Introduction to Text Mining

Hongning Wang

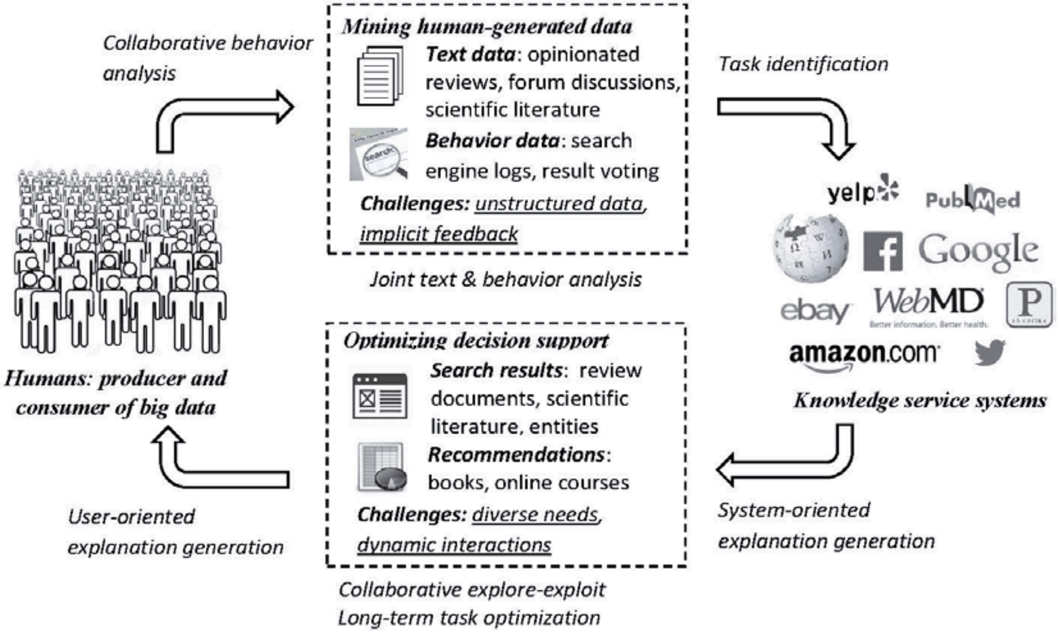
CS@UVa

Who Am I?

- Hongning Wang
  - Assistant professor in CS@UVa since August 2014
  - Research areas
    - Information retrieval
    - Data mining
    - Machine learning

Who Am I?

- Hongning Wang
  - Assistant professor in CS@UVa since August 2014



Collaborative behavior analysis

Mining human-generated data

Text data: opinionated reviews, forum discussions, scientific literature

Behavior data: search engine logs, result voting

Challenges: unstructured data, implicit feedback

Task identification

Joint text & behavior analysis

Knowledge service systems

System-oriented explanation generation

Optimizing decision support

Search results: review documents, scientific literature, entities

Recommendations: books, online courses

Challenges: diverse needs, dynamic interactions

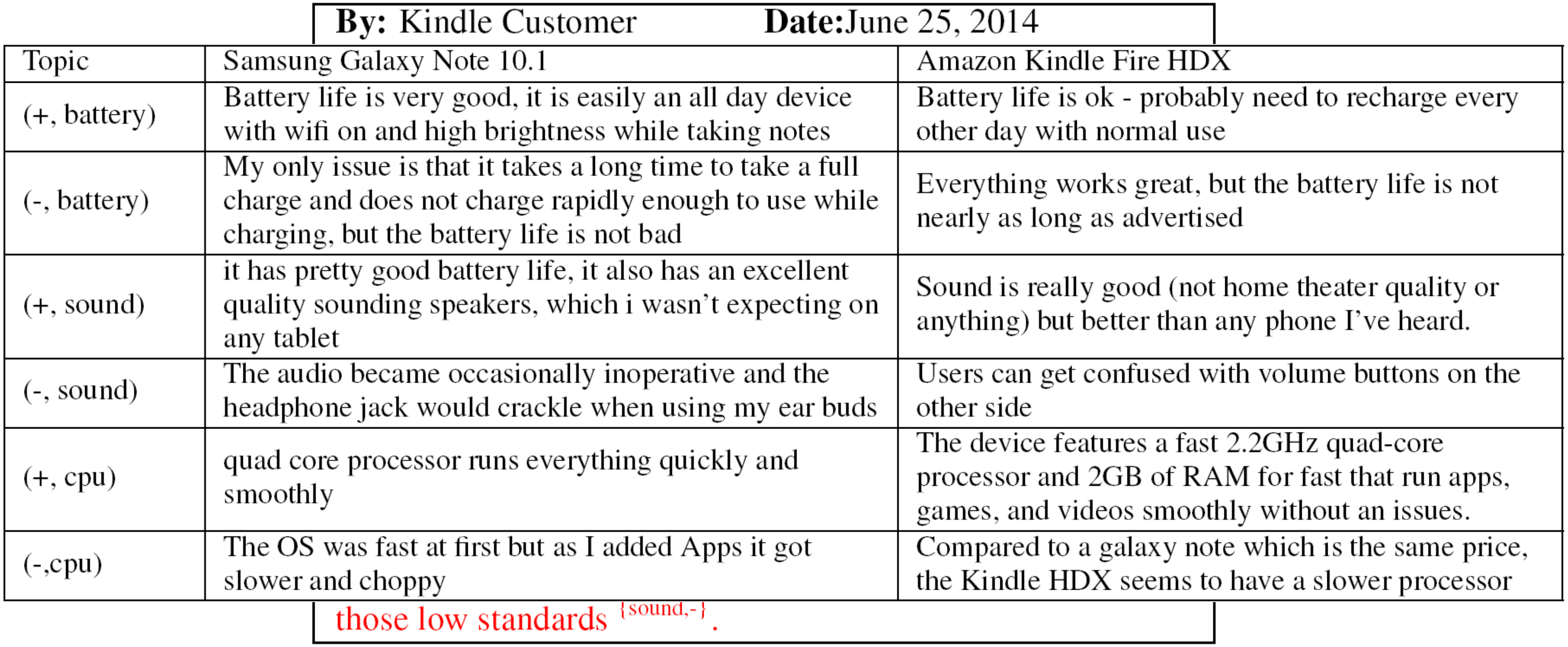
Collaborative explore-exploit
Long-term task optimization

User-oriented explanation generation

Humans: producer and consumer of big data

What Am I Doing at UVa?

- Sentiment analysis with topic modeling



By: Kindle Customer **Date:**June 25, 2014

Topic	Samsung Galaxy Note 10.1	Amazon Kindle Fire HDX
(+, battery)	Battery life is very good, it is easily an all day device with wifi on and high brightness while taking notes	Battery life is ok - probably need to recharge every other day with normal use
(-, battery)	My only issue is that it takes a long time to take a full charge and does not charge rapidly enough to use while charging, but the battery life is not bad	Everything works great, but the battery life is not nearly as long as advertised
(+, sound)	it has pretty good battery life, it also has an excellent quality sounding speakers, which i wasn't expecting on any tablet	Sound is really good (not home theater quality or anything) but better than any phone I've heard.
(-, sound)	The audio became occasionally inoperative and the headphone jack would crackle when using my ear buds	Users can get confused with volume buttons on the other side
(+, cpu)	quad core processor runs everything quickly and smoothly	The device features a fast 2.2GHz quad-core processor and 2GB of RAM for fast that run apps, games, and videos smoothly without an issues.
(-,cpu)	The OS was fast at first but as I added Apps it got slower and choppy	Compared to a galaxy note which is the same price, the Kindle HDX seems to have a slower processor

those low standards $^{\{sound,-\}}$.

What Am I Doing at UVa?

- Interactive online recommendation
  - Modeling recommendation as a two-party game

Strategy?

amazon

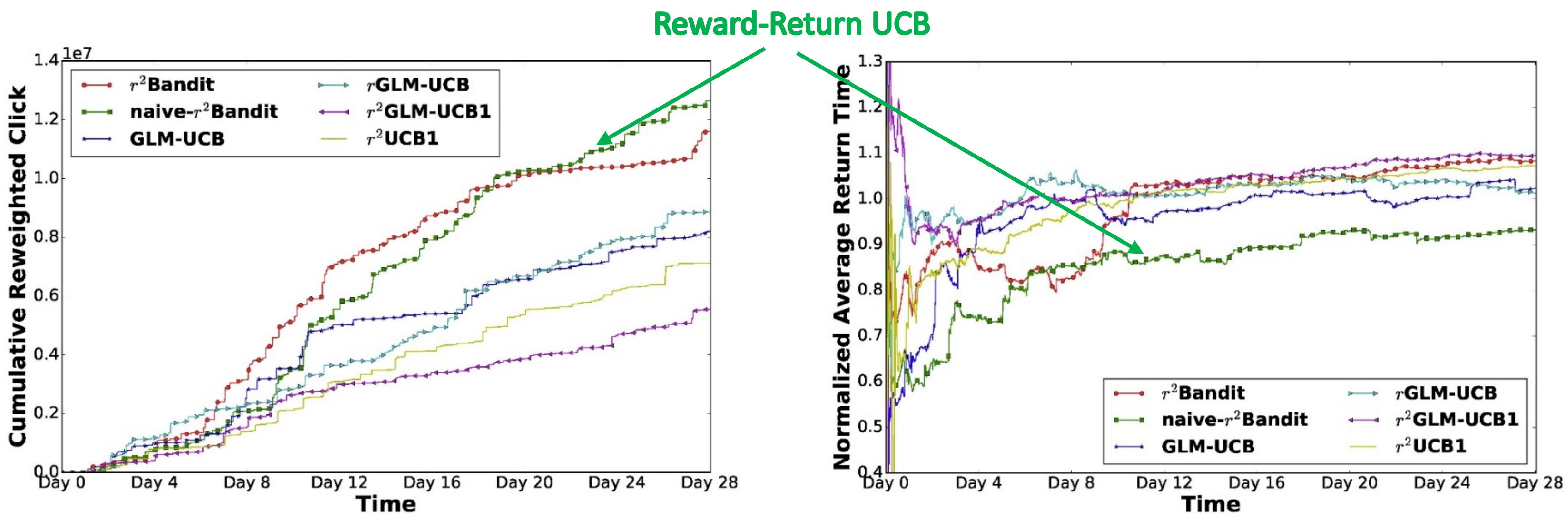
Goal:

Challenge:

1) **Unknown preference;**
2) **Feedback is acquired on the fly, and it is not free!**

What Am I Doing at UVa?

- Yahoo frontpage news recommendation



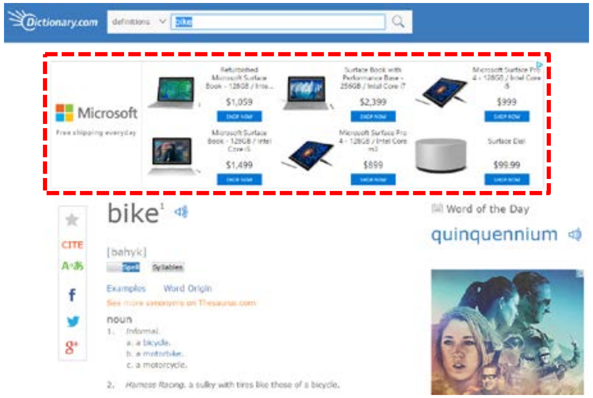
Reward-Return UCB

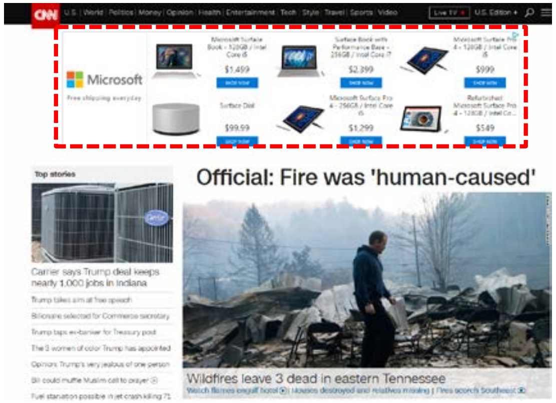
18,882 users, 188,384 articles, and 9,984,879 logged events segmented into 1,123,583 sessions.

