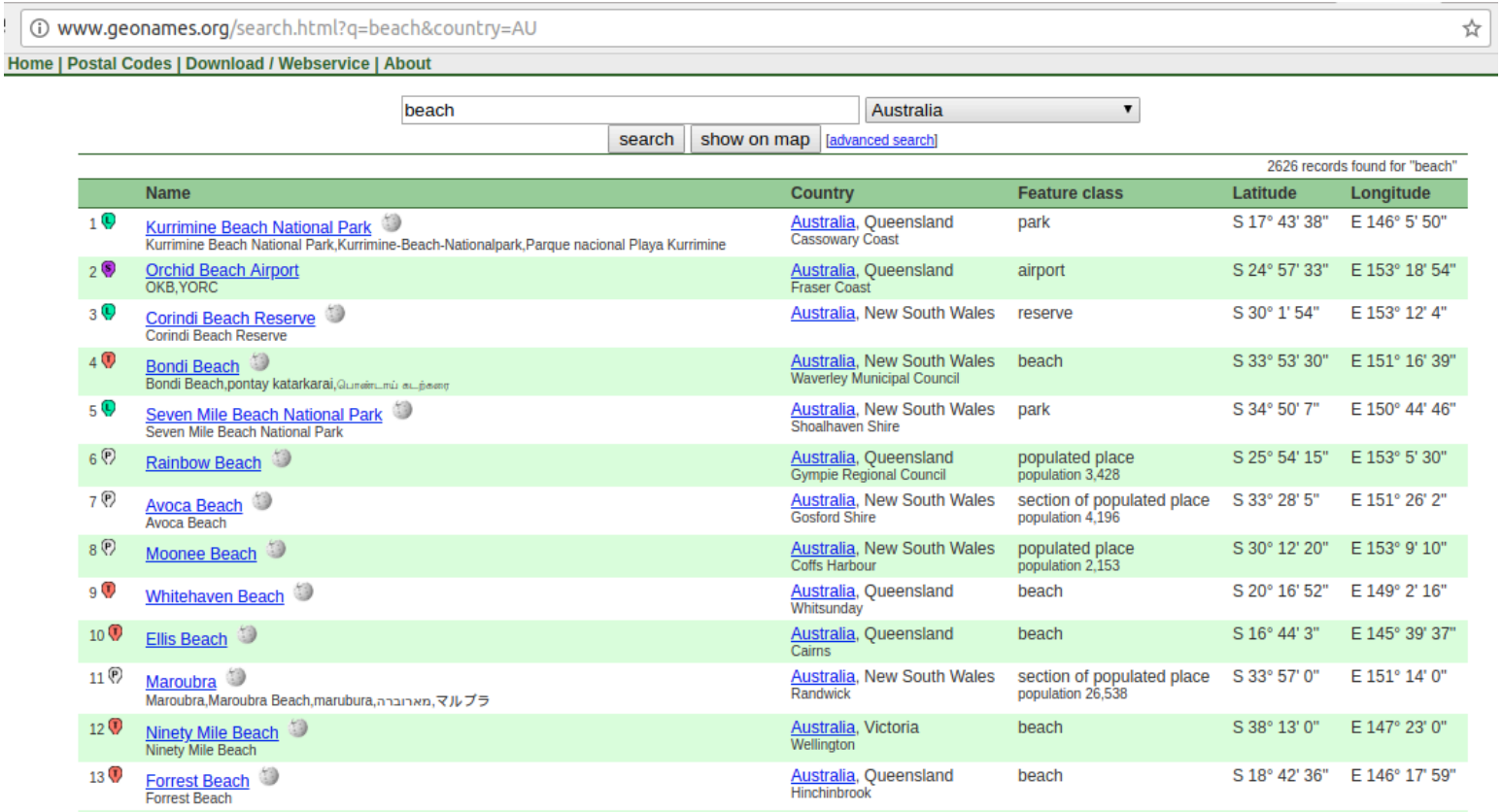
Video

source : https://talkroute.com/automatic-speech-recognition-mixing-it-with-ivr/ , https://www.underconsideration.com/brandnew/archives/new_logo_for_youtube_done_in_house.php

Structured Content

source : http://www.3idatascraping.com/how-to-use-data-scraping-to-mine-structured-data-from-the-unstructured-data.php

