

Data Science Foundations

Master in Big Data Solutions 2017-2018



Liana Napalkova
liana.napalkova@bts.tech

Ludovico Boratto
ludovico.boratto@bts.tech

Francisco Gutierrez
francisco.gutierrez@bts.tech

Session4

Introduction to Text Mining

What is “Text Mining”?

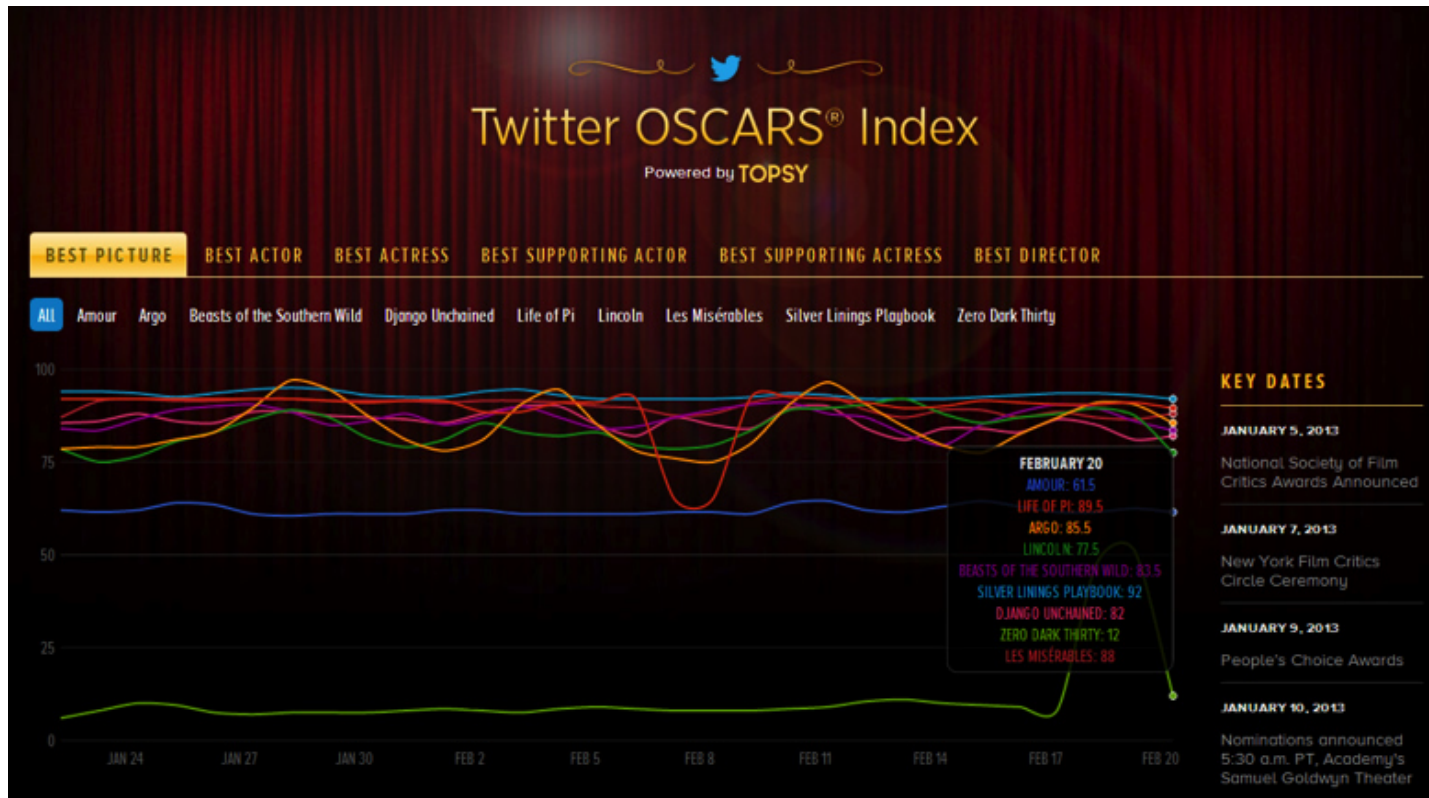
- *“Text mining, also referred to as **text data mining**, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text.” - wikipedia*
- *“Another way to view text data mining is as a process of **exploratory** data analysis that leads to **heretofore unknown** 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