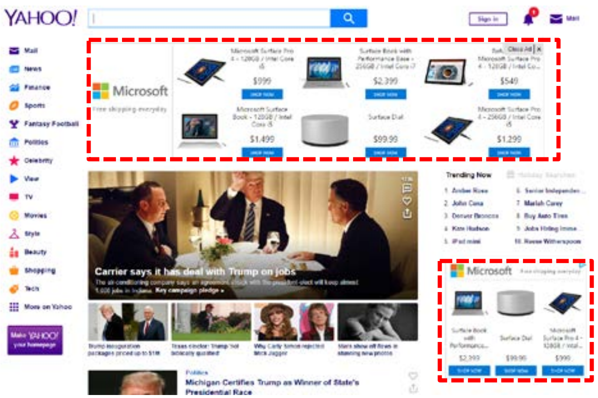
What Am I Doing at UVa?

- Personalization techniques raise serious public concerns about privacy infringement

No means for users to opt-out data collection!



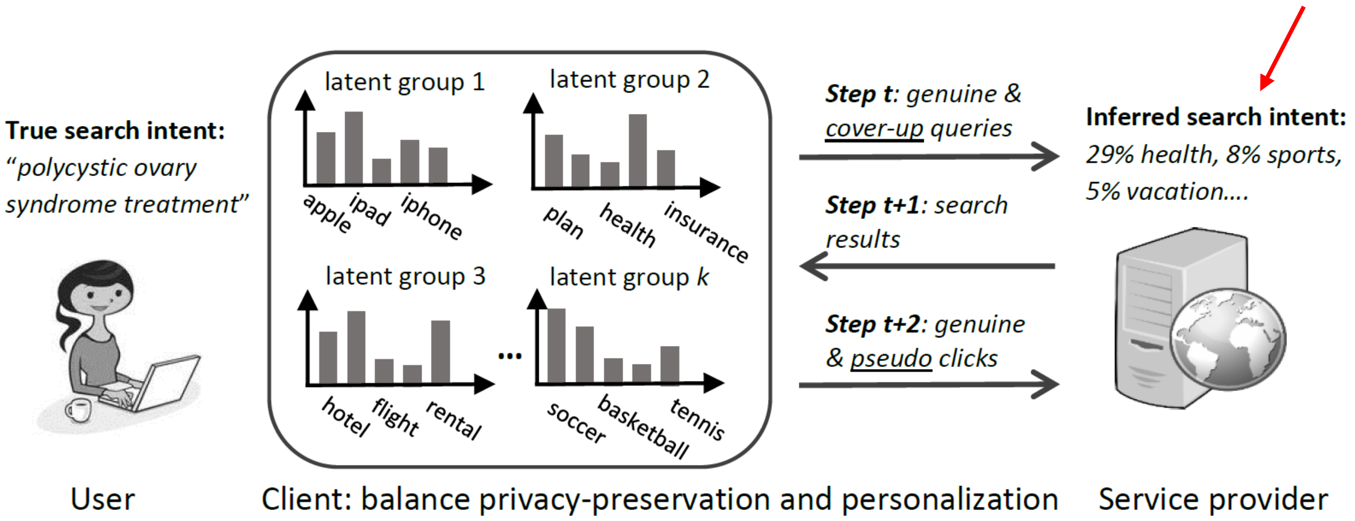

CNN
Microsoft
Top stories
Official: Fire was 'human-caused'
Wildfires leave 3 dead in eastern Tennessee


YAHOO!
Microsoft
Carrier says it has deal with Trump on jobs
Trending Now
Michigan Certifies Trump as Winner of State's Presidential Race

What Am I Doing at UVa?

- Privacy-preserving personalization

Stronger privacy guarantee than *k*-anonymity



True search intent:
"polycystic ovary syndrome treatment"

Step t: *genuine & <u>cover-up</u> queries*

Step t+1: *search results*

Step t+2: *genuine & <u>pseudo</u> clicks*

Inferred search intent:
29% health, 8% sports, 5% vacation….

User

Client: balance privacy-preservation and personalization

Service provider

What are about you?

- Why do you choose this course?
- Anything specific you want me to know?
- What type of text data do you often encounter in your projects?
- What kind of knowledge do you want to extract from it?

What is “Text Mining”?

- *“Text mining, also referred to as* ***text data mining****, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text.”* - wikipedia
- *“Another way to view text data mining is as a process of* ***exploratory*** *data analysis that leads to* ***heretofore unknown*** *information, or to answers for questions for which the answer is not currently known.”* - Hearst, 1999

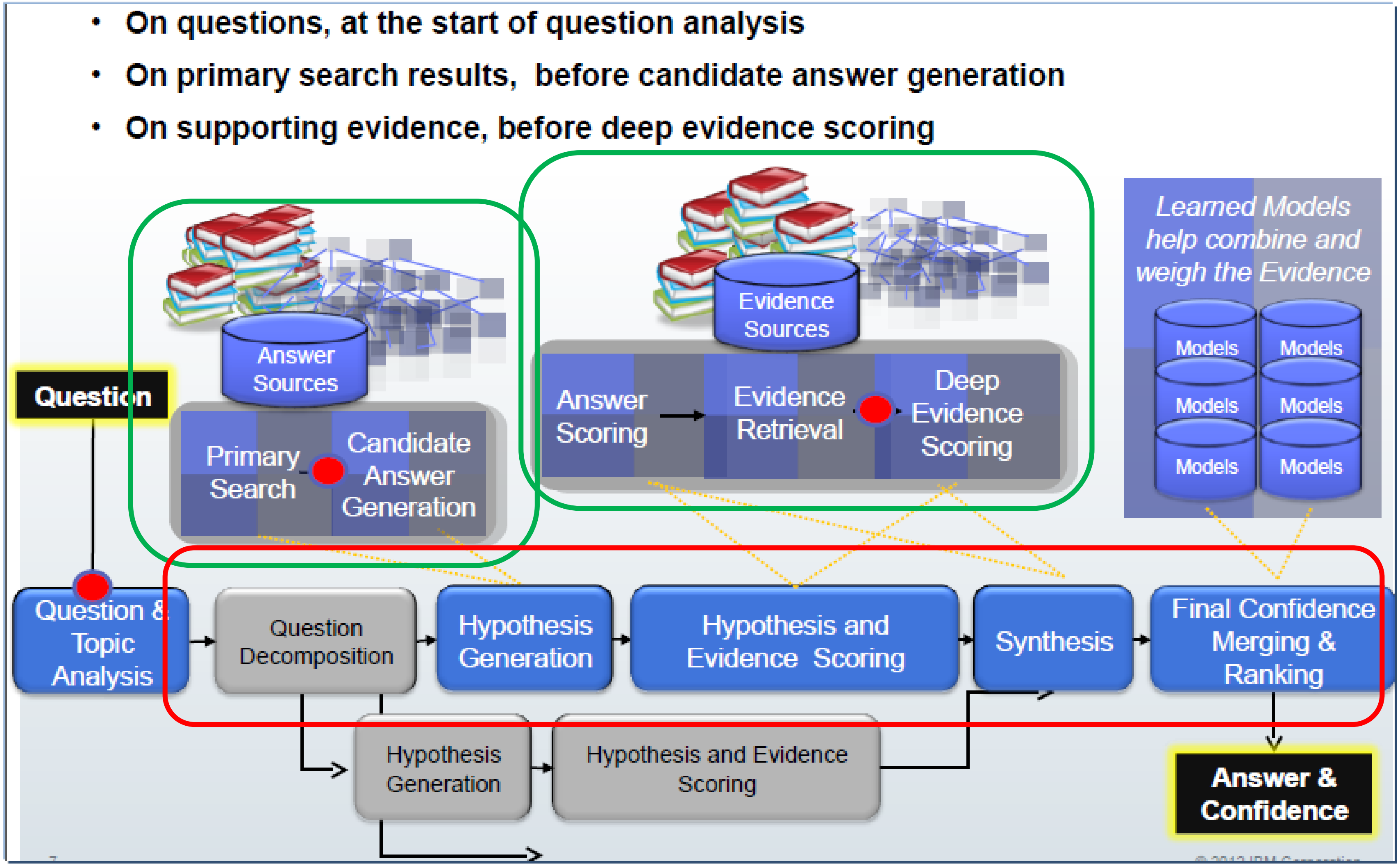
Two different definitions of mining

- Goal-oriented (effectiveness driven)
  - Any process that generates useful results that are non-obvious is called "mining".
  - Keywords: "**useful**" + "**non-obvious**"
  - Data isn't necessarily massive
- Method-oriented (efficiency driven)
  - Any process that involves extracting information from massive data is called "mining"
  - Keywords: "**massive**" + "**pattern**"
  - Patterns aren't necessarily useful

Knowledge discovery from text data

- IBM's Watson wins at Jeopardy! - 2011

An overview of Watson



- On questions, at the start of question analysis
- On primary search results, before candidate answer generation
- On supporting evidence, before deep evidence scoring

What is inside Watson?

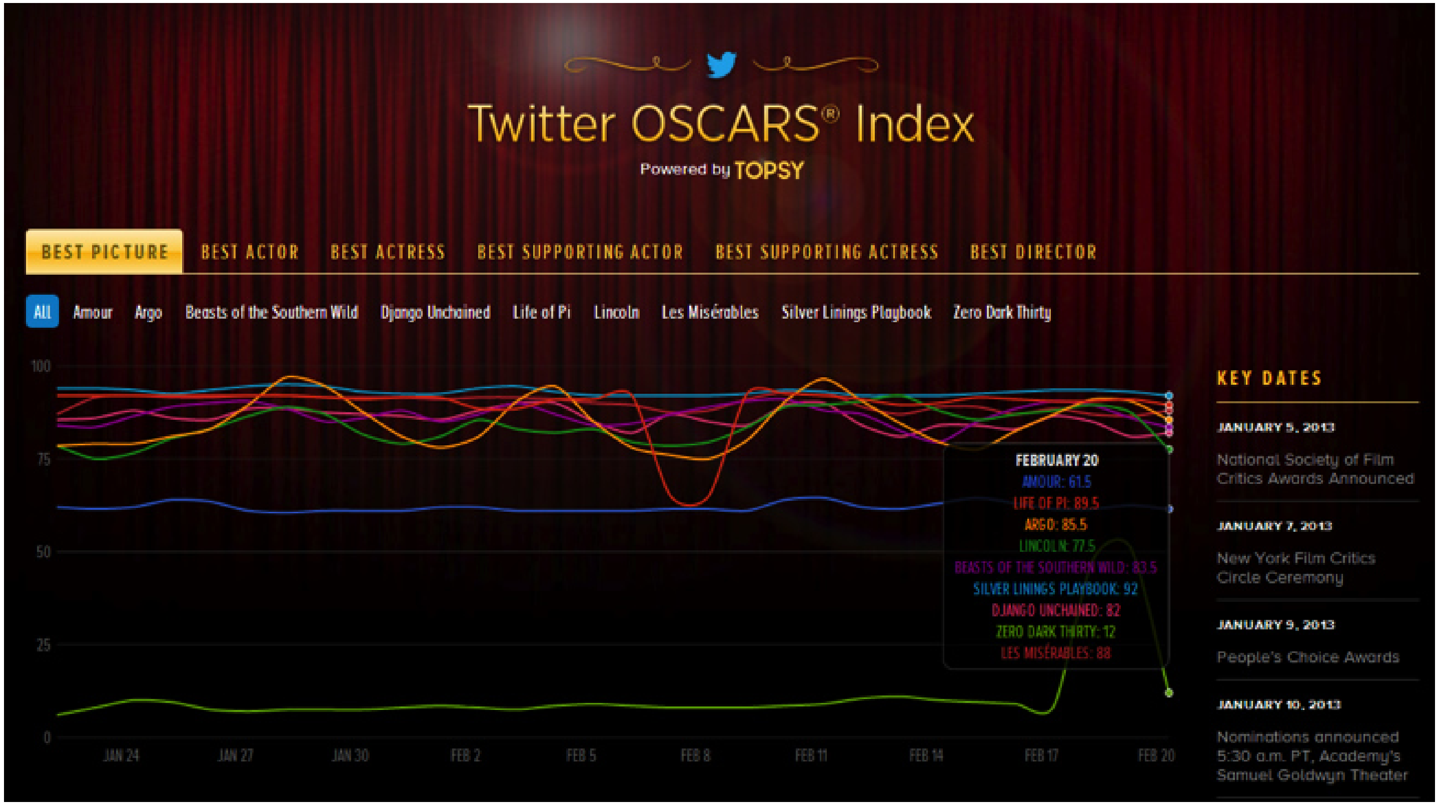
- *"Watson had access to <u>200 million pages</u> of structured and unstructured content consuming four terabytes of disk storage including the full text of Wikipedia" – PC World*
- *"The sources of information for Watson include encyclopedias, dictionaries, thesauri, newswire articles, and literary works. Watson also used databases, taxonomies, and ontologies. Specifically, DBPedia, WordNet, and Yago were used." – AI Magazine*

What is inside Watson?