Geonames geographical gazetteer



www.geonames.org/search.html?q=beach&country=AU

Home | Postal Codes | Download / Webservice | About

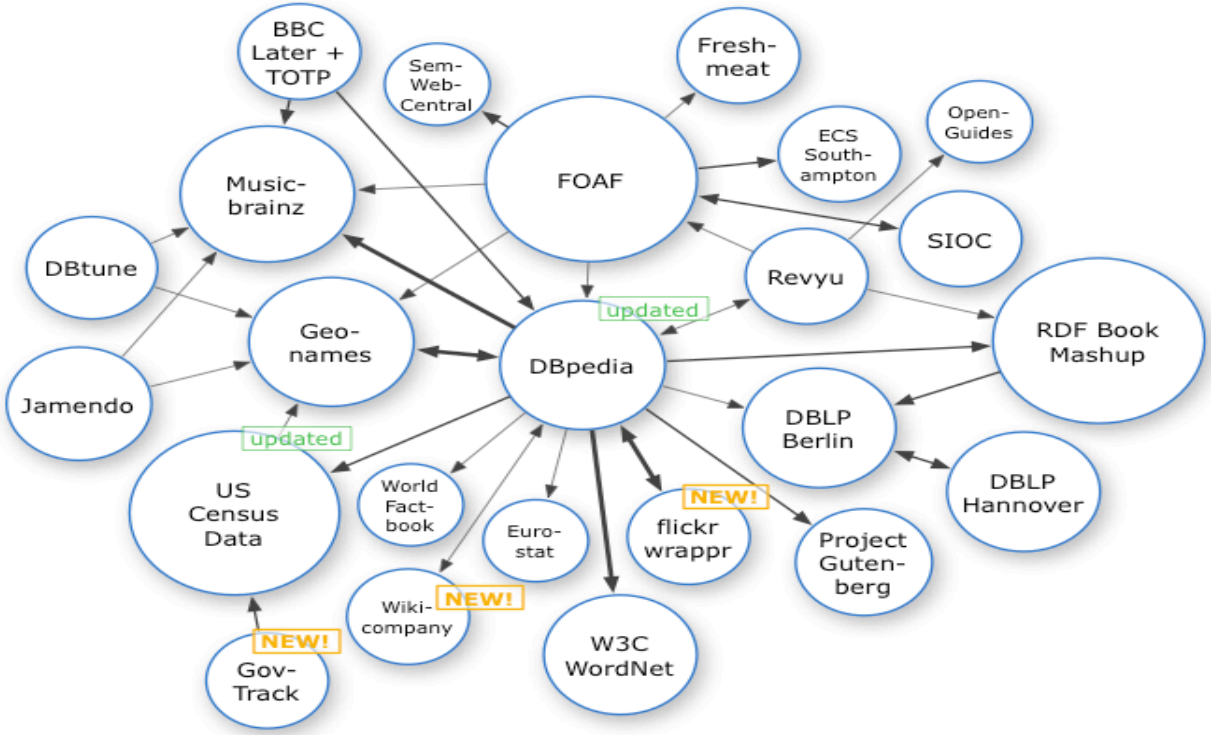
beach — Australia — search — show on map — [advanced search]

2626 records found for "beach"

	Name	Country	Feature class	Latitude	Longitude
1	Kurrimine Beach National Park Kurrimine Beach National Park,Kurrimine-Beach-Nationalpark,Parque nacional Playa Kurrimine	Australia, Queensland Cassowary Coast	park	S 17° 43' 38"	E 146° 5' 50"
2	Orchid Beach Airport OKB,YORC	Australia, Queensland Fraser Coast	airport	S 24° 57' 33"	E 153° 18' 54"
3	Corindi Beach Reserve Corindi Beach Reserve	Australia, New South Wales	reserve	S 30° 1' 54"	E 153° 12' 4"
4	Bondi Beach Bondi Beach,pontay katarkarai,பொண்டாய் கடற்கரை	Australia, New South Wales Waverley Municipal Council	beach	S 33° 53' 30"	E 151° 16' 39"
5	Seven Mile Beach National Park Seven Mile Beach National Park	Australia, New South Wales Shoalhaven Shire	park	S 34° 50' 7"	E 150° 44' 46"
6	Rainbow Beach	Australia, Queensland Gympie Regional Council	populated place population 3,428	S 25° 54' 15"	E 153° 5' 30"
7	Avoca Beach Avoca Beach	Australia, New South Wales Gosford Shire	section of populated place population 4,196	S 33° 28' 5"	E 151° 26' 2"
8	Moonee Beach	Australia, New South Wales Coffs Harbour	populated place population 2,153	S 30° 12' 20"	E 153° 9' 10"
9	Whitehaven Beach	Australia, Queensland Whitsunday	beach	S 20° 16' 52"	E 149° 2' 16"
10	Ellis Beach	Australia, Queensland Cairns	beach	S 16° 44' 3"	E 145° 39' 37"
11	Maroubra Maroubra,Maroubra Beach,marubura,מארובר,マルブラ	Australia, New South Wales Randwick	section of populated place population 26,538	S 33° 57' 0"	E 151° 14' 0"
12	Ninety Mile Beach Ninety Mile Beach	Australia, Victoria Wellington	beach	S 38° 13' 0"	E 147° 23' 0"
13	Forrest Beach Forrest Beach	Australia, Queensland Hinchinbrook	beach	S 18° 42' 36"	E 146° 17' 59"

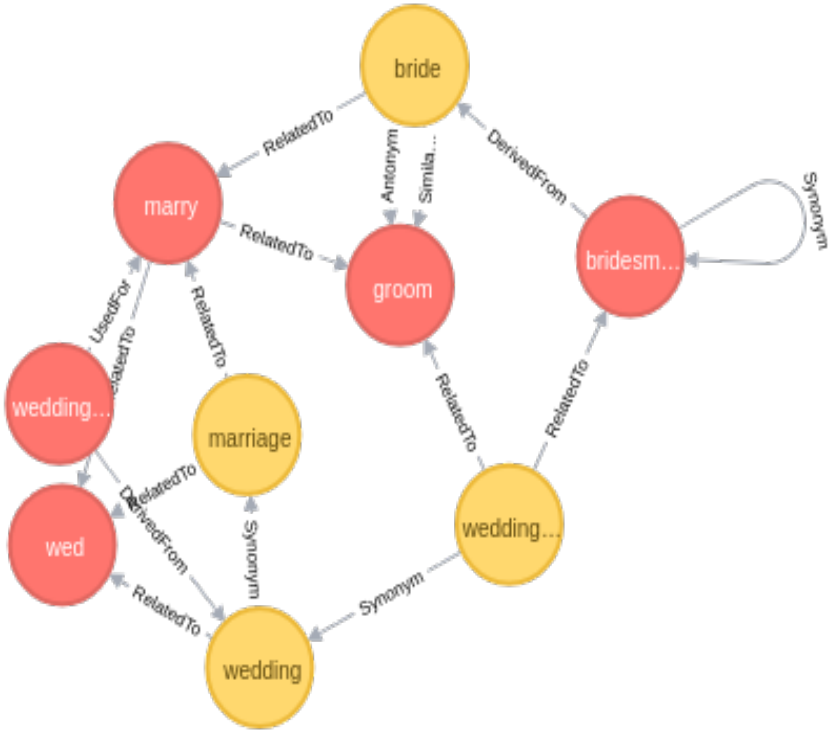
source : http://www.geonames.org/

Dbpedia (Structured data from Wikipedia and other sources)



source : http://lod-cloud.net/

Wordnet - Structured information for NLP



- **WordNet® is a large lexical database of English.**
- **Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept.**
- **Synsets are interlinked by means of conceptual-semantic and lexical relations.**

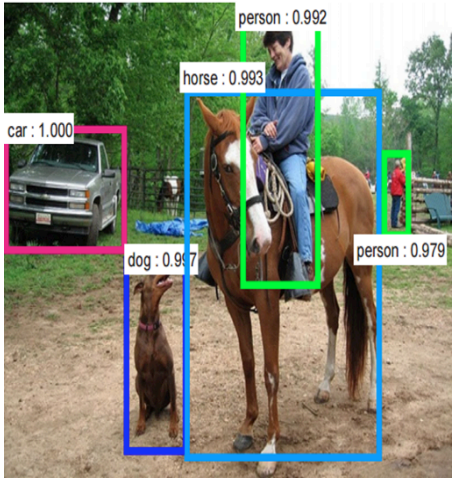
source : http://tomkdickinson.co.uk/2017/05/combining-wordnet-and-conceptnet-in-neo4j/

Web Content Mining

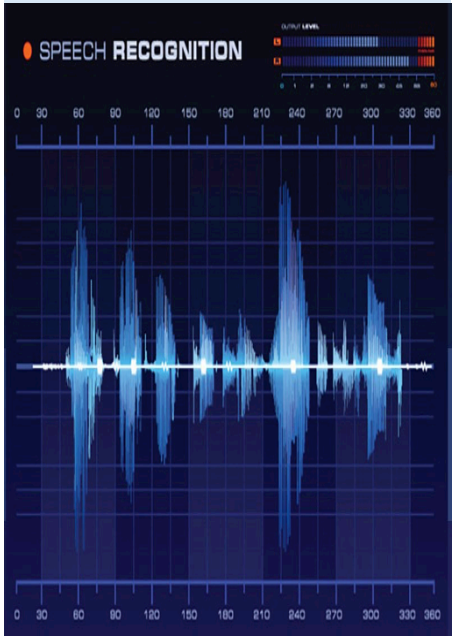
Extraction of useful information from the contents of Web documents

Text

Image



Audio

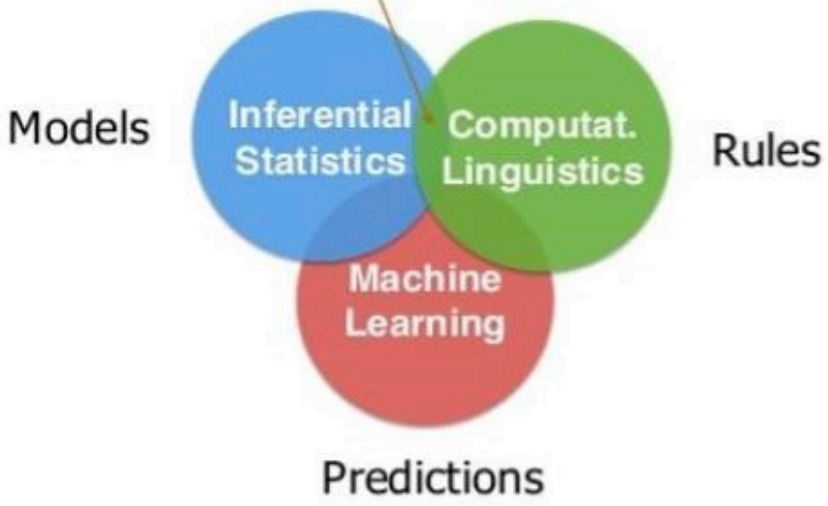


Video

source : https://talkroute.com/automatic-speech-recognition-mixing-it-with-ivr/ , https://www.underconsideration.com/brandnew/archives/new_logo_for_youtube_done_in_house.php

