

# TEXT MINING AND SEARCH (AND THE FIRST PART OF THE IR COURSE – AA 2024/25)

---

Gabriella Pasi

[gabriella.pasi@unimib.it](mailto:gabriella.pasi@unimib.it)

Marco Viviani

[marco.viviani@unimib.it](mailto:marco.viviani@unimib.it)



# Introduction to the course

- Why we do aggregate both TM&S and IR students?
- The **first part** of the course will explain what text mining is, and how it is related to (textual) Information Retrieval.
- It will introduce the basics of *text processing*, which are also employed in IR

# Introduction to the course

- Moreover it will explain the task of Information Retrieval, which is indeed a part of the TM&S course.
- **For these reasons the first part of the two courses will be shared.**
- The shared lessons will be delivered for the first 14 hours.
- **After that the two courses will be diversified ....**

# What the shared part of the two courses will address?

- The focus of both courses will be on ***texts***
- Both courses share the problem related to *text* representation, analysis and processing
- Both courses will also address the task of Information Retrieval, commonly known as Search (all you know Web Search Engines) → students of Computer Science will address more technical issues, while students of Data Science will address more modeling and applicative issues related to Search engines, plus other text mining tasks.

# Organization of the two courses:

## First part – both TM&S and IR

- Definition of Text Mining
- Main differences between Text Mining and Data Mining
- The main tasks related to TM
  - Text classification, text clustering, text summarization
  - Information Retrieval and Information Filtering
- Text pre-processing
- Indexing and Text Representation

# Organization of the two courses:

## Second part – TM&S

- Text Mining tasks:
  - Topic Modeling
  - Text Classification
  - Text Clustering
  - Text Summarization
- Introduction to Information Retrieval
  - Text Based Search Engines and Web Search Engines
  - Information Filtering
- Open Source software for Text Mining and Search
- Lab TM&S (Dr. Luca Celotti Herranz)
  - Introduction to open source software for Text Mining tasks

# Organization of the two courses:

## Second part – IR

- Information Retrieval Models
- Web Search Engines
- The evaluation of Search Engines
- Advanced topics
  
- Lab IR (Dr. Georgios Peikos, [georgios.peikos@unimib.it](mailto:georgios.peikos@unimib.it))
  - Introduction to an open source software platforms for the development of search engines/recommender systems

# Suggested readings

- Suggested readings will be uploaded on the elearning platform



# Exam

- There will be a written exam composed of questions related to the various topics addressed during the course
- Project to be developed by groups of students (up to three)

And if you desire I am available to an oral examination to (possibly 😊 ) increase your mark.



And now let us start !

# Text Mining – the origins

- In 2004 Ian Witten (Weka’s “father”) published a paper titled «Text Mining» in *The Practical Handbook of Internet Computing*.
- I will report some key sentences of this interesting article, which you can find on the Moodle platform related to this course.
- In his paper Ian Witten reports that: “....the first workshops [on text mining] were held at the *International Machine Learning Conference* in July 1999 and the *International Joint Conference on Artificial Intelligence* in August 1999”

# Text Mining

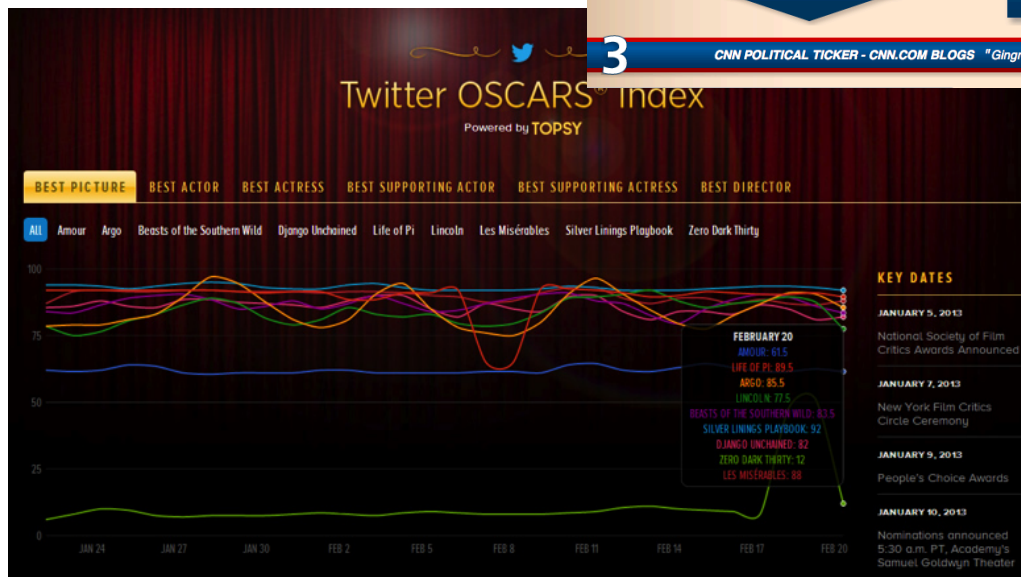
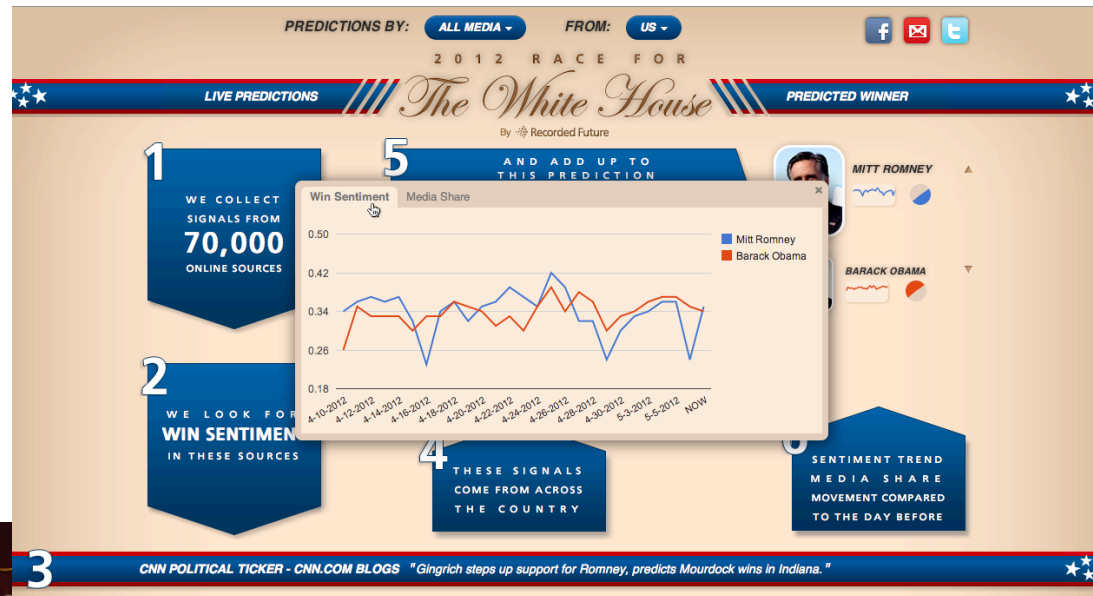
- *“Text mining is a burgeoning new field that attempts to glean meaningful information from natural language text. It may be loosely characterized as the process of analyzing text to extract information that is useful for particular purposes.”*
- *“.....the phrase “text mining” appears 17 times as often as “text data mining” on the Web, according to a popular search engine (and “data mining” occurs 500 times as often).”*