- DeepQA system
  - *“Watson's main innovation was not in the creation of a new algorithm for this operation but rather its ability to **quickly** execute hundreds of proven language analysis algorithms simultaneously to find the correct answer.”* – New York Times
  - The DeepQA Research Team

Text mining around us

- Sentiment analysis



KEY DATES

JANUARY 5, 2013

National Society of Film Critics Awards Announced

JANUARY 7, 2013

New York Film Critics Circle Ceremony

JANUARY 9, 2013

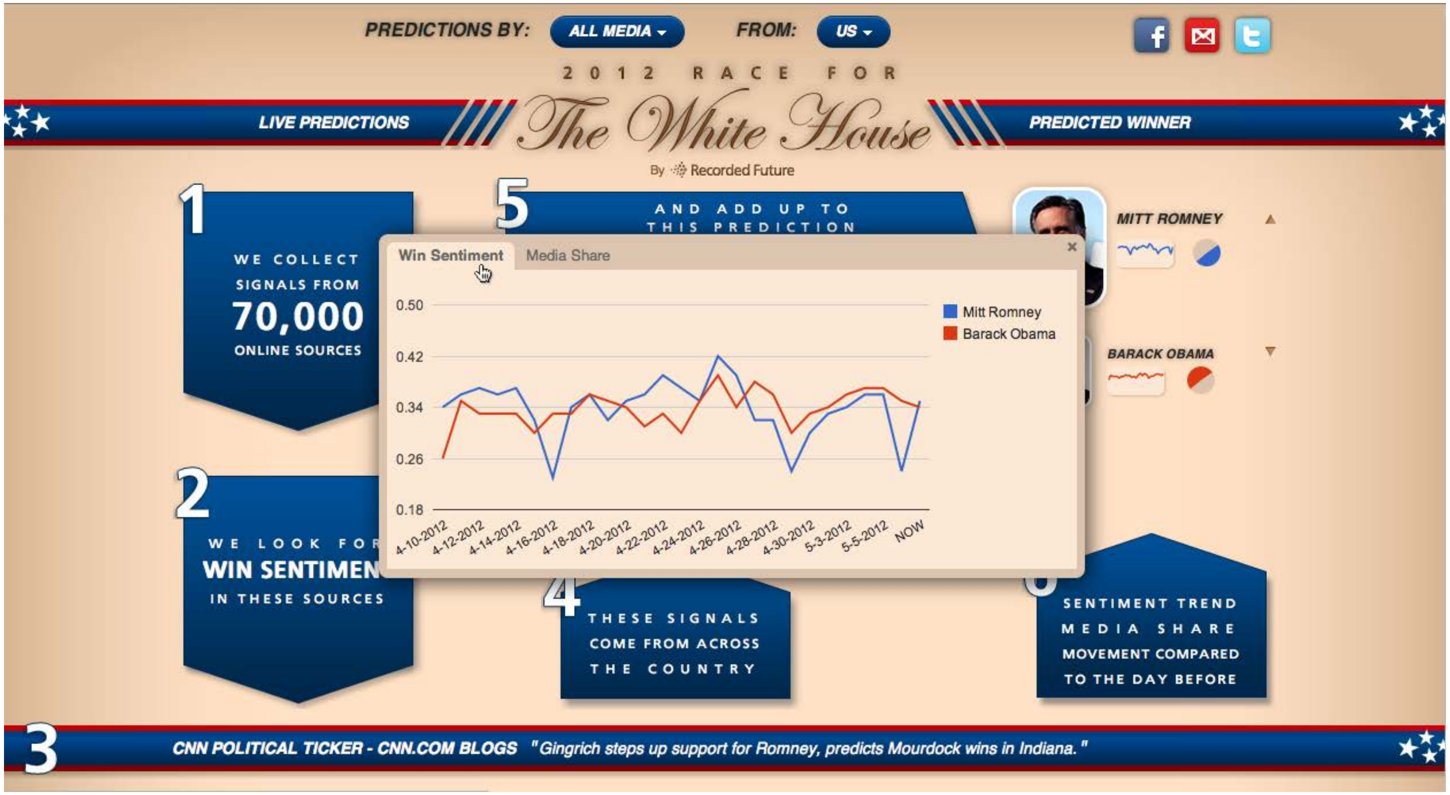
People's Choice Awards

JANUARY 10, 2013

Nominations announced 5:30 a.m. PT, Academy's Samuel Goldwyn Theater

Text mining around us

- Sentiment analysis



PREDICTIONS BY: ALL MEDIA

FROM: US

2012 RACE FOR

The White House

By Recorded Future

1 WE COLLECT SIGNALS FROM 70,000 ONLINE SOURCES

2 WE LOOK FOR WIN SENTIMEN IN THESE SOURCES

3 CNN POLITICAL TICKER - CNN.COM BLOGS "Gingrich steps up support for Romney, predicts Mourdock wins in Indiana."

4 THESE SIGNALS COME FROM ACROSS THE COUNTRY

5 AND ADD UP TO THIS PREDICTION

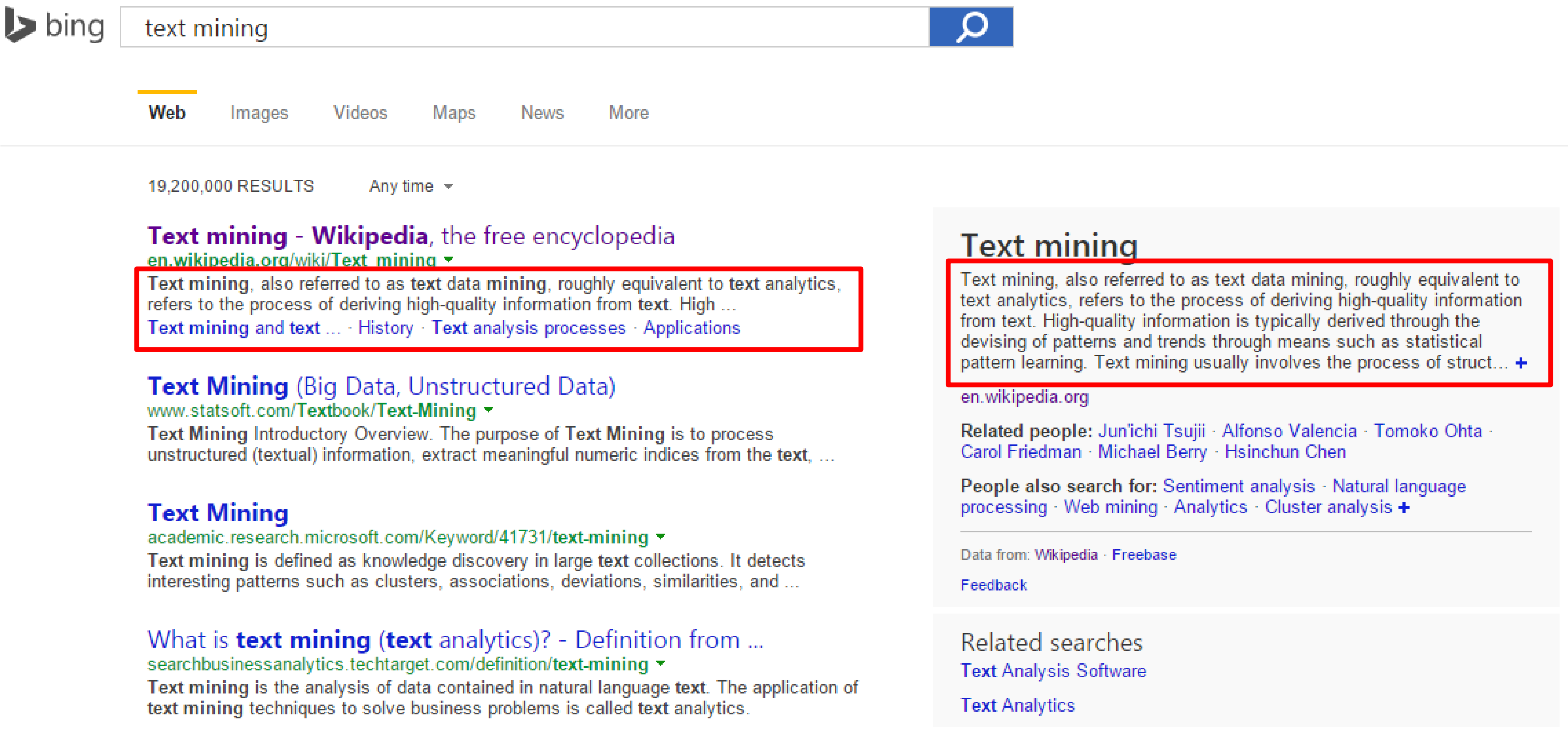
SENTIMENT TREND MEDIA SHARE MOVEMENT COMPARED TO THE DAY BEFORE

Text mining around us

- Document summarization

Text mining around us

- Document summarization



text mining

19,200,000 RESULTS

Text mining - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Text_mining**
Text mining, also referred to as **text** data **mining**, roughly equivalent to **text** analytics, refers to the process of deriving high-quality information from **text**. High …
Text mining and **text** … · History · **Text** analysis processes · Applications

Text Mining (Big Data, Unstructured Data)
www.statsoft.com/**Textbook**/**Text-Mining**
Text Mining Introductory Overview. The purpose of **Text Mining** is to process unstructured (textual) information, extract meaningful numeric indices from the **text**, …

Text Mining
academic.research.microsoft.com/Keyword/41731/**text-mining**
Text mining is defined as knowledge discovery in large **text** collections. It detects interesting patterns such as clusters, associations, deviations, similarities, and …

What is **text mining** (**text** analytics)? - Definition from …
searchbusinessanalytics.techtarget.com/definition/**text-mining**
Text mining is the analysis of data contained in natural language **text**. The application of **text mining** techniques to solve business problems is called **text** analytics.

Text mining
Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of struct… +
en.wikipedia.org

Related people: Jun'ichi Tsujii · Alfonso Valencia · Tomoko Ohta · Carol Friedman · Michael Berry · Hsinchun Chen

People also search for: Sentiment analysis · Natural language processing · Web mining · Analytics · Cluster analysis +

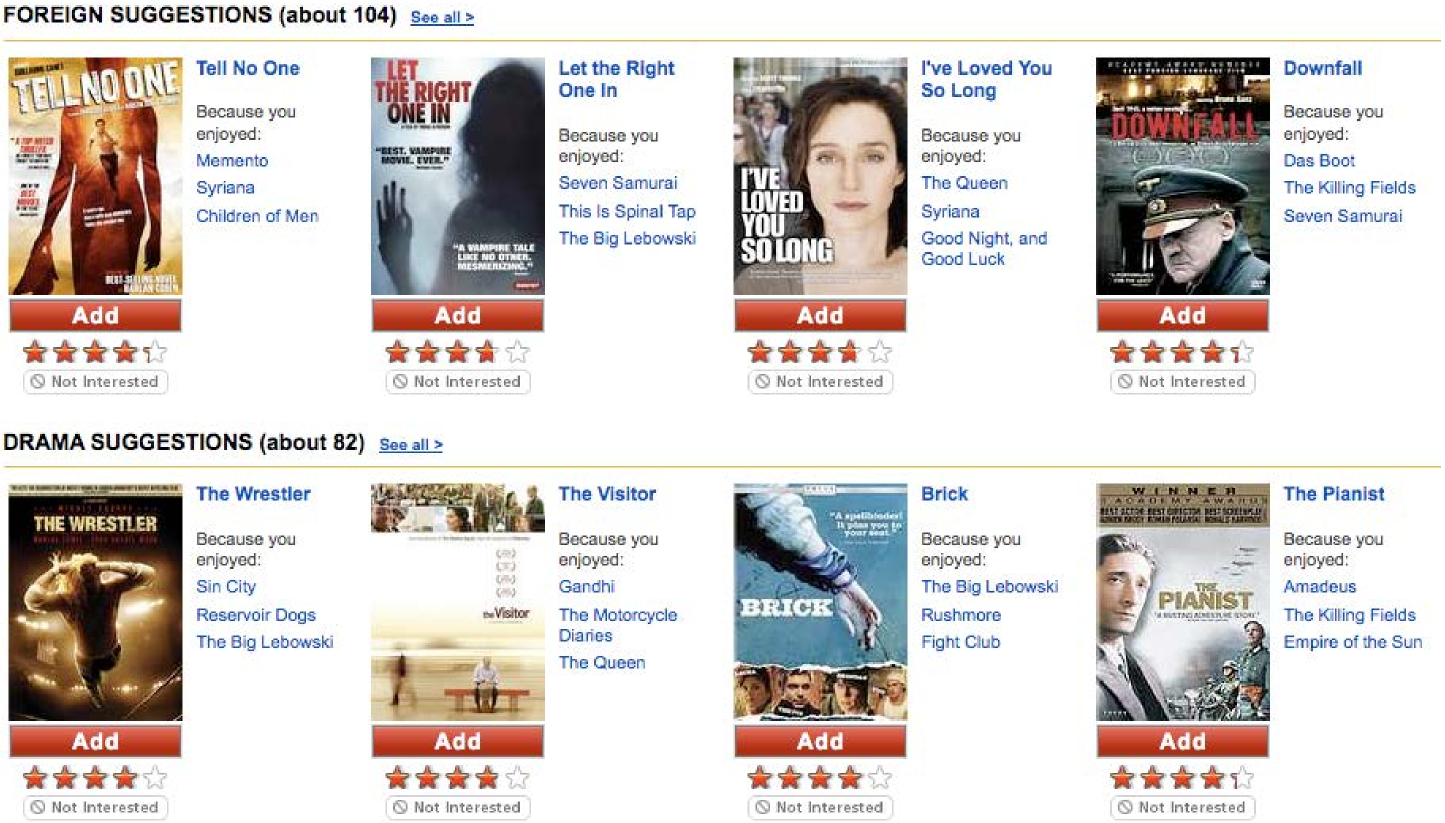
Data from: Wikipedia · Freebase

Feedback

Related searches
Text Analysis Software
Text Analytics

Text mining around us

- Movie recommendation



FOREIGN SUGGESTIONS (about 104)

Tell No One
Because you enjoyed:
Memento
Syriana
Children of Men

Let the Right One In
Because you enjoyed:
Seven Samurai
This Is Spinal Tap
The Big Lebowski

I've Loved You So Long
Because you enjoyed:
The Queen
Syriana
Good Night, and Good Luck

Downfall
Because you enjoyed:
Das Boot
The Killing Fields
Seven Samurai

DRAMA SUGGESTIONS (about 82)

The Wrestler
Because you enjoyed:
Sin City
Reservoir Dogs
The Big Lebowski

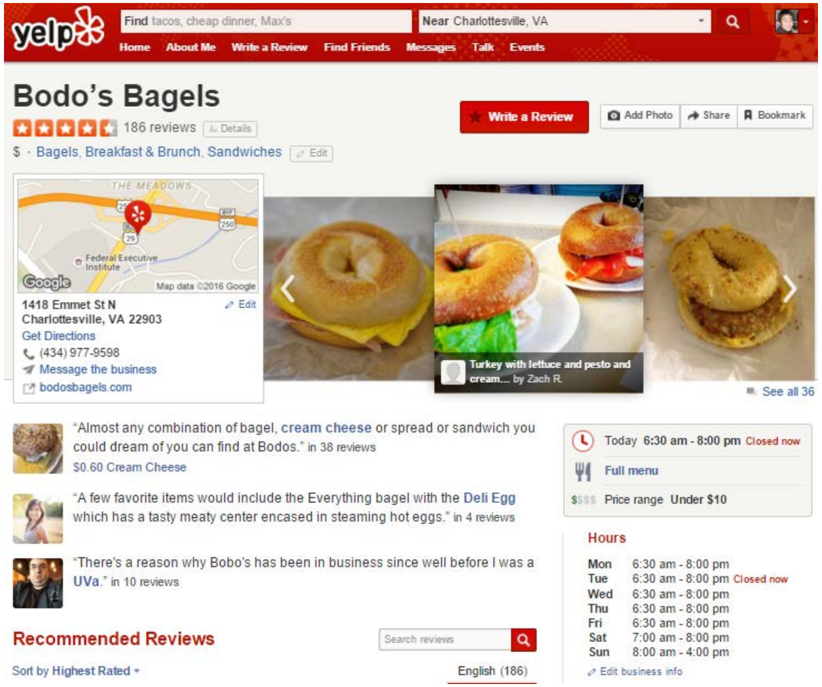
The Visitor
Because you enjoyed:
Gandhi
The Motorcycle Diaries
The Queen

Brick
Because you enjoyed:
The Big Lebowski
Rushmore
Fight Club

The Pianist
Because you enjoyed:
Amadeus
The Killing Fields
Empire of the Sun

Text mining around us

- Restaurant/hotel recommendation



Bodo's Bagels

186 reviews

$ · Bagels, Breakfast & Brunch, Sandwiches

1418 Emmet St N
Charlottesville, VA 22903
Get Directions
(434) 977-9598
Message the business
bodosbagels.com

Turkey with lettuce and pesto and cream... by Zach R.

"Almost any combination of bagel, cream cheese or spread or sandwich you could dream of you can find at Bodos." in 38 reviews
$0.60 Cream Cheese

"A few favorite items would include the Everything bagel with the Deli Egg which has a tasty meaty center encased in steaming hot eggs." in 4 reviews

"There's a reason why Bobo's has been in business since well before I was a UVa." in 10 reviews

Today 6:30 am - 8:00 pm Closed now
Full menu
$$$$ Price range Under $10

Hours

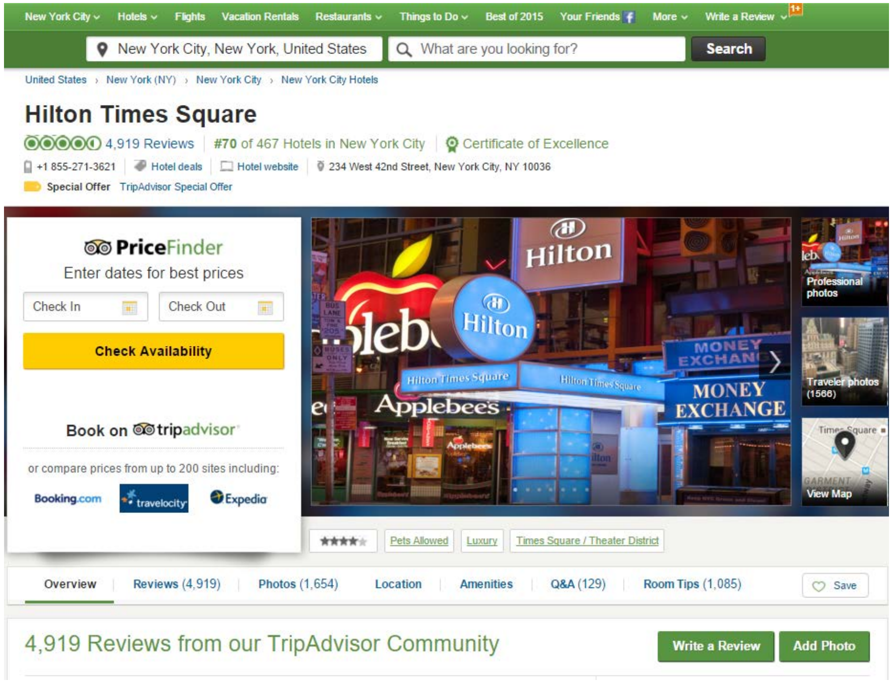
Mon	6:30 am - 8:00 pm
Tue	6:30 am - 8:00 pm Closed now
Wed	6:30 am - 8:00 pm
Thu	6:30 am - 8:00 pm
Fri	6:30 am - 8:00 pm
Sat	7:00 am - 8:00 pm
Sun	8:00 am - 4:00 pm

Edit business info

Recommended Reviews

Sort by Highest Rated

English (186)



United States › New York (NY) › New York City › New York City Hotels

Hilton Times Square

4,919 Reviews | #70 of 467 Hotels in New York City | Certificate of Excellence

+1 855-271-3621 | Hotel deals | Hotel website | 234 West 42nd Street, New York City, NY 10036

Special Offer TripAdvisor Special Offer

PriceFinder
Enter dates for best prices

Book on tripadvisor

or compare prices from up to 200 sites including:

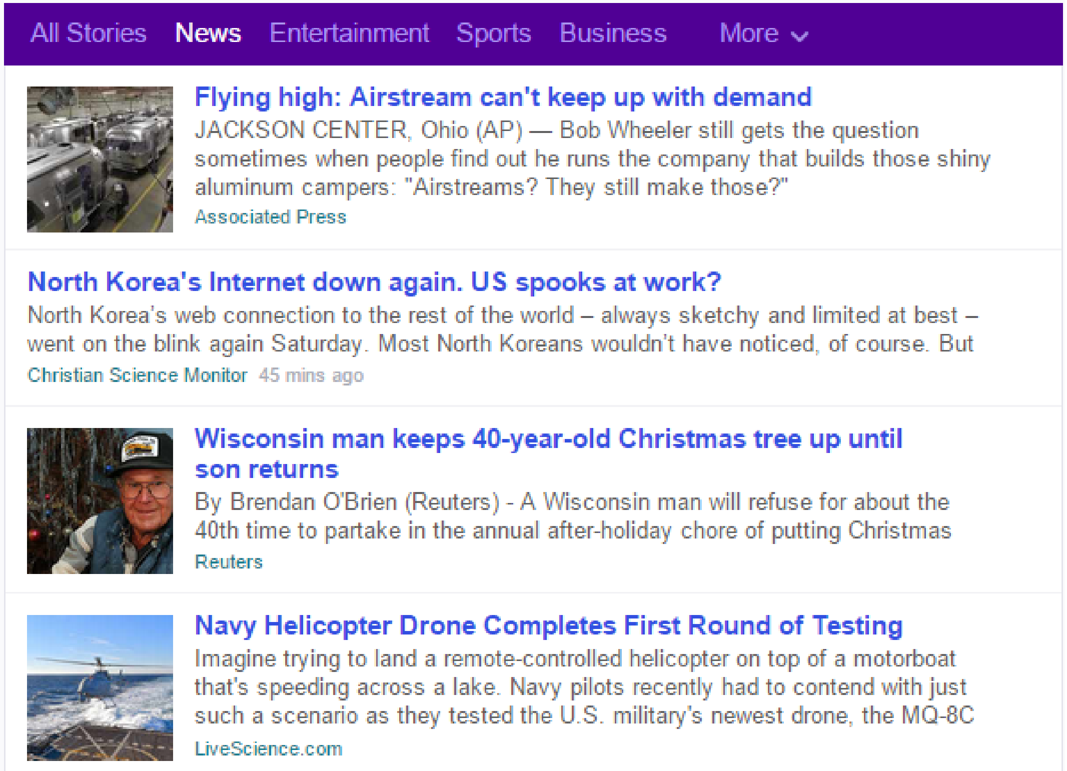
★★★★☆ | Pets Allowed | Luxury | Times Square / Theater District

Overview | Reviews (4,919) | Photos (1,654) | Location | Amenities | Q&A (129) | Room Tips (1,085)

4,919 Reviews from our TripAdvisor Community

Text mining around us

- News recommendation



Flying high: Airstream can't keep up with demand
JACKSON CENTER, Ohio (AP) — Bob Wheeler still gets the question sometimes when people find out he runs the company that builds those shiny aluminum campers: "Airstreams? They still make those?"
Associated Press

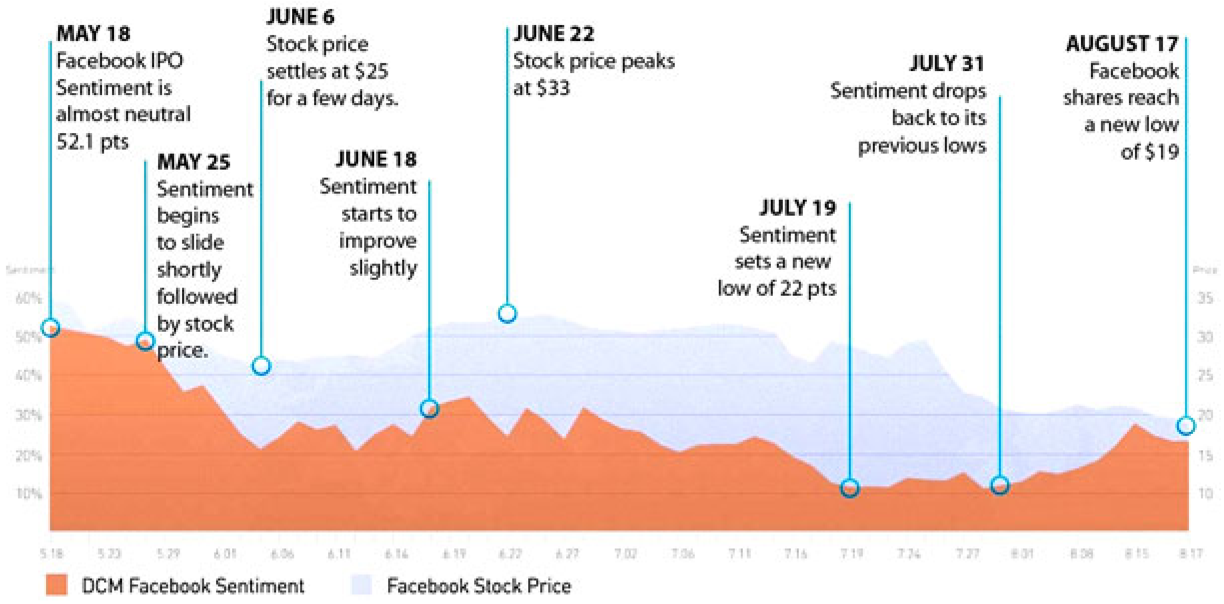
North Korea's Internet down again. US spooks at work?
North Korea's web connection to the rest of the world – always sketchy and limited at best – went on the blink again Saturday. Most North Koreans wouldn't have noticed, of course. But
Christian Science Monitor 45 mins ago

Wisconsin man keeps 40-year-old Christmas tree up until son returns
By Brendan O'Brien (Reuters) - A Wisconsin man will refuse for about the 40th time to partake in the annual after-holiday chore of putting Christmas
Reuters

Navy Helicopter Drone Completes First Round of Testing
Imagine trying to land a remote-controlled helicopter on top of a motorboat that's speeding across a lake. Navy pilots recently had to contend with just such a scenario as they tested the U.S. military's newest drone, the MQ-8C
LiveScience.com

Text mining around us

- Text analytics in financial services



MAY 18
Facebook IPO
Sentiment is
almost neutral
52.1 pts

MAY 25
Sentiment
begins
to slide
shortly
followed
by stock
price.