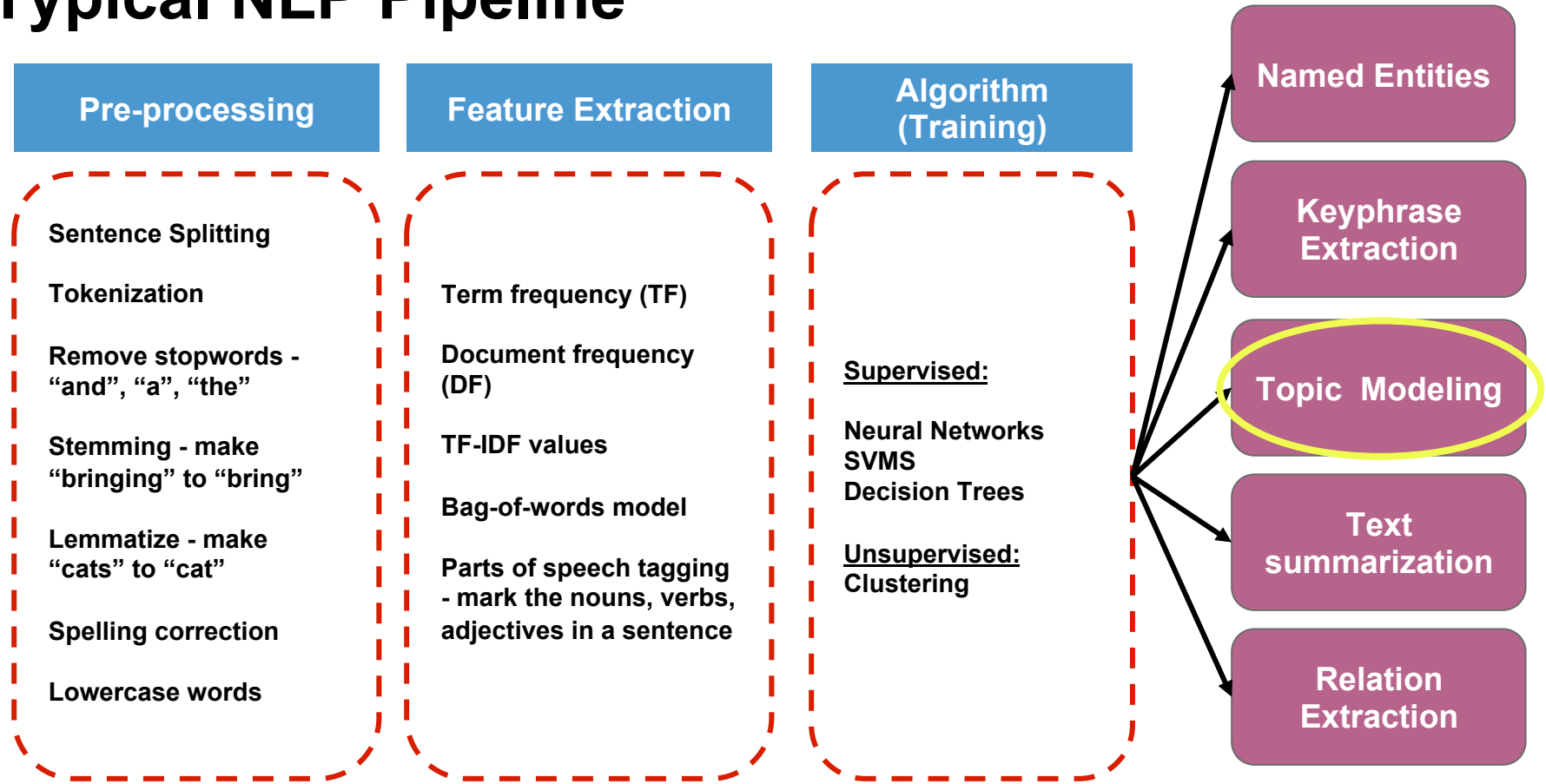
"Text Mining" or "Text Analytics"

The discovery of {new or existing} facts by applying **natural language processing** ("NLP") & statistical learning techniques.

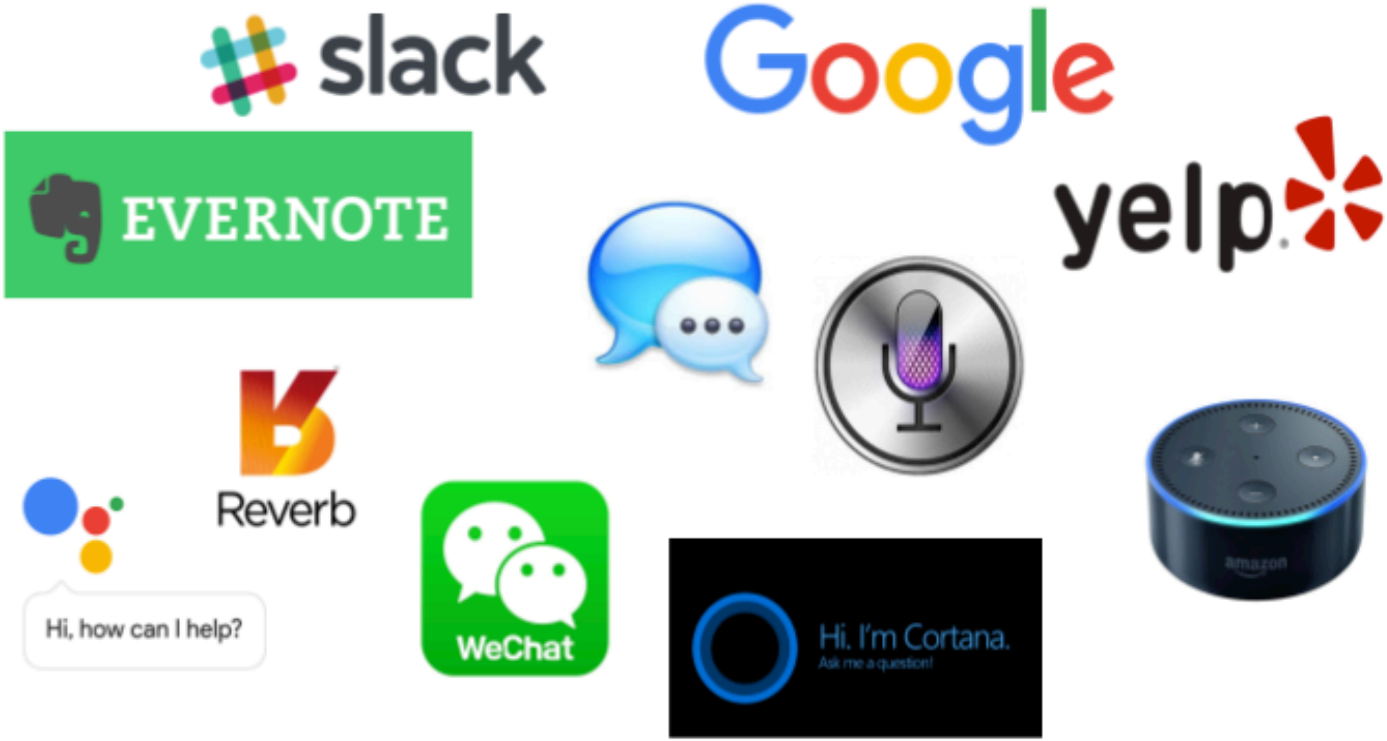


source : https://www.slideshare.net/asdkfjqlwef/statistical-text-mining-introduction-florian-leitner