# Text Mining

- Another definition:
  - “ *The phrase “text mining” is generally used to denote any system that analyzes large quantities of natural language text and detects lexical or linguistic usage patterns in an attempt to extract probably useful (although only probably correct) information*” [Sebastiani, 2002].

# Text mining is around us

- Sentiment analysis





# Text mining around us

- Document summarization

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](https://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](http://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](http://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](http://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**.

*snippets*

**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](https://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 is around us

- Restaurant/hotel recommendation

**Bodo's Bagels**  
 186 reviews  
 1418 Emmet St N, Charlottesville, VA 22903  
 (434) 977-9598  
 bodosbagels.com

**Reviews:**  
 "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

**Hours:**  
 Today 6:30 am - 8:00 pm **Closed now**  
 Mon 6:30 am - 8:00 pm  
 Tue 6:30 am - 8:00 pm  
 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

**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

**PriceFinder:** Enter dates for best prices. Check In, Check Out. Check Availability.

**Book on TripAdvisor:** or compare prices from up to 200 sites including: Booking.com, travelocity, Expedia.

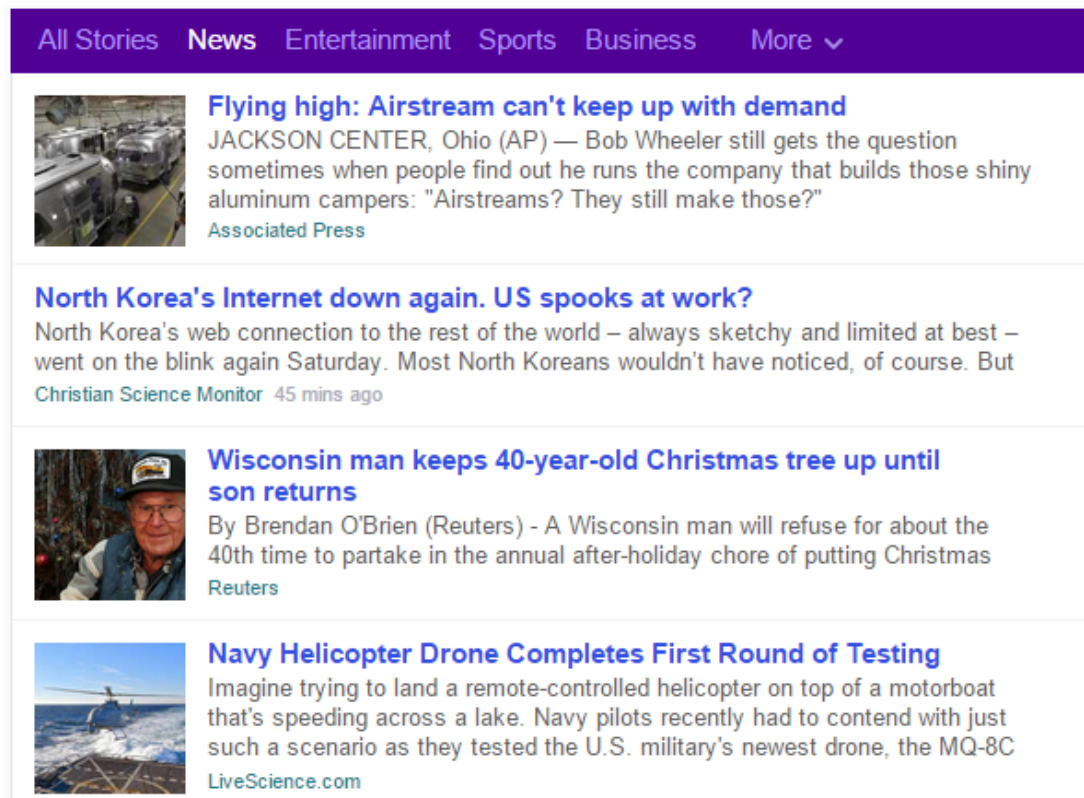
**Photos:** Hilton Times Square, Applebee's, MONEY EXCHANGE, Times Square.