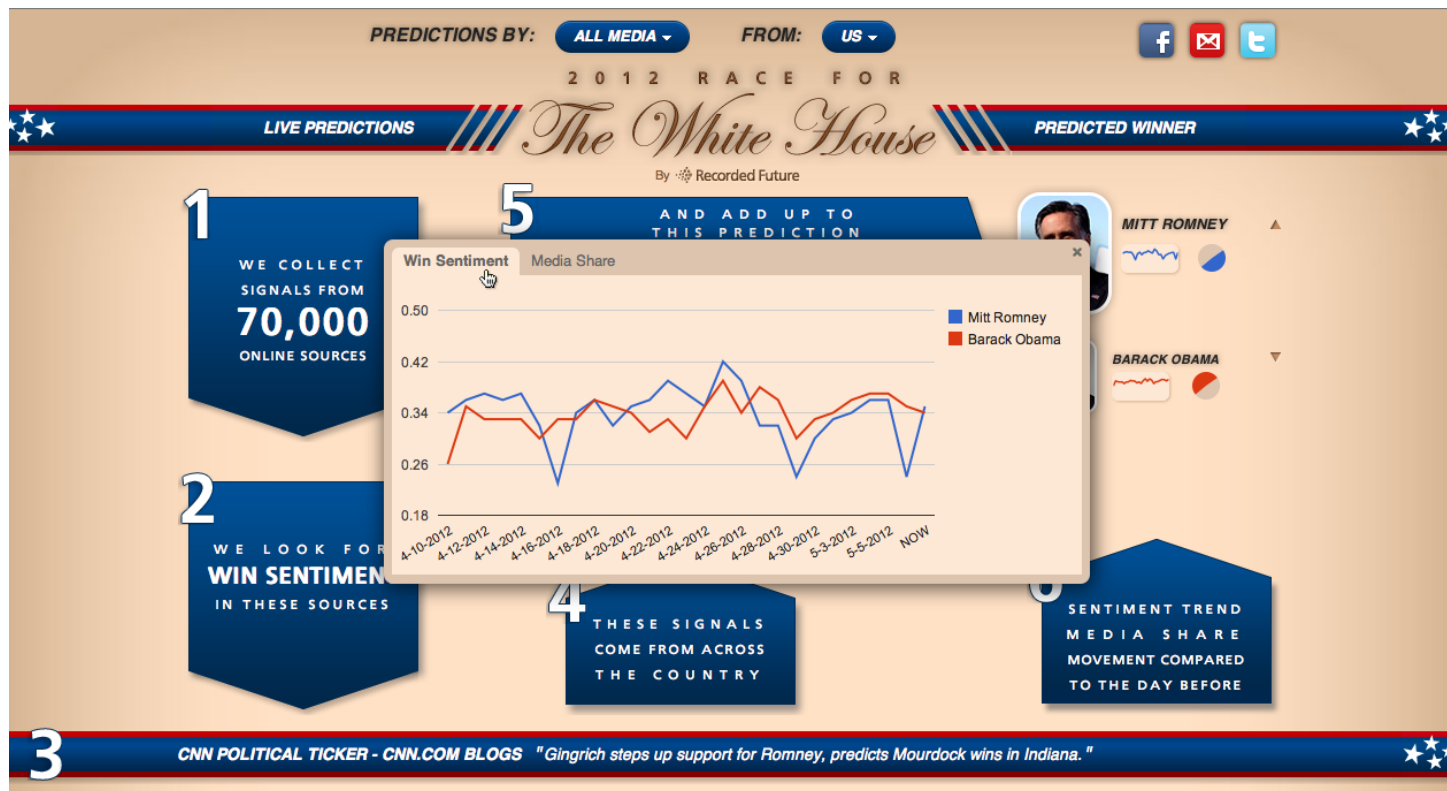
Text mining around us

- Sentiment analysis



Text mining around us

- Sentiment analysis



- Document summarization



Text mining around us

- Document summarization

The image shows a Bing search results page for the query 'text mining'. The search bar at the top shows 'bing' and 'text mining'. Below the search bar, there are tabs for 'Web', 'Images', 'Videos', 'Maps', 'News', and 'More'. The search results show 19,200,000 results, sorted by 'Any time'. The first result is from Wikipedia, titled 'Text mining - Wikipedia, the free encyclopedia'. The snippet for this result is highlighted with a red box: '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'. Below this, there are two more results. The second result is titled 'Text Mining (Big Data, Unstructured Data)' from statsoft.com, with a snippet: 'Text Mining Introductory Overview. The purpose of Text Mining is to process unstructured (textual) information, extract meaningful numeric indices from the text, ...'. The third result is titled 'Text Mining' from academic.research.microsoft.com, with a snippet: 'Text mining is defined as knowledge discovery in large text collections. It detects interesting patterns such as clusters, associations, deviations, similarities, and ...'. On the right side of the page, there is a 'Text mining' summary box. It contains the same snippet as the Wikipedia result, highlighted with a red box: '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... +'. Below this, it lists 'Related people: Jun'ichi Tsujii · Alfonso Valencia · Tomoko Ohta · Carol Friedman · Michael Berry · Hsinchun Chen' and 'People also search for: Sentiment analysis · Natural language processing · Web mining · Analytics · Cluster analysis +'. At the bottom of the page, there are 'Related searches' for 'Text Analysis Software' and 'Text Analytics'.

bing text mining

Web Images Videos Maps News More

19,200,000 RESULTS Any time ▾

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.

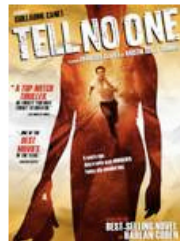
Text mining
 Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern 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) [See all >](#)



Tell No One

Because you enjoyed:
Memento
Syriana
Children of Men

Add



Not Interested



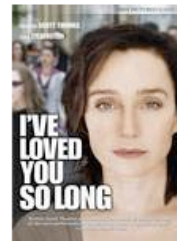