Typical NLP Pipeline



Who is using text mining?



source : http://bytecubed.com/natural-language-processing-for-everyday-people/

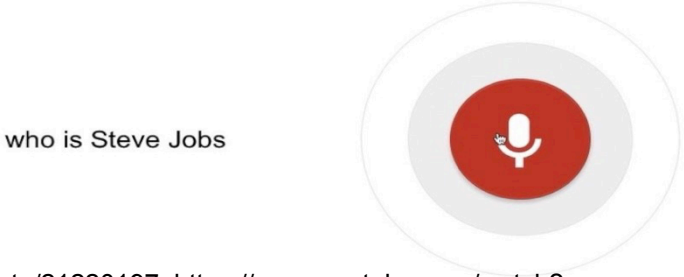
Text mining with Web Data

Spam Filtering

Machine Translation

Trend Analysis

Speech Recognition



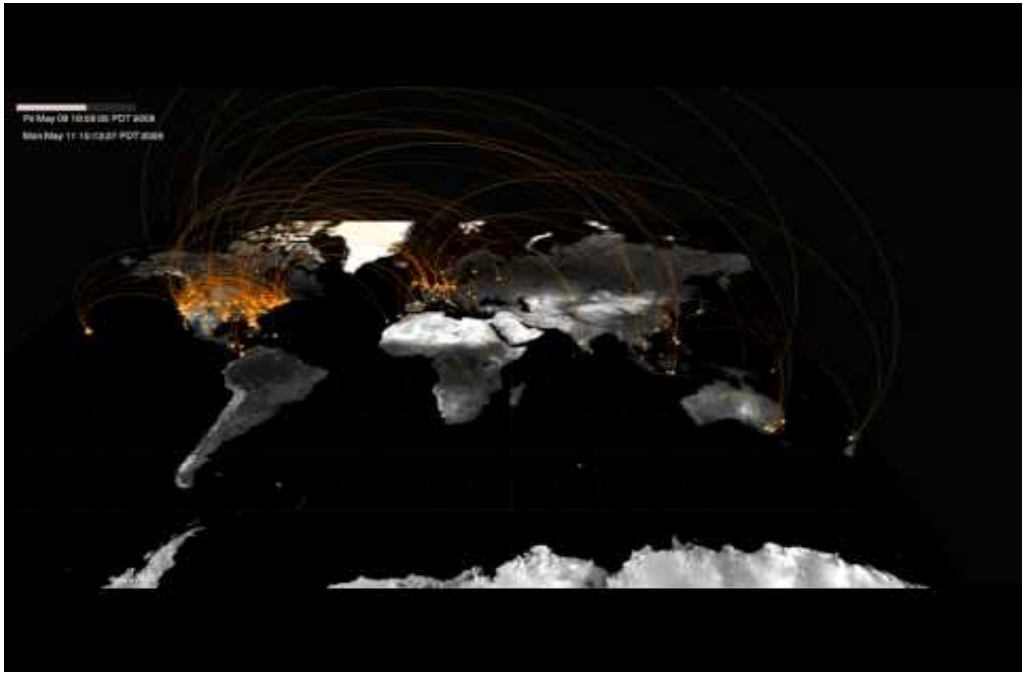
source : https://codecanyon.net/item/easy-google-translate/21220197, https://www.youtube.com/watch?v=Kor-A3mGRnA, https://irchs.org/google-tools/gmail-logo/

Quality insights through social media mining

- Twitter brand monitoring through sentiment analysis of customer tweets
- Customer loyalty analysis by extracting sentiments and topics from user posts on facebook, twitter, instagram
- Disease or epidemic outbreaks from tweets
- Monitoring signs of mental health problems in users from their tweets
- Analyzing social networks for election trends
- New business ventures using big data technologies, visualization dashboards, social media mining
- Digital marketing
- Customer acquisition and customer retention
- Predicting business sales
- Finding latest trends and patterns in population