**Navigation:** 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 is 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

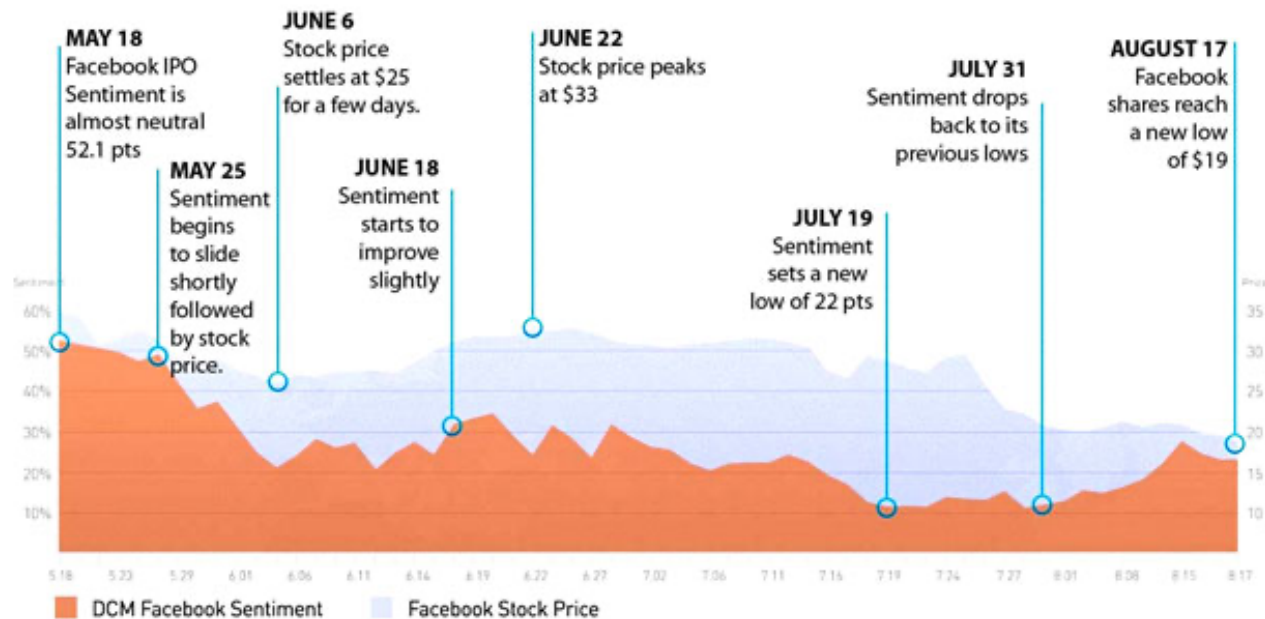
**North Korea's Internet down again. US spooks at work?**  
North Korea's web connection to the rest of the world — always sketchy and limited at best — went on the blink again Saturday. Most North Koreans wouldn't have noticed, of course. But  
Christian Science Monitor 45 mins ago

 **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 is around us

- Text analytics in financial services



# Text mining is around us

- Text analytics in healthcare

|  |  |   |                   |                      |
|--|--|---|-------------------|----------------------|
| REQUEST FOR MEDICAL/DENTAL RECORDS   |  | DATE  | December 20, 1987 |                      |
| 1. PATIENT (Last Name - First Name - Middle Name)  |  | NATIONAL PERSONNEL RECORDS CENTER<br>(Military Personnel Records)<br>9700 Page Boulevard<br>St. Louis, Missouri 63132 |                   |                      |
| 3. TO:   |  | 4. SERVICE NO.(S)   | 5. GRADE OR RATE  |                      |
| Commander<br>V.A. Air Force Hospital<br>Fort AFB, Kansas   |  |   | A 2/c             |                      |
| 7. ORGANIZATION AND PLACE OF TREATMENT   |  | 8. DATES OF TREATMENT (incl)  |                   | 9. DISEASE OR INJURY |
| Your Hospital  |  | 1-23-61 to 3-26-61  |                   | Kidney operation     |
| 10. RECORDS REQUESTED  |  | 11. REMARKS   |                   |                      |
| <input type="checkbox"/> CLINICAL<br><input type="checkbox"/> OUTPATIENT<br><input type="checkbox"/> HEALTH RECORD<br><input type="checkbox"/> DENTAL RECORD<br><input type="checkbox"/> X-RAY<br><input type="checkbox"/> MEDICAL REPORT CARDS, EMERGENCY MEDICAL TAGS,<br><input type="checkbox"/> FIELD MEDICAL CARDS<br><input type="checkbox"/> OTHERS (See remarks)<br><input checked="" type="checkbox"/> ALL AVAILABLE RECORDS (Search will include all<br>Hospital, dispensary, clinic, or other medical<br>facilities) |  | Forward records to<br>address on item 13,<br>below  |                   |                      |
| 13. TO:  |  | 12. SIGNATURE   |                   |                      |
| WARD Liberty Avenue<br>1000 Liberty Avenue<br>Pittsburgh, PA 15222   |  | [Signature]<br>NATIONAL PERSONNEL RECORDS CENTER (MILPERC)<br>ST. LOUIS, MO 63132<br>RTN B. Key                       |                   |                      |
| 15. ENCLOSURES (Number of)   |  | 14. ACTION TAKEN  |                   |                      |
| <input type="checkbox"/> CLINICAL<br><input type="checkbox"/> OUTPATIENT<br><input type="checkbox"/> HEALTH RECORD<br><input type="checkbox"/> DENTAL RECORD<br><input type="checkbox"/> X-RAY<br><input type="checkbox"/> MEDICAL REPORT CARDS, EMERGENCY MEDICAL TAGS,<br><input type="checkbox"/> FIELD MEDICAL CARDS<br><input type="checkbox"/> OTHERS (See remarks)  |  | <input type="checkbox"/> AVAILABLE RECORDS ENCLOSED<br><input type="checkbox"/> NO RECORDS ON FILE                    |                   |                      |
| 16. REMARKS  |  | 17. DATE  |                   |                      |
|  |  |   |                   |                      |
| 18. SIGNATURE  |  |   |                   |                      |