Let the Right One In

Because you enjoyed:
Seven Samurai
This Is Spinal Tap
The Big Lebowski

Add



Not Interested



I've Loved You So Long

Because you enjoyed:
The Queen
Syriana
Good Night, and Good Luck

Add



Not Interested



Downfall

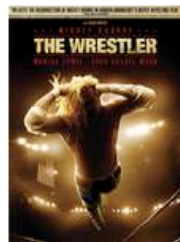
Because you enjoyed:
Das Boot
The Killing Fields
Seven Samurai

Add



Not Interested

DRAMA SUGGESTIONS (about 82) [See all >](#)



The Wrestler

Because you enjoyed:
Sin City
Reservoir Dogs
The Big Lebowski

Add



Not Interested



The Visitor

Because you enjoyed:
Gandhi
The Motorcycle Diaries
The Queen

Add



Not Interested



Brick

Because you enjoyed:
The Big Lebowski
Rushmore
Fight Club

Add



Not Interested



The Pianist

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


Add



Not Interested



Text mining around us


- Restaurant/hotel recommendation



[Home](#)
[About Me](#)
[Write a Review](#)
[Find Friends](#)
[Messages](#)
[Talk](#)
[Events](#)

Bodo's Bagels








186 reviews


 Details

[\\$ - Bagels, Breakfast & Brunch, Sandwiches](#)
[Edit](#)




1418 Enmet St N
Charlottesville, VA 22903

[Get Directions](#)


 (434) 977-9598
 [Message the business](#)
 [bodosbagels.com](#)




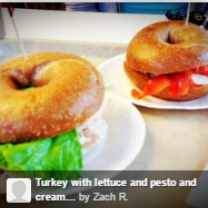

"Almost any combination of bagel, [cream cheese](#) or spread or sandwich you could dream of you can find at Bodos." in 38 reviews
\$0.60 Cream Cheese




"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

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

 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	
Wed	6:30 am - 8:00 pm	Closed now
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**

Search reviews



English (186)

New York City › Hotels › Flights › Vacation Rentals › Restaurants › Things To Do › Best of 2015 › Your Friends › More › Write a Review

📍 New York City, New York, United States 🔍 What are you looking for? 🔍 Search

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

Check In Check Out

Check Availability

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 \(1,085\)](#) [Save](#)


4,919 Reviews from our TripAdvisor Community

[Write a Review](#) [Add Photo](#)

Text mining around us


- News recommendation

[All Stories](#) [News](#) [Entertainment](#) [Sports](#) [Business](#) [More](#) ▼