source : http://www.xsjjys.com/facebook-wallpapers.html

Tracking Flu and People Movement from Twitter



Source: http://www.youtube.com/watch?v=rUuPBfEkiJs

“You Are What You Tweet” : Analyzing Twitter for Public Health

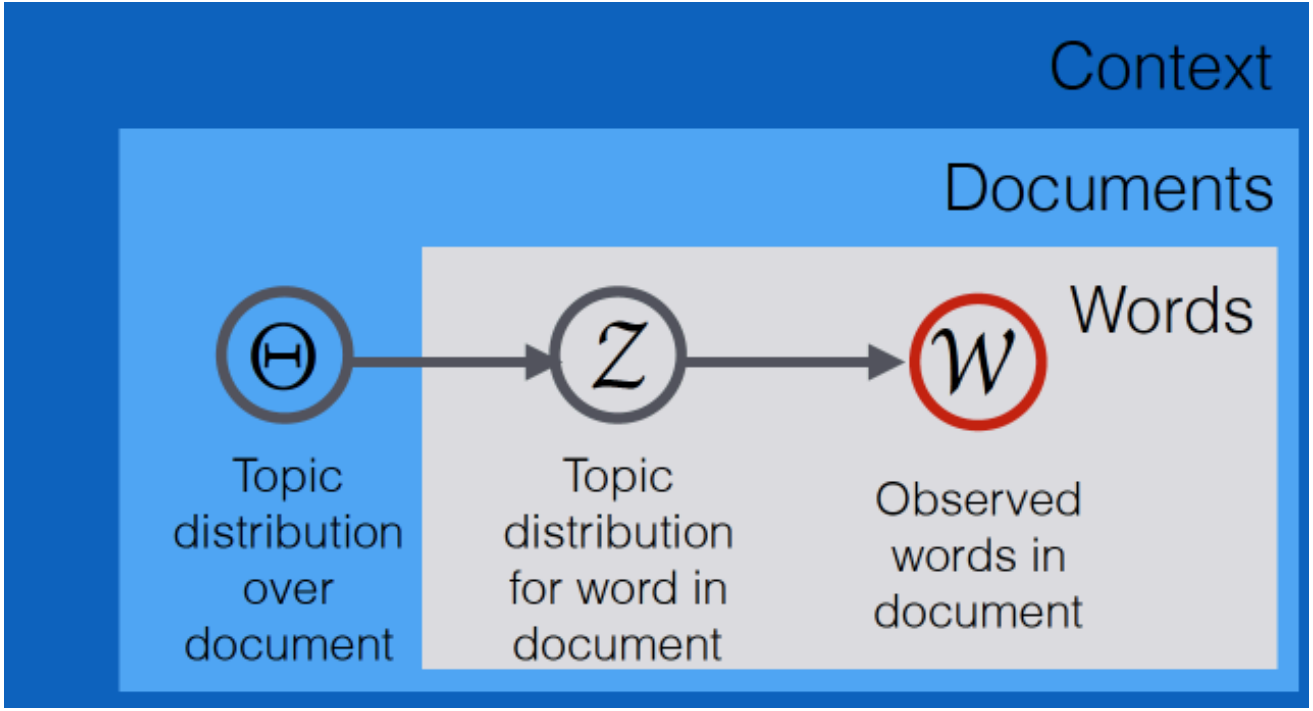
Authors : Paul, Michael J., and Mark Dredze.

AAAI Publications, Fifth International AAAI Conference on Weblogs and Social Media, 2011

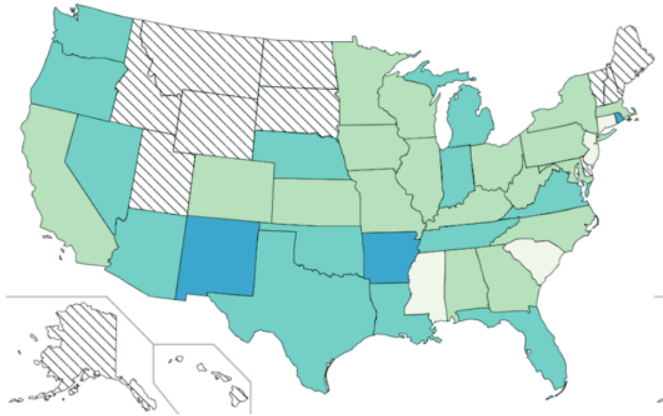
1. Analyze Twitter data to find public health characteristics.
2. Applies Ailment Topic Aspect Model over sentiments
3. Prior distribution using articles related to ailments
4. Measuring behavioral risk factors, localizing illness by geography, analyze symptoms and medical usage
5. Quantitative public health data
6. Qualitative evaluations

Ailment Topic Aspect Model

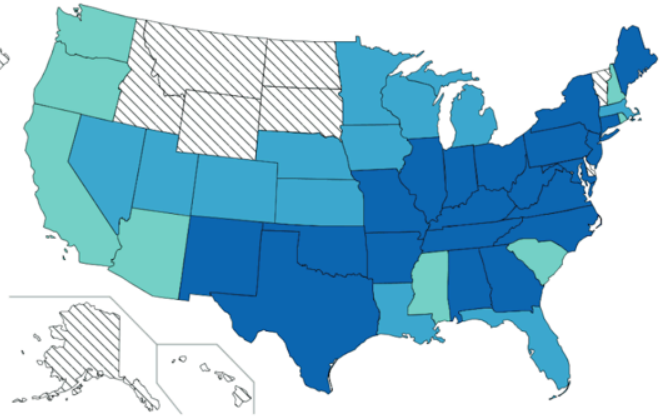
LDA (Latent Dirichlet allocation) - Each document is a mixture of topics. Each words is attributable to one of the topics.



watery
helsinki
mold
watering
faucet
lolss
sneezes
sneezy
teary
bloom



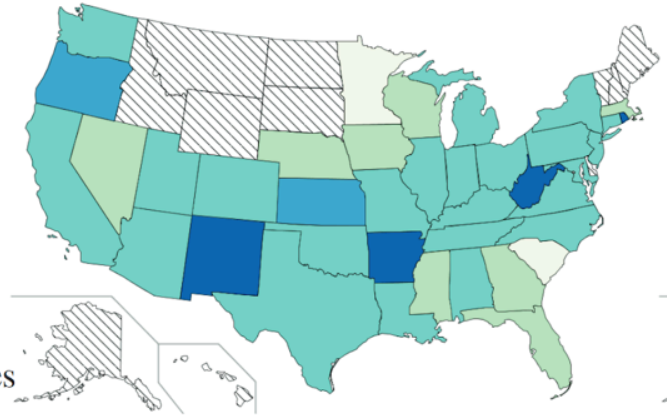
(a) February



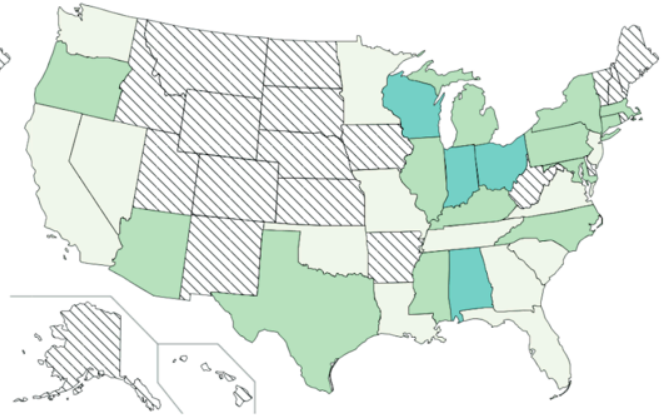
(b) April

pollen
zyrtec
claritin
spring
watering
trees
watery
itching
bloom
grass

hayfever
grass
watering
watery
claritin
humidity
zyrtec
bonkers
mold
antihistamines