NATIONAL ARCHIVES AND RECORDS ADMINISTRATION NA FORM 13042-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

What's Happening Now  
See All Discussions | Tips | Resources

Search This Community

Popular Discussions

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...  
 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  
 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

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 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

# Text mining and Data Mining

- **Data mining** can be more fully characterized as the extraction of implicit, previously unknown, and potentially useful information from data [Witten and Frank, 2000].
  - *The information is implicit in the input data: it is hidden, unknown, and could hardly be extracted without recourse to automatic techniques of data mining.*
- With **text mining**, the information to be extracted is clearly and explicitly stated in the text. It is not hidden at all
  - *Text mining strives to bring it out of the text in a form that is suitable for consumption by computers directly, with no need for a human intermediary.*

# Text Mining and Data Mining

- *“Mining implies extracting precious nuggets of ore from otherwise worthless rock”.*
- *If data mining really followed this metaphor, it would mean that people were discovering new factoids within their inventory databases. However, in practice this is not really the case. Instead, data mining applications tend to be (semi)automated discovery of trends and patterns across very large datasets, usually for the purposes of decision making*

From: Marti A. Hearst. 1999. Untangling text data mining. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics (ACL '99).

# Text Mining and Data Mining

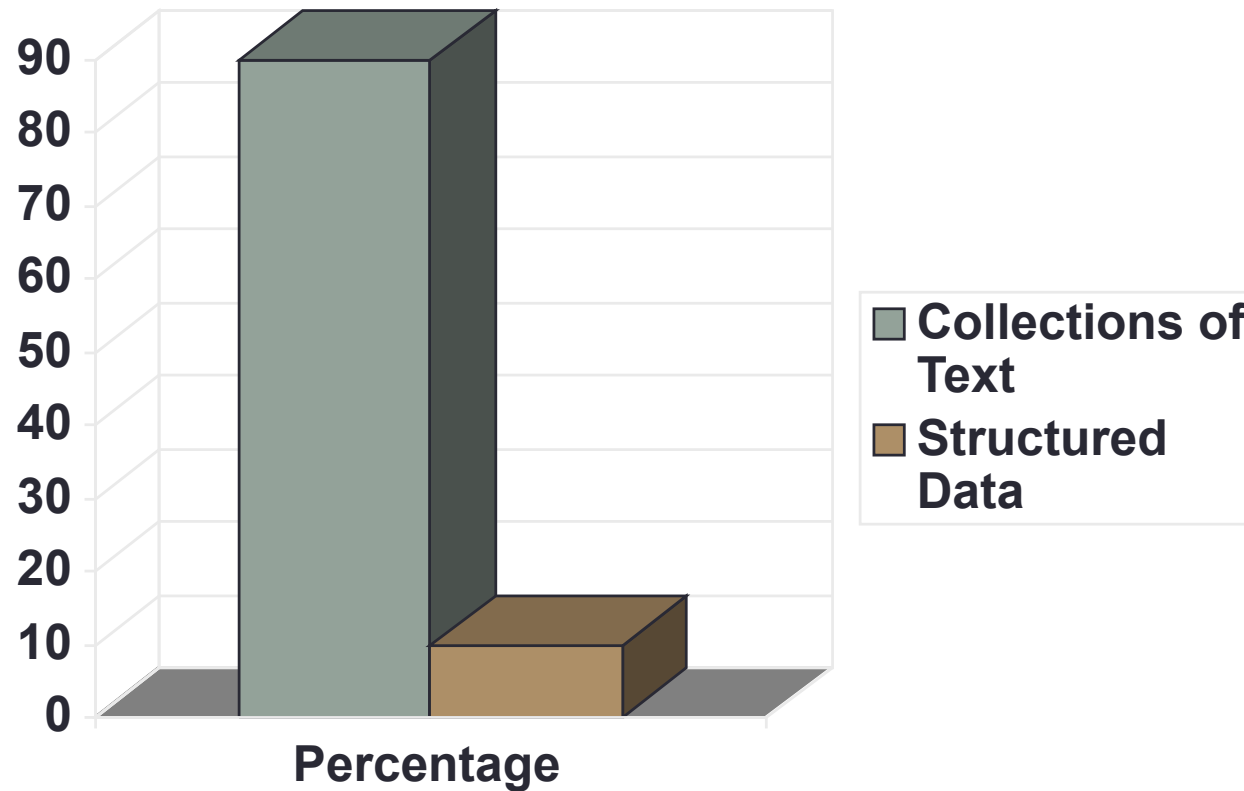
|                  | Finding Patterns          | Finding Nuggets |                       |
|------------------|---------------------------|-----------------|-----------------------|
|                  |                           | Novel           | Non-Novel             |
| Non-textual data | standard data mining      | ?               | database queries      |
| Textual data     | computational linguistics | real TDM        | information retrieval |

Table 1: A classification of data mining and text data mining applications.