JUNE 6
Stock price
settles at $25
for a few days.

JUNE 18
Sentiment
starts to
improve
slightly

JUNE 22
Stock price peaks
at $33

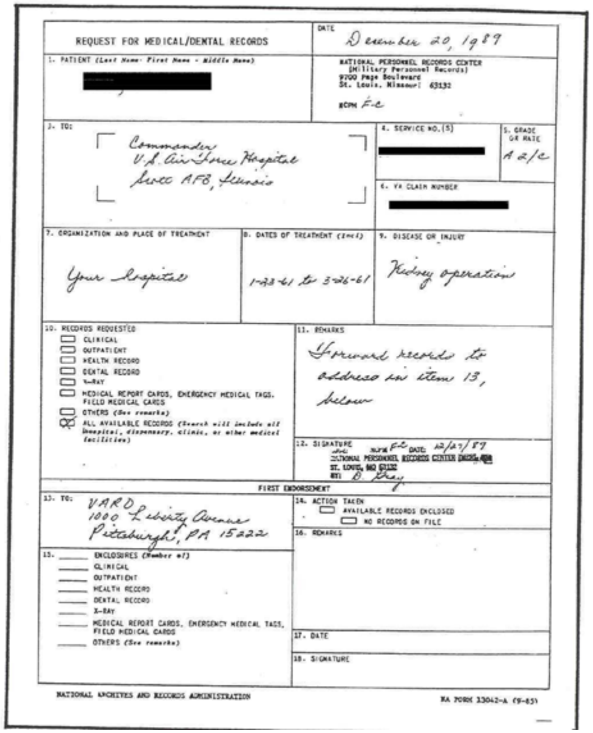
JULY 19
Sentiment
sets a new
low of 22 pts

JULY 31
Sentiment drops
back to its
previous lows

AUGUST 17
Facebook
shares reach
a new low
of $19

Text mining around us

- Text analytics in healthcare



REQUEST FOR MEDICAL/DENTAL RECORDS

DATE December 20, 1989

1. PATIENT (Last Name - First Name - Middle Name)

[redacted]

NATIONAL PERSONNEL RECORDS CENTER
(Military Personnel Records)
9700 Page Boulevard
St. Louis, Missouri 63132

NCPM F-C

3. TO:

Commander
U.S. Air Force Hospital
Scott AFB, Illinois

4. SERVICE NO.(S)

[redacted]

5. GRADE OR RATE

A 2/c

6. VA CLAIM NUMBER

[redacted]

7. ORGANIZATION AND PLACE OF TREATMENT

Your hospital

8. DATES OF TREATMENT (Incl)

1-23-61 to 3-26-61

9. DISEASE OR INJURY

Kidney operation

10. RECORDS REQUESTED

□ CLINICAL
□ OUTPATIENT
□ HEALTH RECORD
□ DENTAL RECORD
□ X-RAY
□ MEDICAL REPORT CARDS, EMERGENCY MEDICAL TAGS, FIELD MEDICAL CARDS
□ OTHERS (See remarks)
☒ ALL AVAILABLE RECORDS (Search will include all hospital, dispensary, clinic, or other medical facilities)

11. REMARKS

Forward records to address in item 13, below

12. SIGNATURE

NCPM F-C DATE: 12/27/89
NATIONAL PERSONNEL RECORDS CENTER (NCPM)
ST. LOUIS, MO 63132
BY: D. Gray

FIRST ENDORSEMENT

13. TO:

VARO
1000 Liberty Avenue
Pittsburgh, PA 15222

14. ACTION TAKEN

□ AVAILABLE RECORDS ENCLOSED
□ NO RECORDS ON FILE

15. ______ ENCLOSURES (Number of)

______ CLINICAL
______ OUTPATIENT
______ HEALTH RECORD
______ DENTAL RECORD
______ X-RAY
______ MEDICAL REPORT CARDS, EMERGENCY MEDICAL TAGS, FIELD MEDICAL CARDS
______ OTHERS (See remarks)

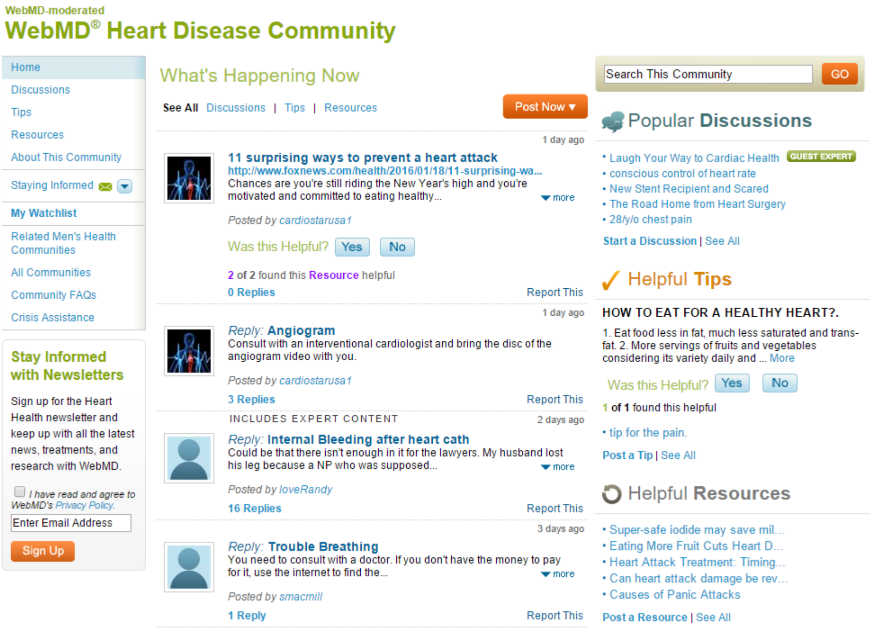
16. REMARKS

17. DATE

18. SIGNATURE

NATIONAL ARCHIVES AND RECORDS ADMINISTRATION

NA FORM 13042-A (9-85)



WebMD-moderated

WebMD® Heart Disease Community

Stay Informed with Newsletters

Sign up for the Heart Health newsletter and keep up with all the latest news, treatments, and research with WebMD.

□ I have read and agree to WebMD's Privacy Policy.

What's Happening Now

1 day ago

11 surprising ways to prevent a heart attack
http://www.foxnews.com/health/2016/01/18/11-surprising-wa...
Chances are you're still riding the New Year's high and you're motivated and committed to eating healthy... ▾ more

Posted by cardiostarusa1

Was this Helpful? Yes No

2 of 2 found this Resource helpful

0 Replies Report This

1 day ago

Reply: Angiogram
Consult with an interventional cardiologist and bring the disc of the angiogram video with you.

Posted by cardiostarusa1

3 Replies Report This

INCLUDES EXPERT CONTENT 2 days ago

Reply: Internal Bleeding after heart cath
Could be that there isn't enough in it for the lawyers. My husband lost his leg because a NP who was supposed... ▾ more

Posted by loveRandy

16 Replies Report This

3 days ago

Reply: Trouble Breathing
You need to consult with a doctor. If you don't have the money to pay for it, use the internet to find the... ▾ more

Posted by smacmill

1 Reply Report This

Popular Discussions

- Laugh Your Way to Cardiac Health GUEST EXPERT
- conscious control of heart rate
- New Stent Recipient and Scared
- The Road Home from Heart Surgery
- 28/y/o chest pain

Start a Discussion | See All

✓ Helpful Tips

HOW TO EAT FOR A HEALTHY HEART?.

1. Eat food less in fat, much less saturated and trans-fat. 2. More servings of fruits and vegetables considering its variety daily and ... More

Was this Helpful? Yes No

1 of 1 found this helpful

- tip for the pain.

Post a Tip | See All

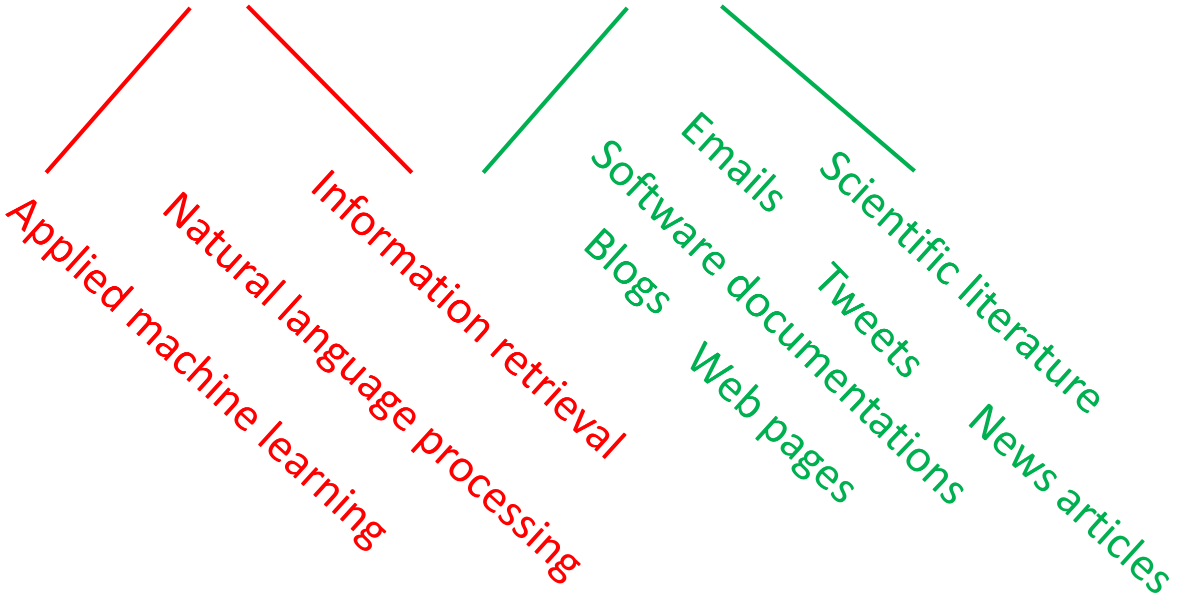
Helpful Resources

- Super-safe iodide may save mil...
- Eating More Fruit Cuts Heart D...
- Heart Attack Treatment: Timing...
- Can heart attack damage be rev...
- Causes of Panic Attacks

Post a Resource | See All

How to perform text mining?

- As computer scientists, we view it as
  - Text Mining = Data Mining + Text Data



Applied machine learning

Natural language processing

Information retrieval

Software documentations

Emails

Blogs

Scientific literature

Tweets

Web pages

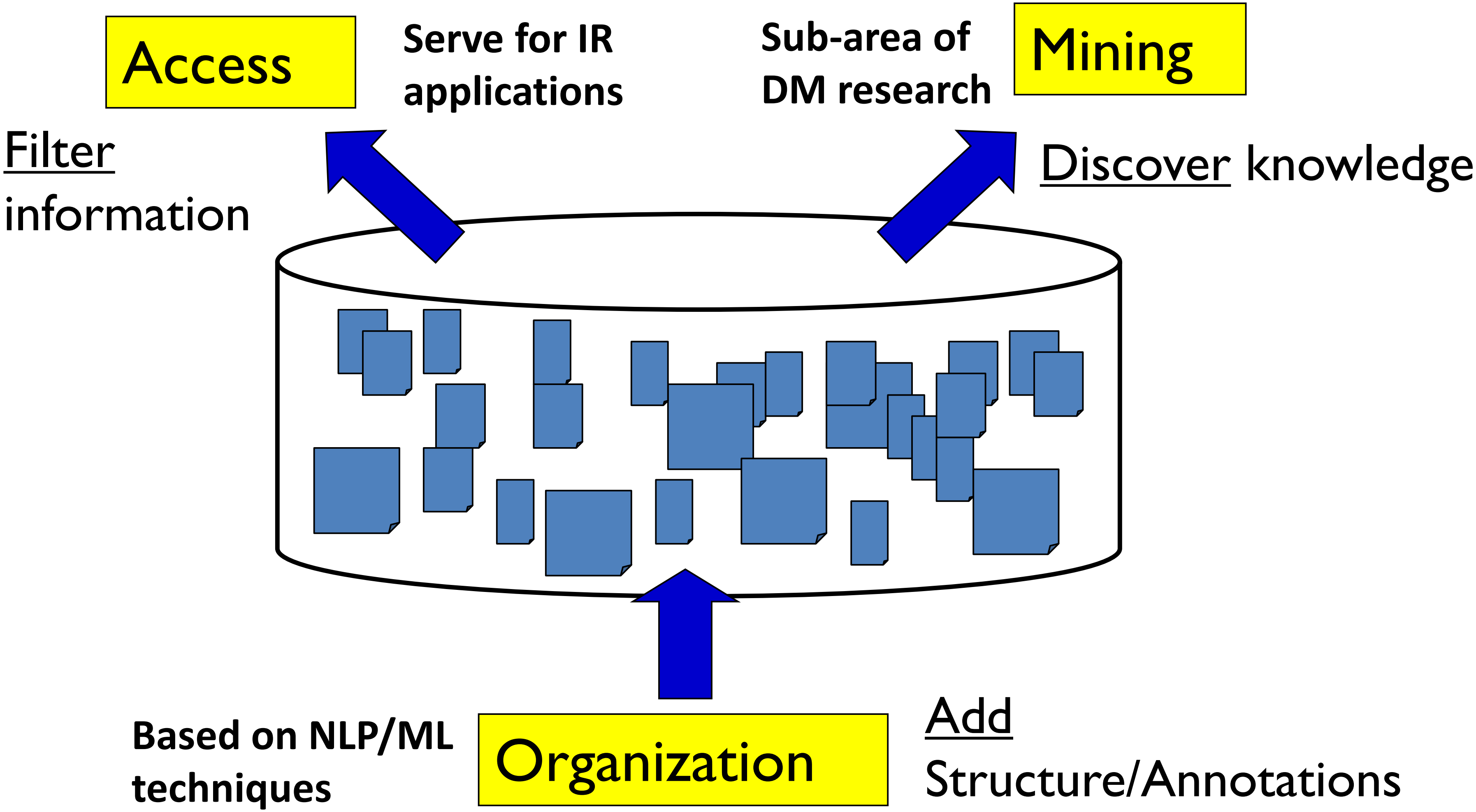
News articles

Text mining v.s. NLP, IR, DM…

- How does it relate to data mining in general?
- How does it relate to computational linguistics?
- How does it relate to information retrieval?

	Finding Patterns	Finding “Nuggets”	
		Novel	Non-Novel
Non-textual data	General data-mining	Exploratory 'ysis	Database queries
Textual data	Comp Ling	Text Mining	Information retrieval

Text mining in general



Access

Serve for IR applications

Filter information

Sub-area of DM research

Mining

Discover knowledge

Based on NLP/ML techniques

Organization

Add Structure/Annotations

Challenges in text mining

- Data collection is "free text"
  - Data is not well-organized
    - Semi-structured or unstructured
  - Natural language text contains ambiguities on many levels
    - Lexical, syntactic, semantic, and pragmatic
  - Learning techniques for processing text typically need annotated training examples
    - Expensive to acquire at scale
- What to mine?

Text mining problems we will solve

- Lexical semantics and word senses
  - Identifying which sense of a word (i.e. meaning) is used in a sentence, when the word has multiple meanings

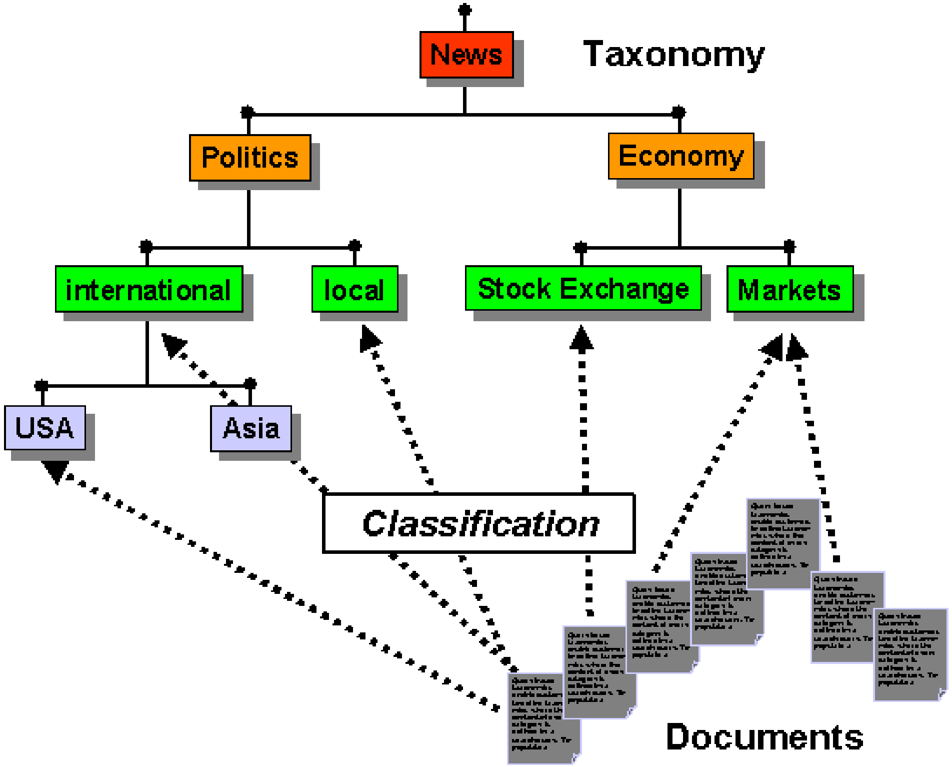
Bass: fish

???

Bass: instrument

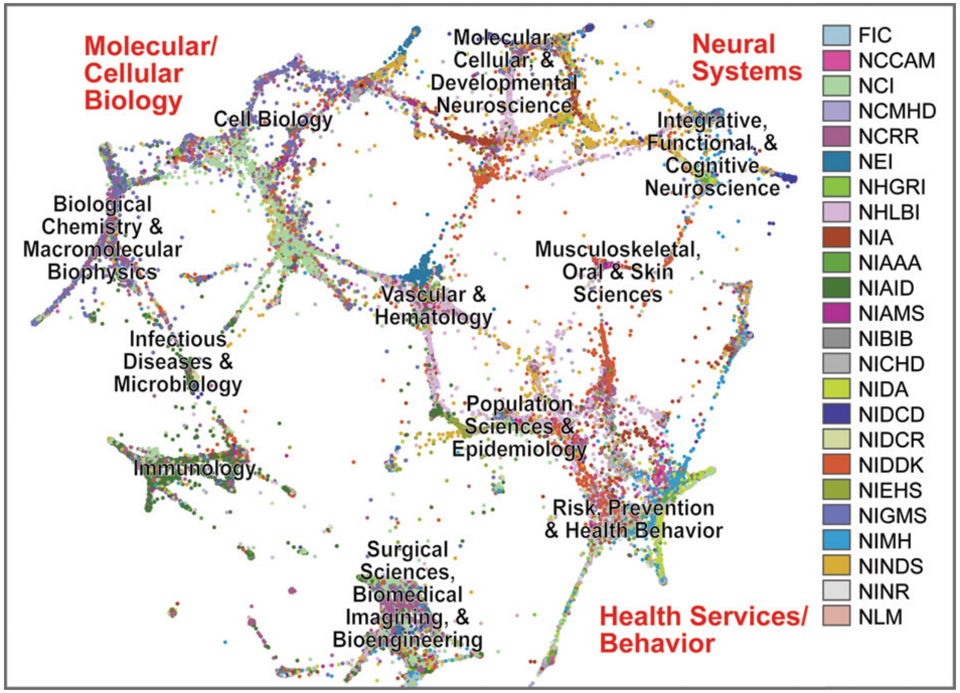
Text mining problems we will solve

- Document categorization
  - Adding structures to the text corpus



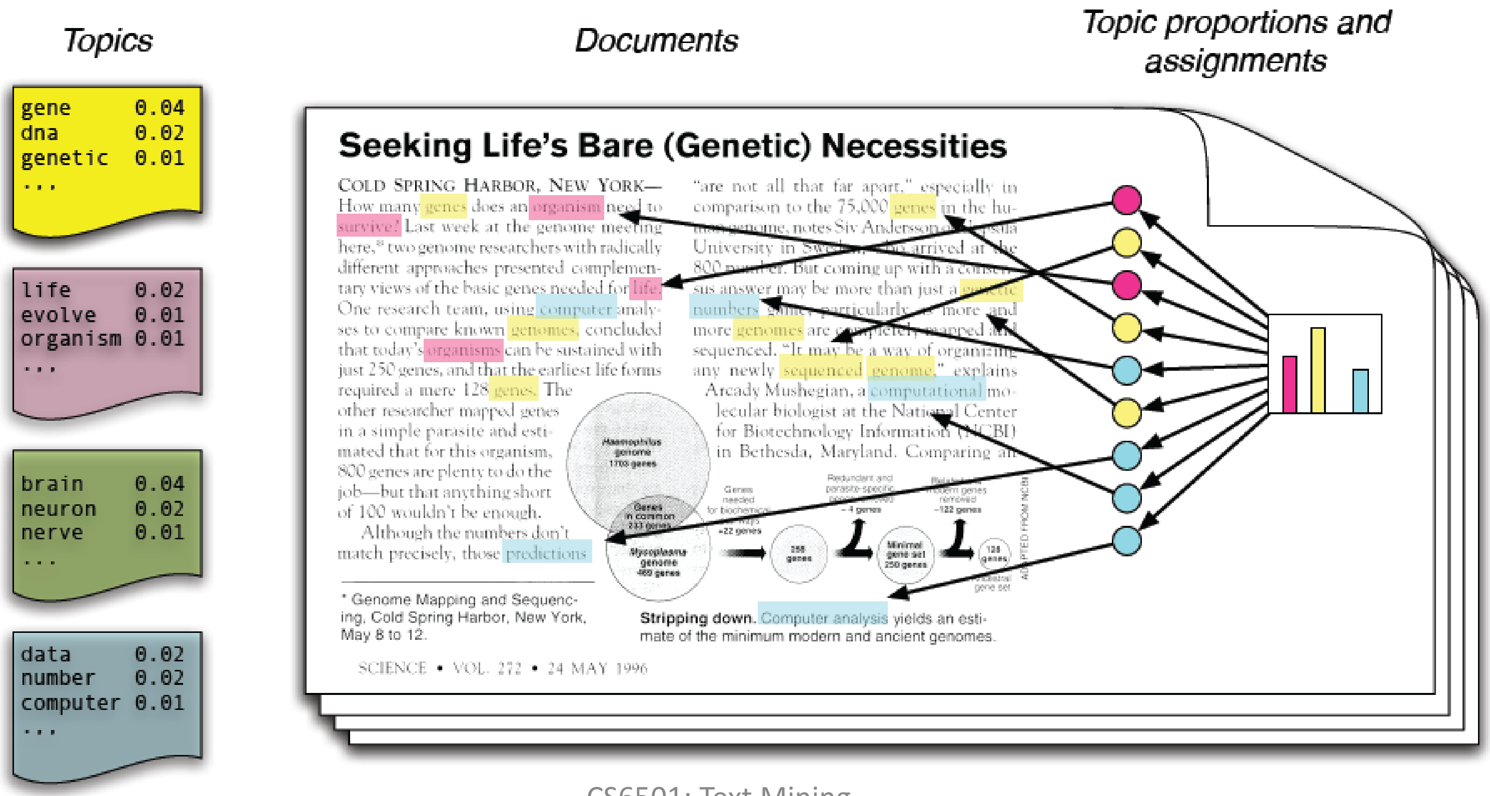
Text mining problems we will solve

- Text clustering
  - Identifying structures in the text corpus



Text mining problems we will solve

- Topic modeling
  - Identifying structures in the text corpus



Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 468 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related to modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

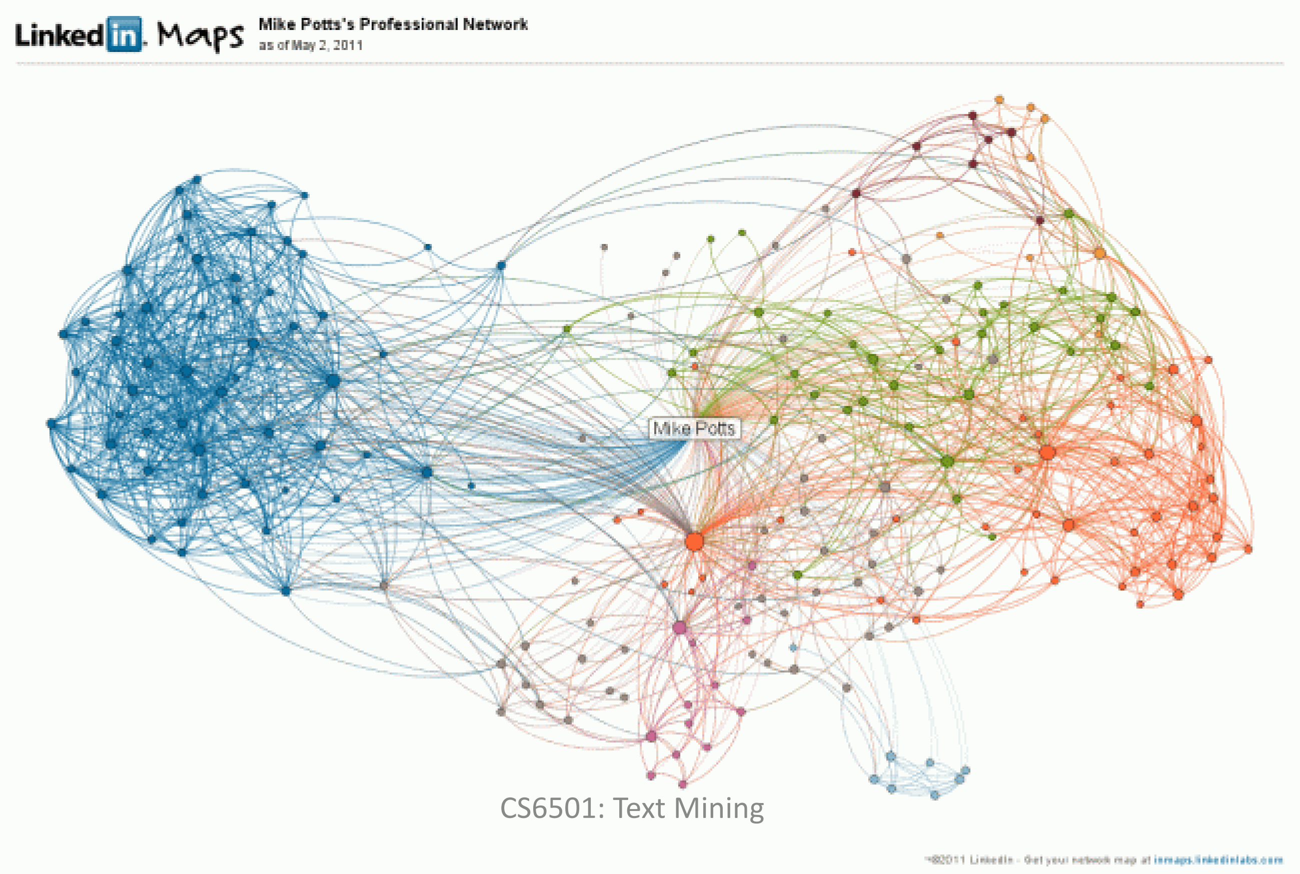
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Text mining problems we will solve

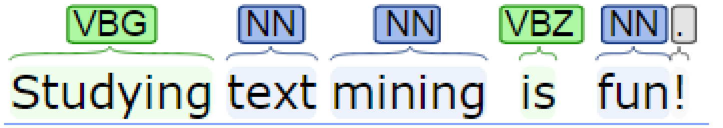
- Social media and social network analysis
  - Exploring additional structure in the text corpus



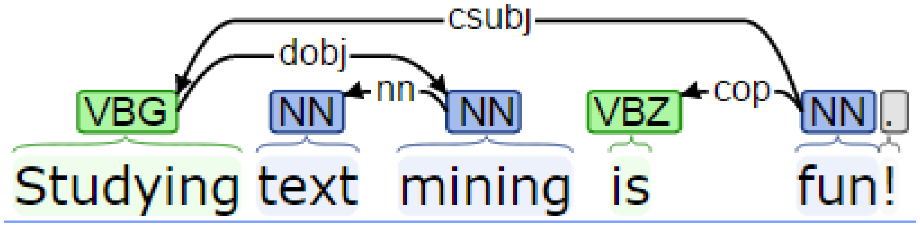
©2011 LinkedIn - Get your network map at inmaps.linkedinlabs.com

We will also briefly cover

- Natural language processing pipeline
  - Tokenization
    - “Studying text mining is fun!” -> “studying” + “text” + “mining” + “is” + “fun” + “!”
  - Part-of-speech tagging
    - “Studying text mining is fun!” ->



  - Dependency parsing
    - “Studying text mining is fun!” ->



We will also briefly cover

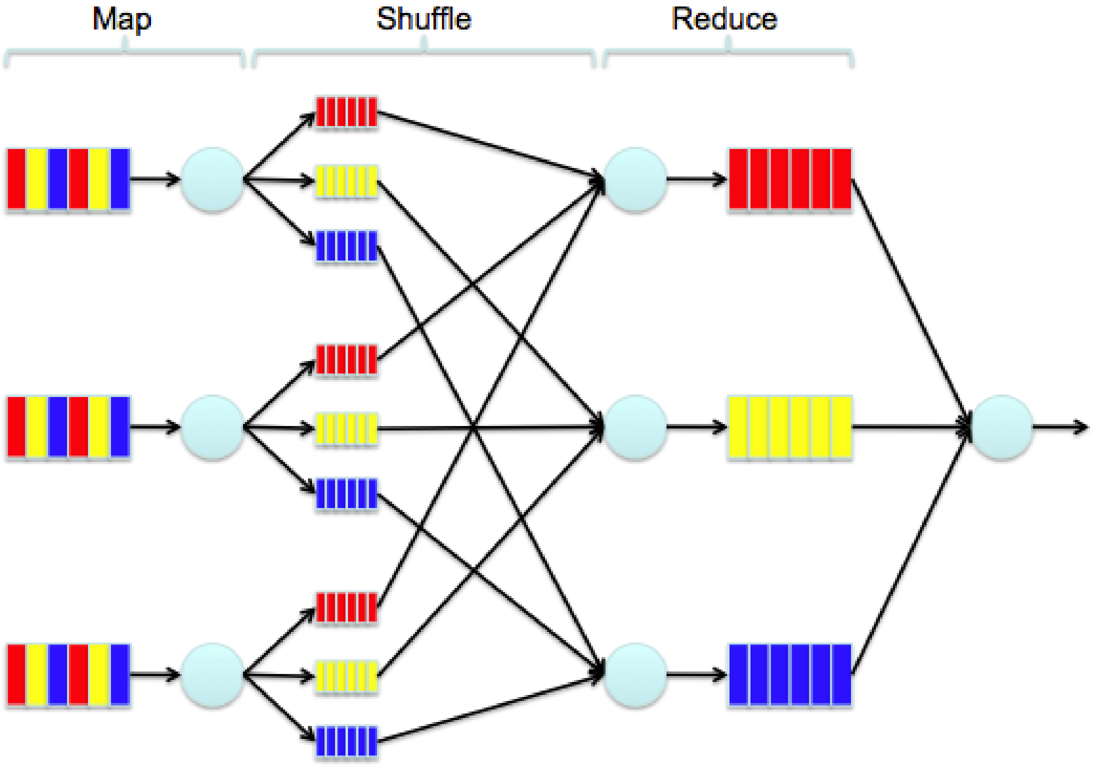
- Machine learning techniques
  - Supervised methods
    - Naïve Bayes, k Nearest Neighbors, Logistic Regression
  - Unsupervised methods
    - K-Means, hierarchical clustering, topic models
  - Semi-supervised methods
    - Expectation Maximization

Text mining in the era of Big Data

- Huge in size
  - Google processes 5.13B queries/day (2013)
  - Twitter receives 340M tweets/day (2012)
  - Facebook has 2.5 PB of user data + 15 TB/day (4/2009)
  - eBay has 6.5 PB of user data + 50 TB/da
- 80% data is unstructured (IBM, 2010)

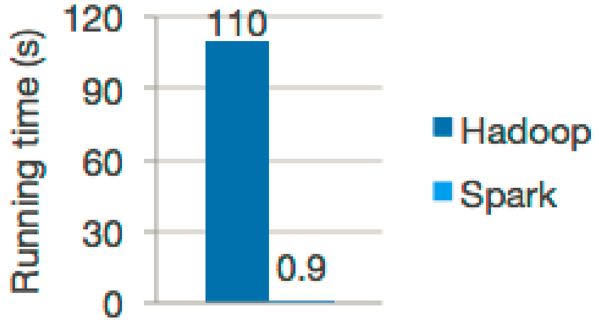
Scalability is crucial

- Large scale text processing techniques
  - MapReduce framework



State-of-the-art solutions

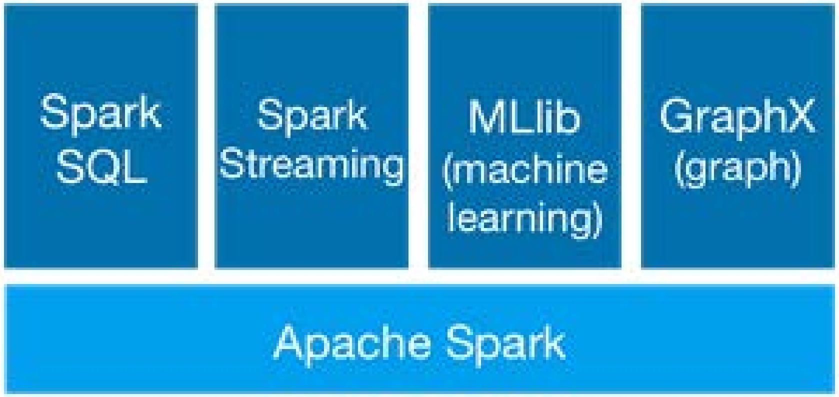
- Apache Spark (spark.apache.org)
  - In-memory MapReduce
    - Specialized for machine learning algorithms
  - Speed
    - 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.



Logistic regression in Hadoop and Spark

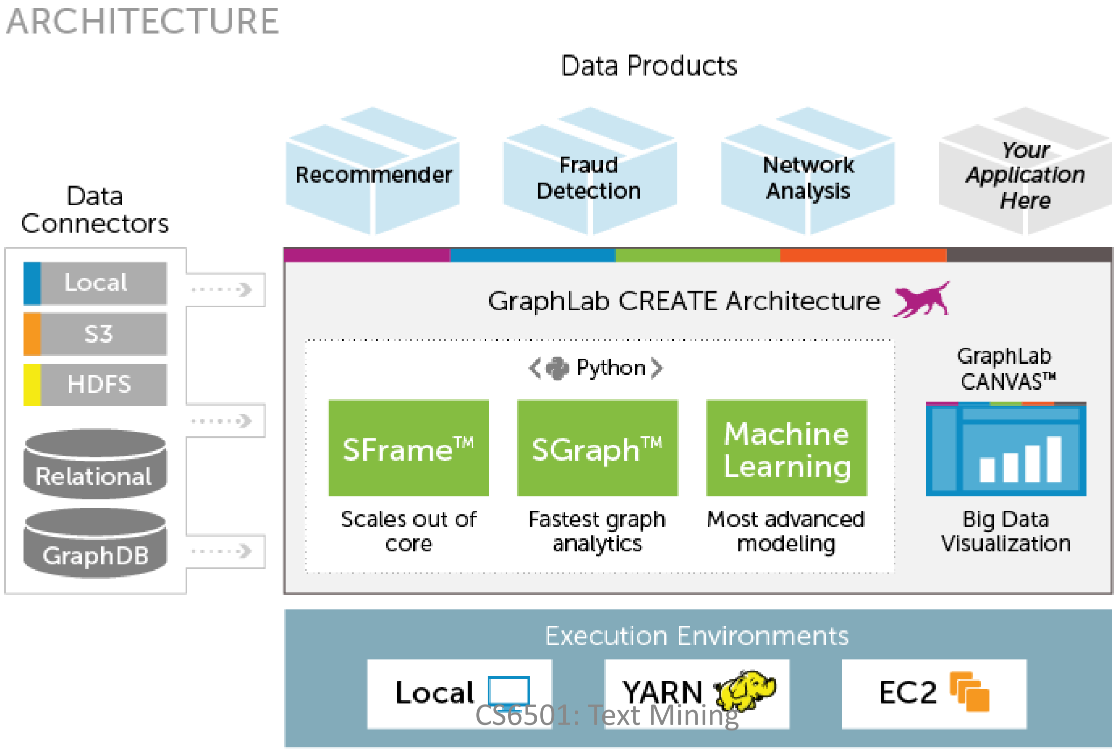
State-of-the-art solutions

- Apache Spark (spark.apache.org)
  - In-memory MapReduce
    - Specialized for machine learning algorithms
  - Generality
    - Combine SQL, streaming, and complex analytics



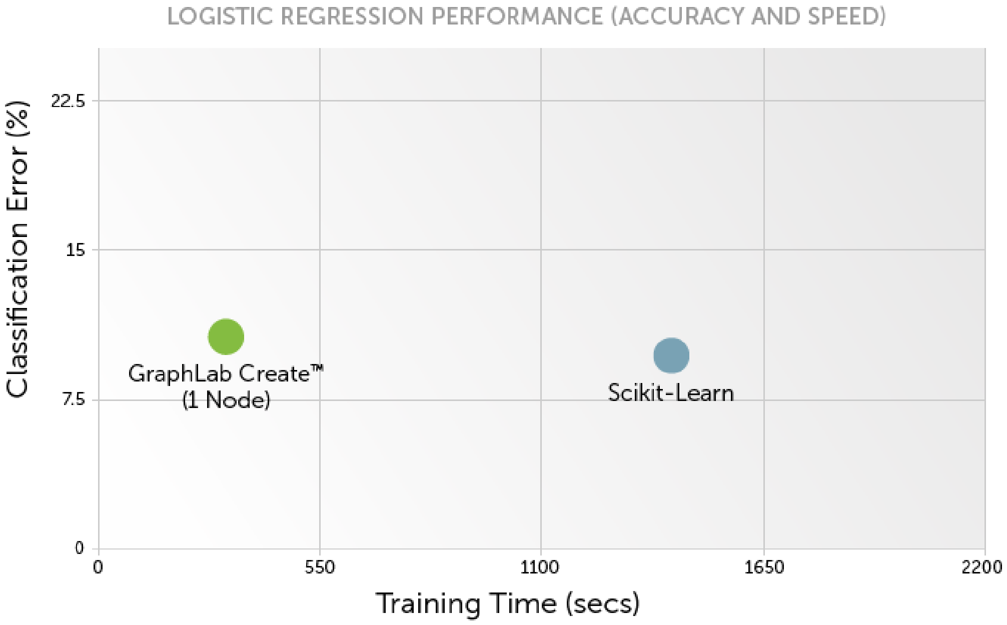
State-of-the-art solutions

- GraphLab (graphlab.com)
  - Graph-based, high performance, distributed computation framework

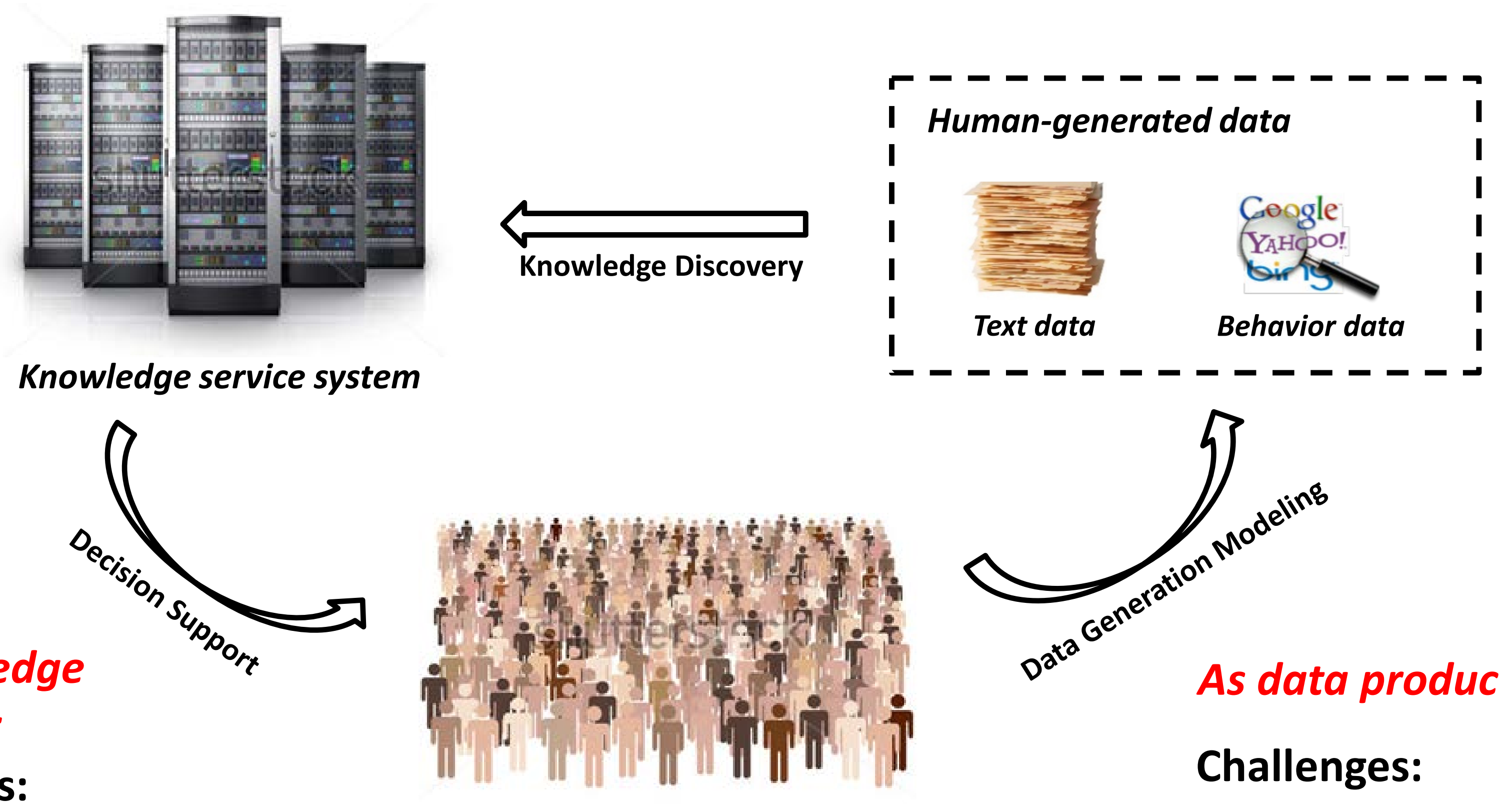


State-of-the-art solutions

- GraphLab (graphlab.com)
  - Specialized for sparse data with local dependencies for iterative algorithms



Text mining in the era of Big Data



Knowledge service system

Knowledge Discovery

Human-generated data

Text data

Behavior data

Decision Support

Data Generation Modeling

Human: big data producer and consumer

As knowledge consumer

Challenges:

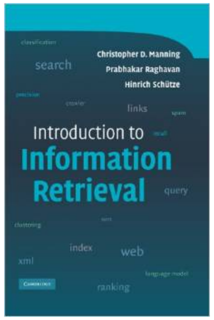
1. Implicit feedback
2. Diverse and dynamic

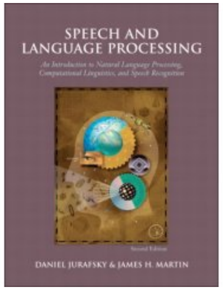
As data producer

Challenges:

1. Unstructured data
2. Rich semantic

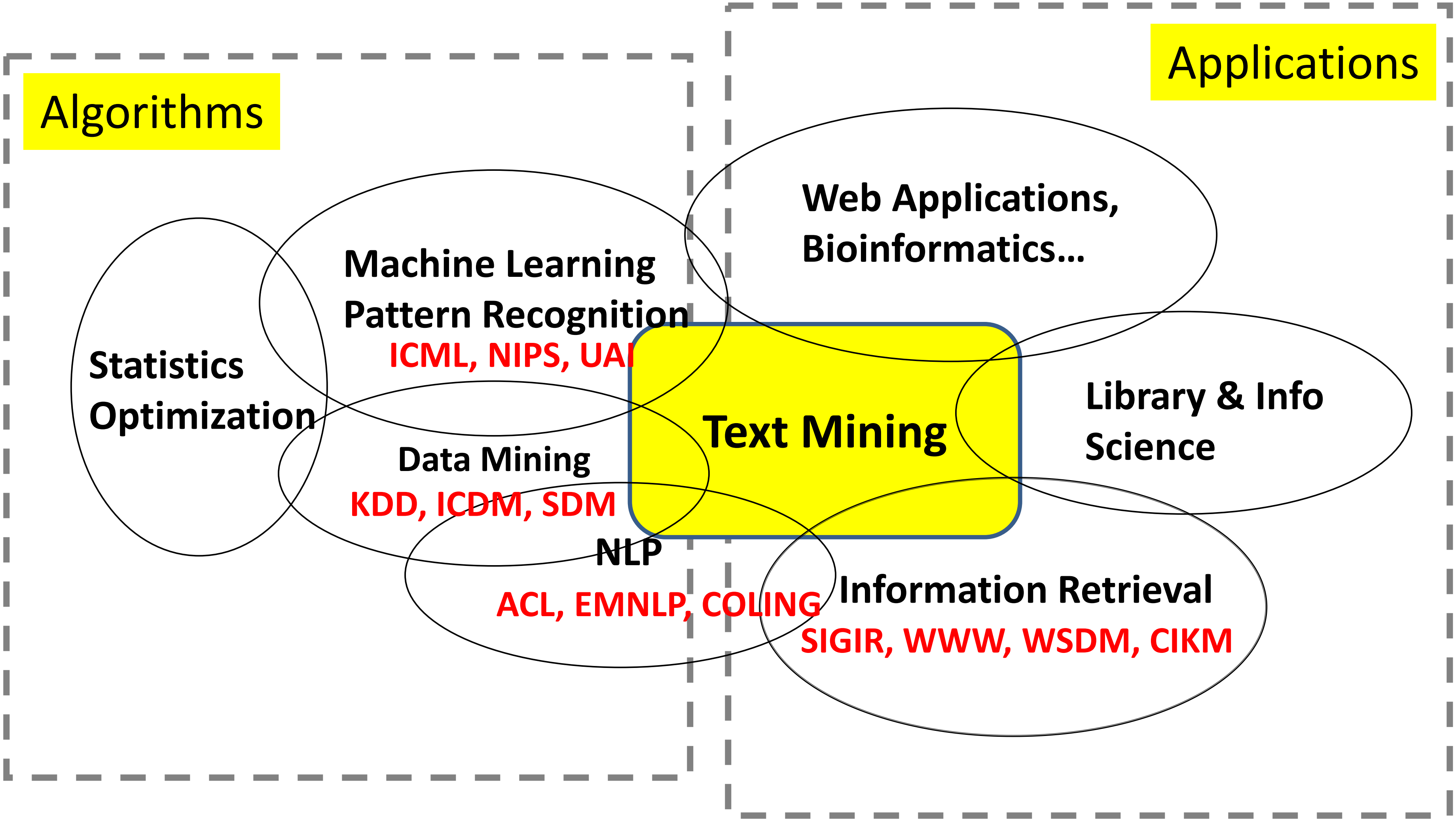
Text books





- ***Introduction to Information Retrieval***. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schuetze, Cambridge University Press, 2007.
- ***Speech and Language Processing***. Daniel Jurafsky and James H. Martin, Pearson Education, 2000.
- ***Mining Text Data***. Charu C. Aggarwal and ChengXiang Zhai, Springer, 2012.

What to read?



- Find more on course website for resource

Welcome to the class of “Text Mining”!