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



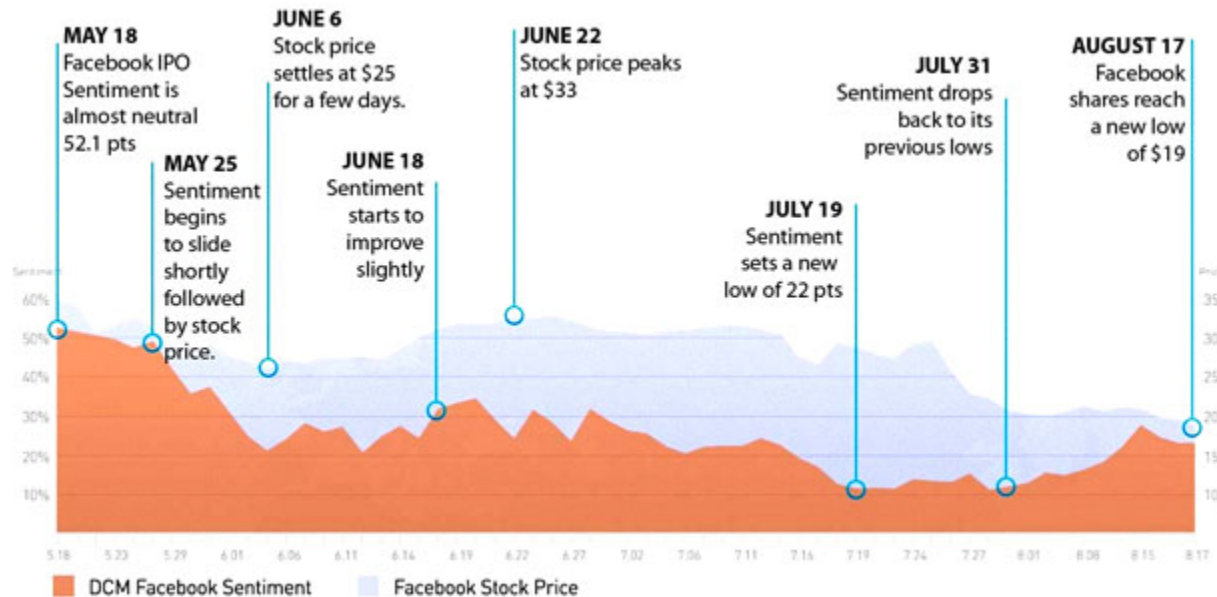
Wisconsin man keeps 40-year-old Christmas tree up until son returns
By Brendan O'Brien (Reuters) - A Wisconsin man will refuse for about the 40th time to partake in the annual after-holiday chore of putting Christmas
[Reuters](#)



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



Text mining around us

- Text analytics in healthcare

REQUEST FOR MEDICAL/DENTAL RECORDS		DATE
1. PATIENT (Last Name - First Name - Middle Name)		December 20, 1989
2. NATIONAL PERSONNEL RECORDS CENTER (Military Personnel Records) 9700 Pappe Boulevard St. Louis, Missouri 63132		RCMP #
3. TO:	4. SERVICE NO.(S)	5. GRADE OR RATE
Commander V.A. Air Force Hospital Scott AFB, Illinois		A 2/c
6. VA CLAIM NUMBER		
7. ORGANIZATION AND PLACE OF TREATMENT	8. DATES OF TREATMENT (mm/dd)	9. DISEASE OR INJURY
Your Hospital	1-23-61 to 3-28-61	Kidney operation
10. RECORDS REQUESTED: <input type="checkbox"/> CLINICAL <input type="checkbox"/> OUTPATIENT <input type="checkbox"/> HEALTH RECORD <input type="checkbox"/> DENTAL RECORD <input type="checkbox"/> X-RAY <input checked="" type="checkbox"/> MEDICAL REPORT CARDS, EMERGENCY MEDICAL TAGS, FIELD MEDICAL CARDS OTHERS (See remarks)	11. REMARKS Forward records to address in item 13, below	
12. SIGNATURE J. M. P. DATE: 10/17/89 ST. LOUIS, MO 63132 RTI: B. Ray		
FIRST ENDORSEMENT		
13. TO:	14. ACTION TAKEN	
VARD 1000 Liberty Avenue Pittsburgh, PA 15222	<input type="checkbox"/> AVAILABLE RECORDS ENCLOSED <input type="checkbox"/> NO RECORDS ON FILE	
15. ENCLOSURES (Number of) <input type="checkbox"/> CLINICAL <input type="checkbox"/> OUTPATIENT <input type="checkbox"/> HEALTH RECORD <input type="checkbox"/> DENTAL RECORD <input type="checkbox"/> X-RAY <input type="checkbox"/> MEDICAL REPORT CARDS, EMERGENCY MEDICAL TAGS, FIELD MEDICAL CARDS OTHERS (See remarks)	16. REMARKS	
17. DATE		18. SIGNATURE

NATIONAL ARCHIVES AND RECORDS ADMINISTRATION RA FORM 33042-A (9-85)

WebMD-moderated WebMD® Heart Disease Community

- Home
- Discussions
- Tips
- Resources
- About This Community
- Staying Informed
- My Watchlist
- Related Men's Health Communities
- All Communities
- Community FAQs
- Crisis Assistance

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.
Enter Email Address

Sign Up

What's Happening Now

See All Discussions | Tips | Resources



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

Posted by cardiostarus1

Was this Helpful?