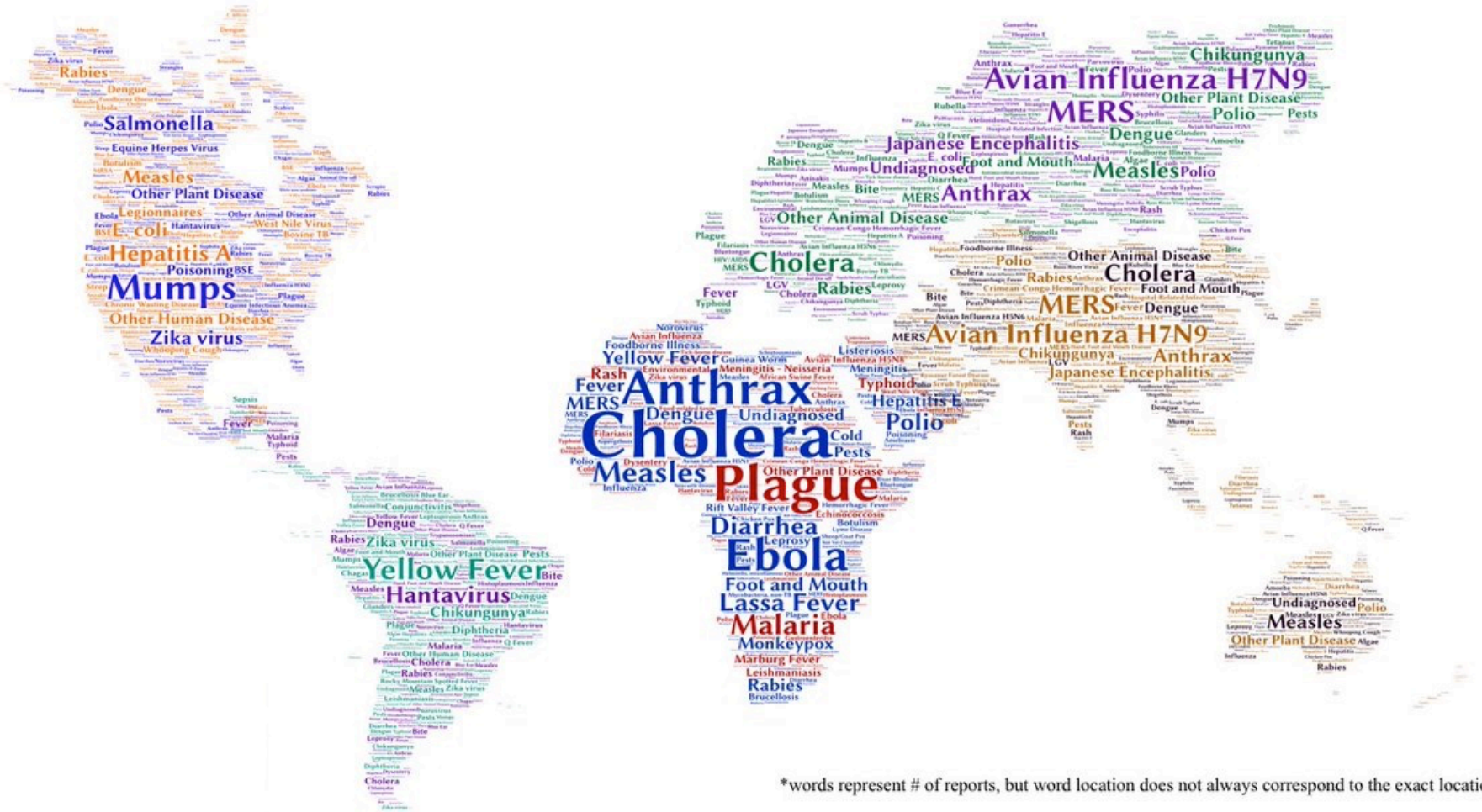


(c) June

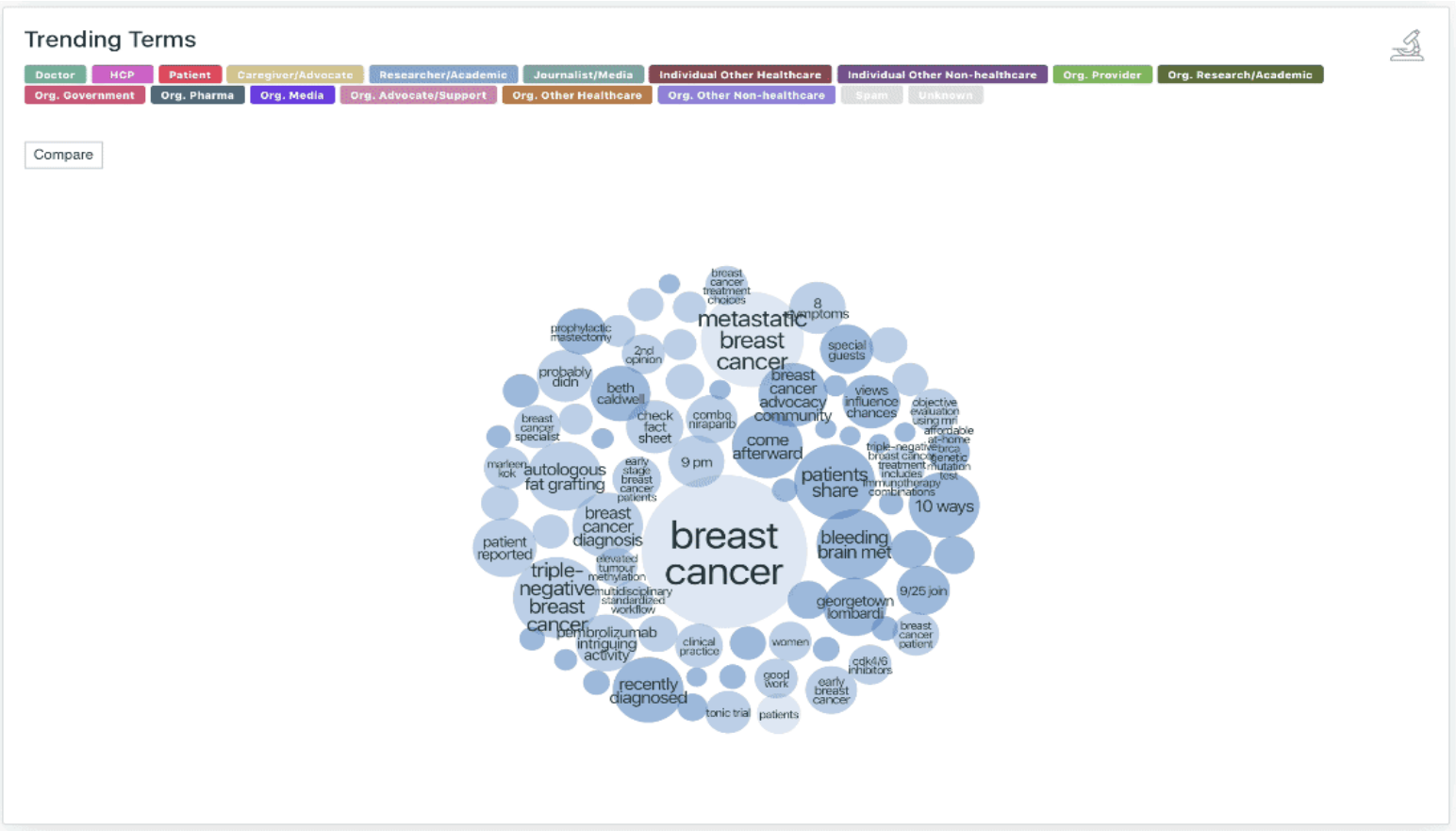


(d) August

dust
catherine
pinche
buildings
mold
hadd
gato
cessation
meditating
ragweed



*words represent # of reports, but word location does not always correspond to the exact location



https://dashboard.symplur.com/dashboard

Social media mining - Approach

A relative new domain closely related to web-mining: Web content mining + Web Usage mining

Capturing the data -

1. REST APIs exposed by platforms (Twitter, Facebook, etc)

2. Web crawling

Data architecture -

Primarily big data architecture - Petabytes of data on social media (TBs generated daily - 2.7 billion FB likes, 98 million posts, per day)

Can be real time or batch processing system

Social-media/web mining in news

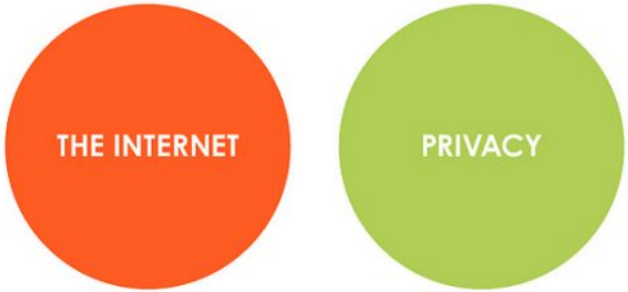
1. Facebook & Cambridge analytica - Using social media data to influence elections.
2. China - Using social media data to create credit profiles of users.
3. USA to check social media history for VISA applicants for background checks
4. Customized advertisements based on Social media content
5. LiveRamp - Customer data platform (Aggregates all data from Online/Offline customer interaction points

Privacy Issues

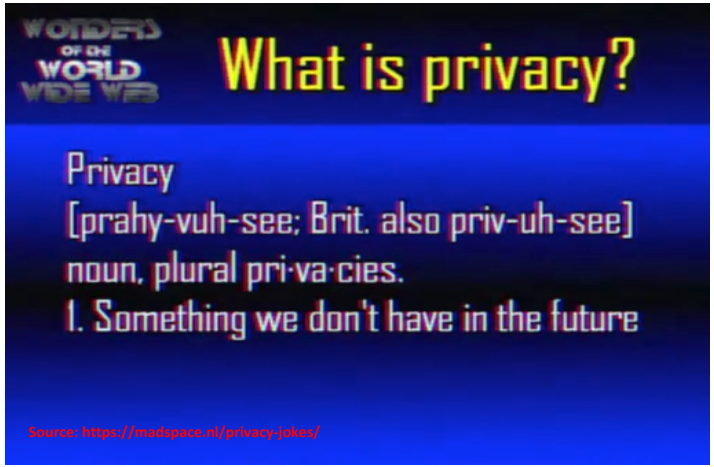
- Privacy is a sensitive topic which has been attracting a lot of attention recently due to rapid growth of ecommerce and social media.
- Users want to maintain strict anonymity on the Web.
- On the other hand, site administrators are interested in finding out the demographics of users as well as the usage statistics of different sections of their Web site.
- The main challenge is to come up with guidelines and rules such that site administrators can perform analytics on the usage data without compromising the identity of an individual user.

Privacy Issues

- Furthermore, there should be strict regulations to prevent the usage data from being exchanged/sold to other sites.
- The users should be made aware of the privacy policies followed by any given site.



A HELPFUL VENN DIAGRAM



Source: https://madspace.nl/privacy-jokes/