***Information Retrieval was founded well before the appearance of the Expression Text Mining***

It contributed to the ***basis of the analysis of texts***, as we will see later

# Interest in Text Mining





# Examples of texts

- Email
- Insurance claims
- News articles
- Web pages
- Patent portfolios
- User generated content in Social media (*course on Social Media Analytics*)
- Customer complaint letters
- Contracts
- Transcripts of phone calls with customers
- Technical documents
- Scientific papers
- Health related information ....

# Challenges in Text Mining

- Documents in an unstructured textual form are not readily accessible to be used by computers
- Dealing with huge *collections of documents* or *streams of texts*
- Data is not well-organized
  - Semi-structured or unstructured
- Natural language text contains ambiguities on many levels
  - Lexical, syntactic, semantic, and pragmatic

# Text Mining and Search (Information Retrieval)

- Information Retrieval:
  - Make it easier to find things on the Web.
  - You ask and the collection is “mined” to find useful answers
  - Its roots date back to 70ties (and even before)
- The metaphor of extracting ore from rock:
  - extracting documents of interest from a huge pile (Extraction of useful information from huge data repositories)
  - based on analysis of texts to find *correspondence* with a user query

We will go deeper inside IR

# TASKS AFFECTED BY TEXT MINING

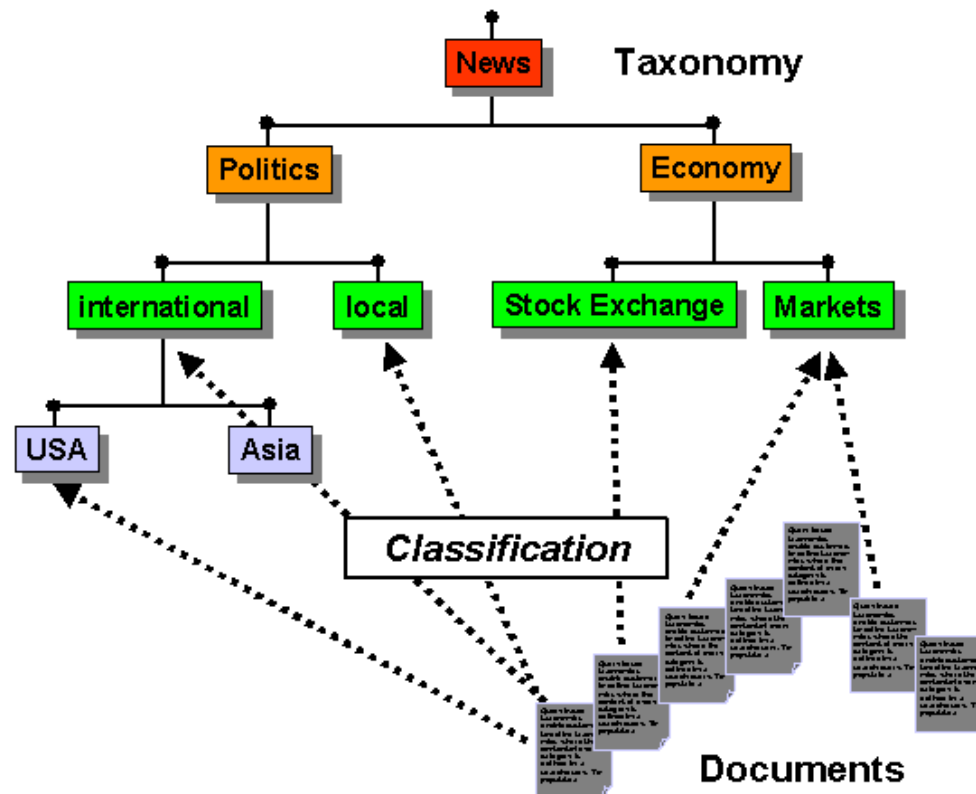
- **Text Summarization:** A text summarizer strives to produce a condensed representation of its input, intended for human consumption
- **Information Retrieval:** given a corpus of documents and a user's information need expressed by a query, IR is the task of identifying and returning the most relevant documents to the query. Web search engine also apply text summarization stage that focuses on the query posed by the user to provide a short synthesis of the retrieved documents.
- **Content Based Recommender Systems:** textual contents produced in a stream are pushed to a user to fulfill the user preferences as represented in a user model, also called user profile

# TASKS AFFECTED BY TEXT MINING

- **Text Classification** : Text classification (or text categorization) is the assignment of natural language documents to pre-defined categories according to their content [Sebastiani, 2002]. It has a variety of applications (e.g. sentiment analysis) hot topic in machine learning (supervised learning)
- **Document clustering**: document clustering is “unsupervised” learning in which there is no predefined category or “class,” but groups of documents that share the similar topics are sought.
- Topic identification and tracking

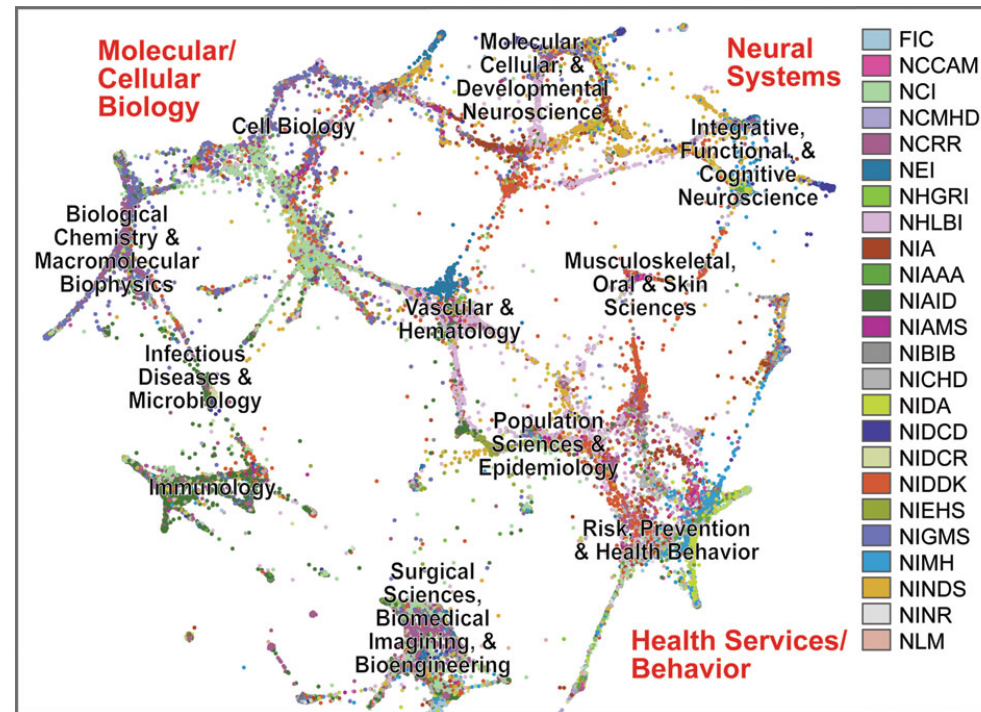
# Document classification

- Document classification
  - Possible application: adding structure to the text corpus



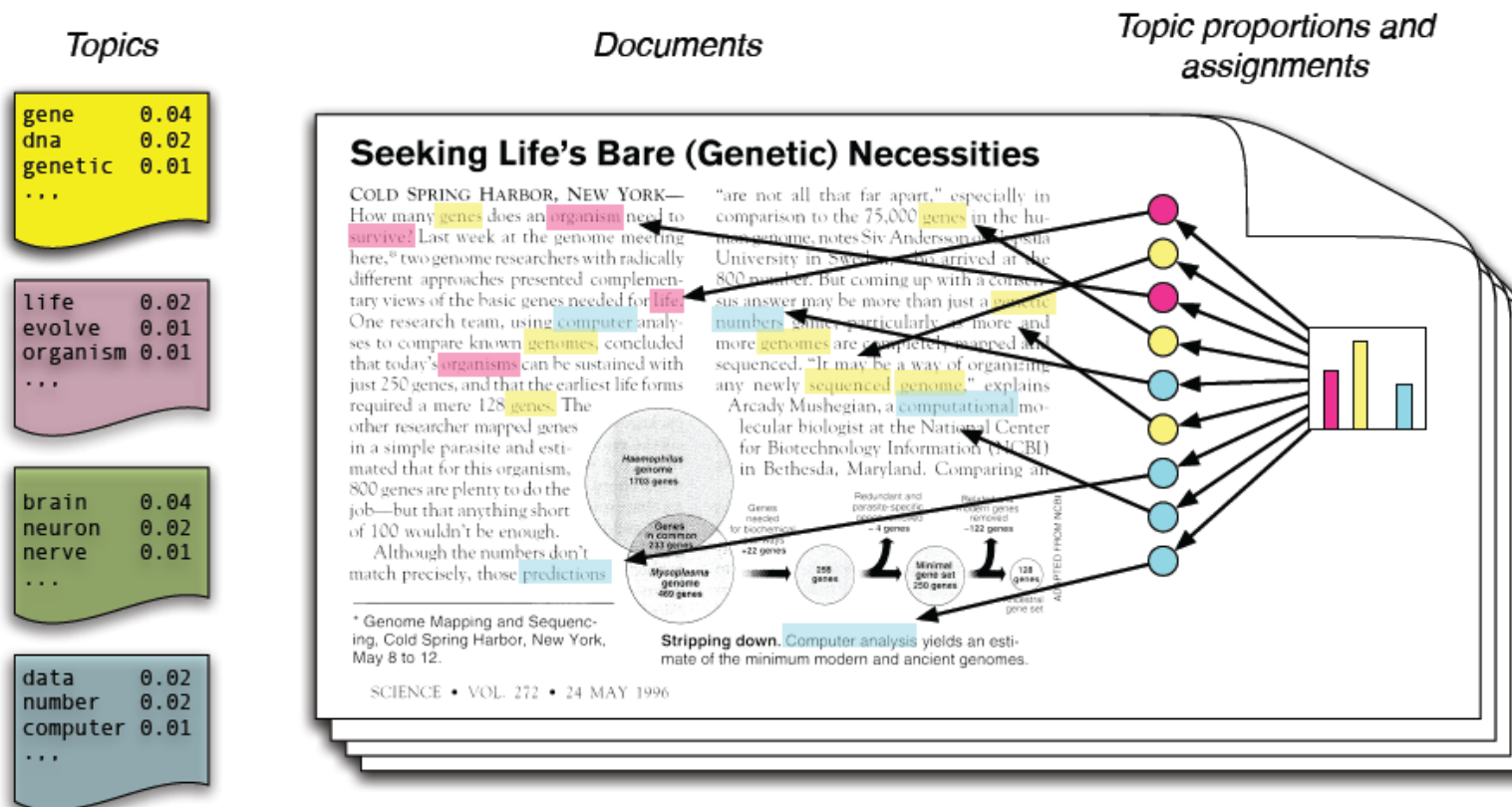
# Text clustering

- Text clustering
  - Possible application: identifying structures in a text corpus



# Topic Modeling

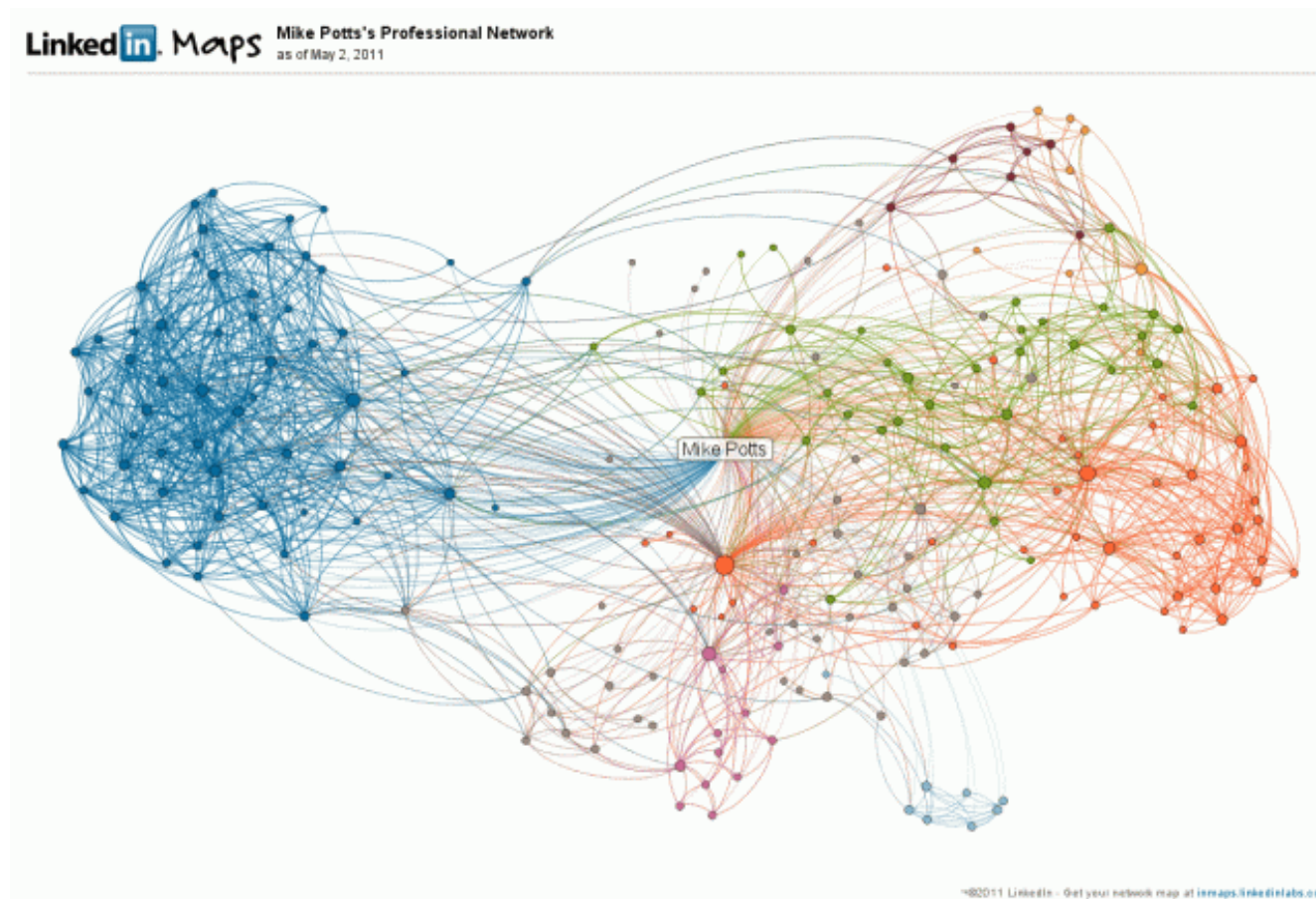
Identifying topics in the text corpus (or in single texts)





# Social Media Analytics

- Exploring additional structure in the text corpus



# Mining structured information from texts

## **Mining structured data within texts.**

- Entity extraction: many practical tasks involve identifying linguistic constructions that stand for objects or “entities” in the world (e.g. names of people, places, etc.)
- Information Extraction: the task of filling templates from natural language input
- Learning rules from texts: extracting rules that characterize the content of the text itself.

# Predictive and Exploratory Analysis of Text

- Predictive Analysis of Text
  - developing computer programs that automatically recognize or detect *a particular concept* within a span of text.
- Exploratory Analysis of Text:
  - developing computer programs that automatically discover *interesting and useful patterns or trends* in text collections.

# Predictive Analysis of Text: examples

- *Opinion Mining*
  - automatically detecting whether a span of opinionated text expresses a positive or negative opinion about the item being judged
- *Sentiment/Affect Analysis*
  - automatically detecting the emotional state of the author of a span of text (usually from a set of *pre-defined* emotional states).
- *Bias Detection*
  - automatically detecting whether the author of a span of text favors a particular viewpoint (usually from a set of *pre-defined* viewpoints)

# Opinion Mining: movie reviews

“Great movie! It kept me on the edge of my seat the whole time. I IMAX-ed it and have no regrets.” **positive**

“Waste of time! It sucked!” **negative**

“This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can't hold up.” **negative**

“Trust me, this movie is a masterpiece .... after you've seen it 4+ times.” **???**

# Sentiment Analysis in support group posts

- “[I] also found out that the radiologist is doing the biopsy, not a breast surgeon. I am more scared now than when I ...”  
fear
- “... My radiologist ‘assured’ me my scan was NOT going to be cancer...she was wrong.”  
despair
- “ ... My radiologist did my core biopsy. Not a problem and he did a super job of it.”  
hope
- “It's pretty standard for the radiologist to do the biopsy so I wouldn't be concerned on that score.”  
hope

# Bias Detection

- “Nationalizing businesses, nationalizing banks, is not a solution for the democratic party, it's the objective.” -- Rush Limbaugh **conservative (vs. liberal)**
- “If you're keeping score at home, so far our war in Iraq has created a police state in that country and socialism in Spain. So, no democracies yet, but we're really getting close.” -- Jon Stewart **against war in iraq (vs. in favor of)**

# Predictive Analysis of Text: examples

- *Information Extraction*

- automatically detecting that a short sequence of words belongs to (or is an instance of) a particular entity type, for example:
  - ▶ Person(X)
  - ▶ Location(X)
  - ▶ TennisPlayer(X)


- *Relation Learning*

- automatically detecting pairs of entities that share a particular relation, for example:
  - ▶ CEO(<person>, <company>)
  - ▶ Capital(<city>, <country>)
  - ▶ Mother(<person>, <person>)
  - .....



# Relation Learning

CEO(<person>, <company>)

[Know Yahoo's Marissa Mayer in 11 facts - CNN.com](#)

[www.cnn.com/2012/07/17/...marissa-mayer/index.html](http://www.cnn.com/2012/07/17/...marissa-mayer/index.html)



by John D. Sutter - in 846,411 Google+ circles - More by John D. Sutter

Jul 19, 2012 – Here's a quick guide to some of the most interesting and water-cooler-worthy facts about **Marissa Mayer**, who was named CEO of **Yahoo** on

...

<person>, who was named CEO of <company>

# Relation Learning

## CEO(<person>,<company>)

Search query: ",who was named CEO of"



### [DailyTech - Fisker Appoints New CEO, Eliminates Battery/Engine ...](#)

[www.dailytech.com/article.aspx?newsid=25412](http://www.dailytech.com/article.aspx?newsid=25412)

4 days ago – Tom LaSorda, **who was named CEO of Fisker** back in February 2012 when founder Henrik Fisker stepped down, is leaving the company, but ...

CEO(Tom LaSorda, Fisker)

### [who was named CEO of Yahoo on Monday. Christian Science Monitor](#)

[gtp123.com/.../who-was-named-ceo-of-yahoo-on-monday-christian-...](http://gtp123.com/.../who-was-named-ceo-of-yahoo-on-monday-christian-...)

Jul 17, 2012 – You are browsing the archive for **who was named CEO of Yahoo** on Monday. Christian Science Monitor. Avatar of Garland E. Harris ...

### [CEO of renamed Sara Lee meat biz chooses Winnetka - Residential ...](#)

[www.chicagorealestatedaily.com](http://www.chicagorealestatedaily.com) › Home › Residential News

Aug 7, 2012 – Sean Connolly, **who was named CEO of Hillshire Brands Co.** in January, declines to comment through a company spokesman. Records show ...

CEO(Sean Connolly, Hillshire Brands)

### [Who is the woman who was named CEO of Gilt Groupe in Septemb...](#)

[askville.amazon.com](http://askville.amazon.com) › Miscellaneous › Popular News

Askville Question: Who is the woman **who was named CEO of Gilt Groupe** in September? : Popular News.

CEO(woman, Gilt Groupe)

### [Tom McKillop - Wikipedia, the free encyclopedia](#)

[en.wikipedia.org/wiki/Tom\\_McKillop](http://en.wikipedia.org/wiki/Tom_McKillop)

Sir Thomas Fulton Wilson "Tom" McKillop, FRS (born 19 March 1943) is a Scottish chemist, **who was named CEO of AstraZeneca PLC** in 1999 (retired 1 January ...

CEO(scottish chemist, AstraZeneca)

### [Harrison adjusts to view from top at First Hawaiian - Pacific Business ...](#)

[www.bizjournals.com/.../harrison-adjusts-to-view-from-top-at.html?...](http://www.bizjournals.com/.../harrison-adjusts-to-view-from-top-at.html?...)

Jan 27, 2012 – Bob Harrison, **who was named CEO of First Hawaiian Bank** on Jan. 1, says he'll spend a lot of time focusing on his people and community ...

CEO(Bob Harrison, First Hawaiian Bank)

# Predictive Analysis of Text

- Text-driven Forecasting
  - monitoring incoming text (e.g., tweets) and making predictions about external, real- world events or trends
    - a presidential candidate's poll rating
    - a company's stock value change
    - a movie's box office earnings
    - side-effects for a particular drug
    - ...
- Temporal Summarization
  - monitoring incoming text (e.g., tweets) about a news event and predicting whether a sentence should be included in an on-going summary of the event

# Exploratory analysis of text

- Text clustering
- Topic modeling
- ...

# Mining structured text

- Several Web resources have a structure: for example Web pages are written in HTML. XML is another markup language that provides a “logical” structure to a text.
- Many software systems use external online resources by hand-coding simple parsing modules, commonly called “wrappers,” to analyze the page structure and extract the required information.

# Welcome to the class of “Text Mining”!