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 cardiostarus1

3 Replies
INCLUDES EXPERT CONTENT

Report This

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...

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...

Posted by smacmill

1 Reply

Report This

Search This Community

GO

Post Now

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
- 28y/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?

1 of 1 found this helpful

tip for the pain.

Post a Tip | See All

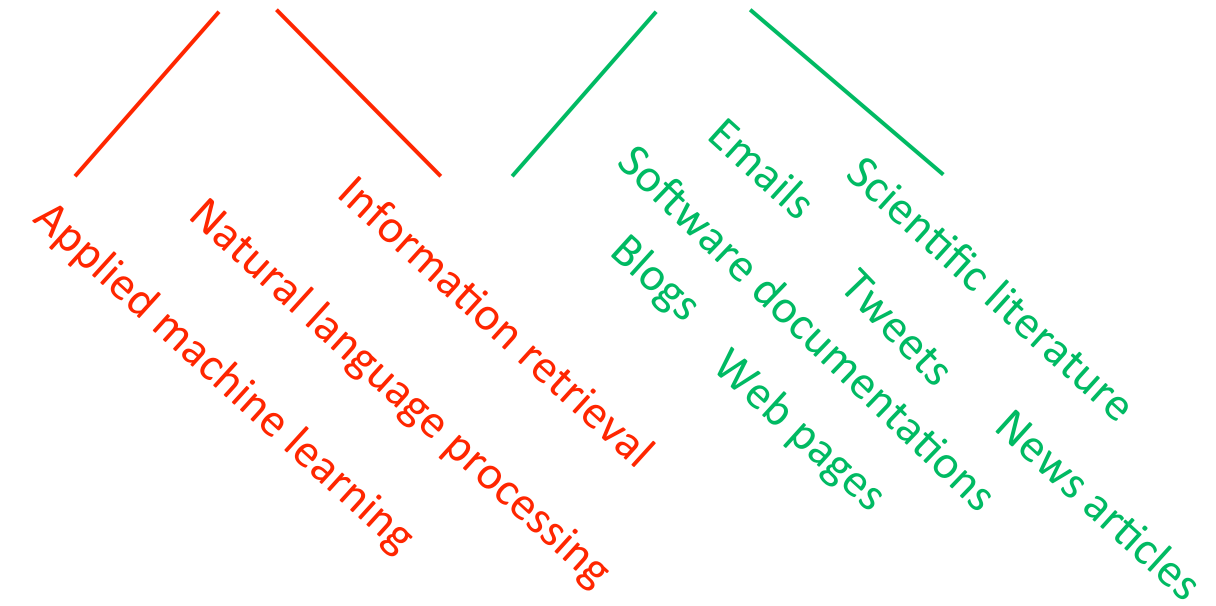
Helpful Resources

- Super-safe iodine 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**



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-Novels
Non-textual data	General data-mining	Exploratory analysis	Database queries
Textual data	Computational Linguistics		Information retrieval

**Text
Mining**

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?

The NLTK library

TEXT MINING IN PYTHON

The NLTK library

- [NLTK](#) (*Natural Language Toolkit*) is the most famous Python Natural Language Processing Toolkit
- Description from the website:
 - NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning.

Installing NLTK

- **Install Pip:** run `sudo easy_install pip`
- **Install Numpy:**
run `sudo pip install -U numpy`
- **Install PyYAML and NLTK:** run `sudo pip install -U pyyaml nltk`
- **Test installation:** run `python` then type `import nltk`

Installing NLTK Data

- After installing NLTK, you need install NLTK Data which include a lot of corpora, grammars, models and etc.
- You can find the complete nltk data list here:
http://nltk.org/nltk_data/
- The simplest way to install NLTK Data is run the Python interpreter and type the commands

```
>>> import nltk
>>> nltk.download( )
```
- A new window should open, showing the NLTK Downloader

Installing NLTK Data

Collections	Corpora	Models	All Packages
Identifier	Name	Size	Status
all	All packages	n/a	not installed
all-corpora	All the corpora	n/a	not installed
book	Everything used in the NLTK Book	n/a	not installed

Download

Refresh

Server Index:

Download Directory: