TEXT MINING AND SEARCH (and the first part part of INFORMATION RETRIEVAL)

Gabriella Pasi
gabriella.pasi@unimib.it

Marco Viviani
marco.viviani@unimib.it

Introduction to the Course (1)

- Why do we 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 (2)

- 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 till mid-October (further information will be provided via the e-learning platform)
- After that, the two courses will be diversified…

Shared Part of the two Courses

- 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
  - Students of Data Science will address more modeling and applicative issues related to search engines, plus other text mining tasks

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, topic modeling, text summarization
  - Information Retrieval and Information Filtering
- Text pre-processing
- Indexing and representation (from sparse to dense representations)

Second Part – TM&S

- Text Mining tasks
  - Text Classification and Clustering
  - Topic Modeling
  - Text Summarization

- Introduction to Information Retrieval and Information Filtering
  - Text Based Search Engines, Web Search Engines, Recommender Systems

- Lab (Joseph Muddle – j.muddle@campus.unimib.it)
  - Introduction to open source software for Text Mining tasks

Second Part – IR

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

Exam (Overall Information)

- There will be a **written exam** composed of open questions related to the various topics addressed during the course (grade: max 30/30)

- **Project** to be developed by groups of students (up to three students; grade: from 0 to max 4 points to be added to the written exam grade)
  - Sufficiency in the written exam is mandatory to add these points (i.e., you must obtain at least 18/30 in the written exam)
  - The “lode” is assigned with a score of at least 31,5/30
  - Further details on the project will be provided during the course and shared on the e-learning platform

- **Suggested readings** will be uploaded on the e-learning platform

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*."

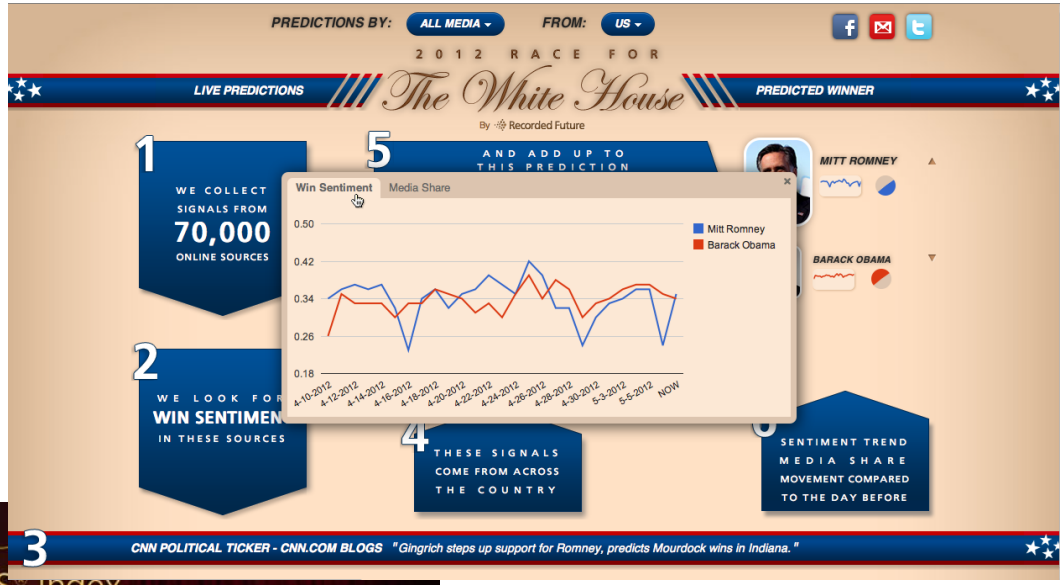
- *"…the phrase "text mining" appears 17 times as often as "text data mining" on the Web, according to a popular search engine (and "data mining" occurs 500 times as often*)."

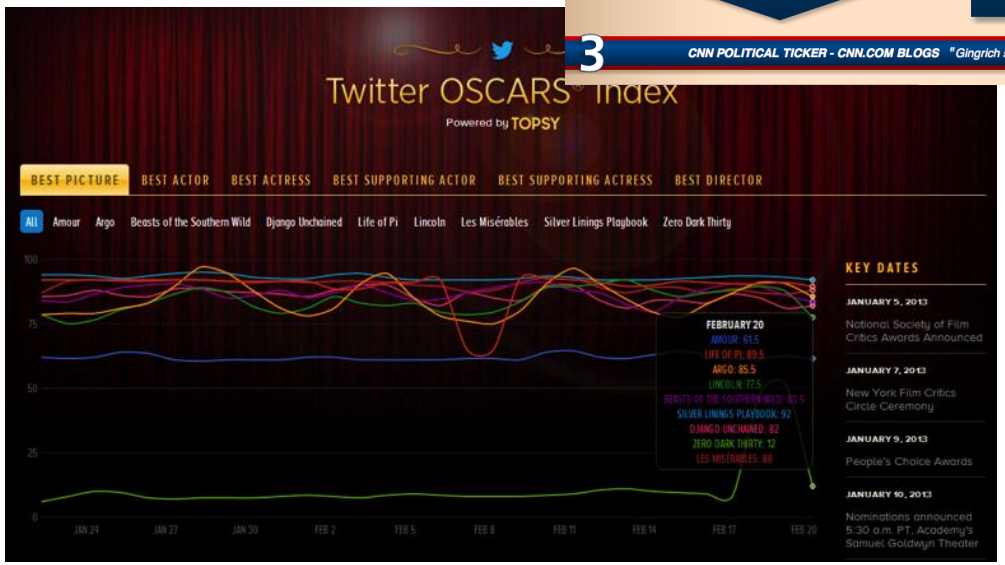
Text Mining

- 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 (1)

- Sentiment analysis



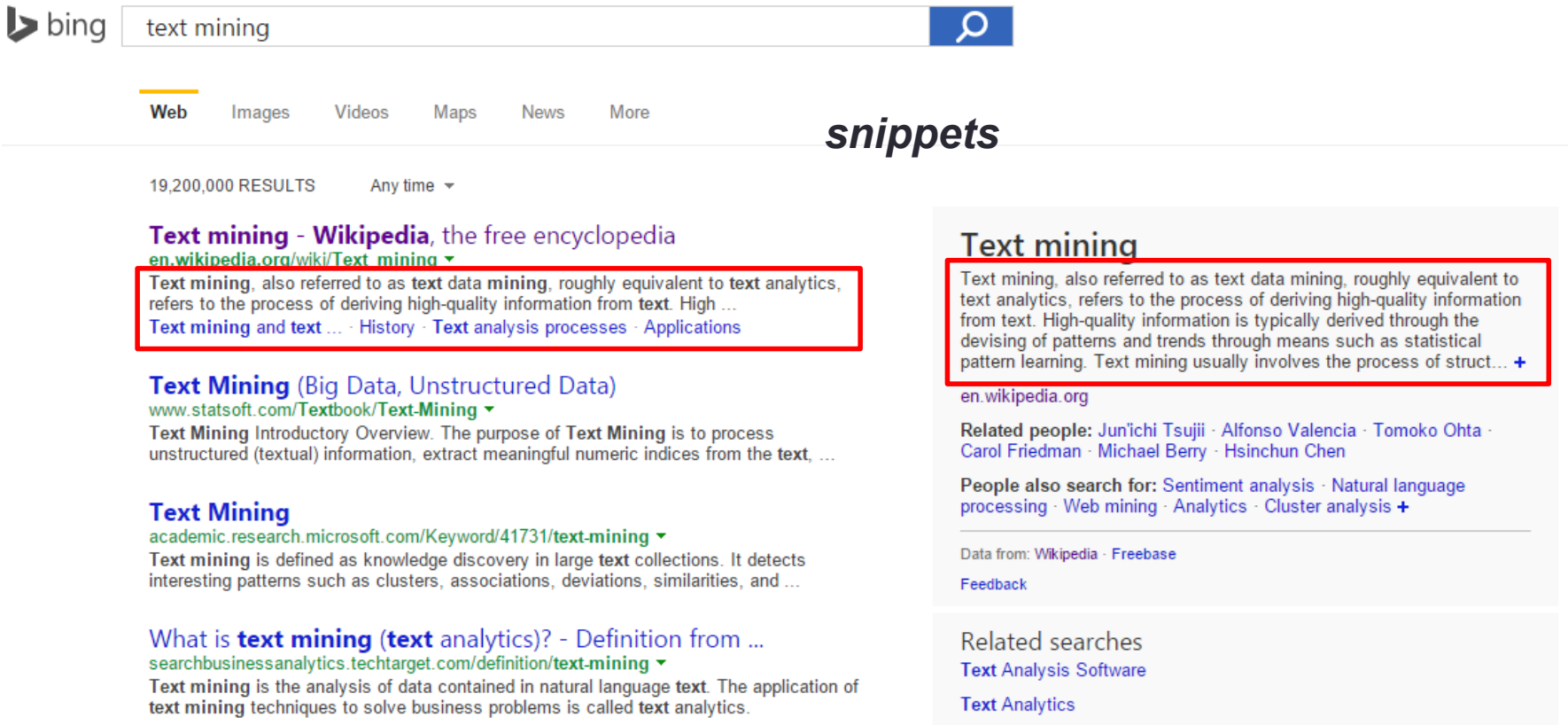


Text Mining is Around Us (2)

- Document summarization

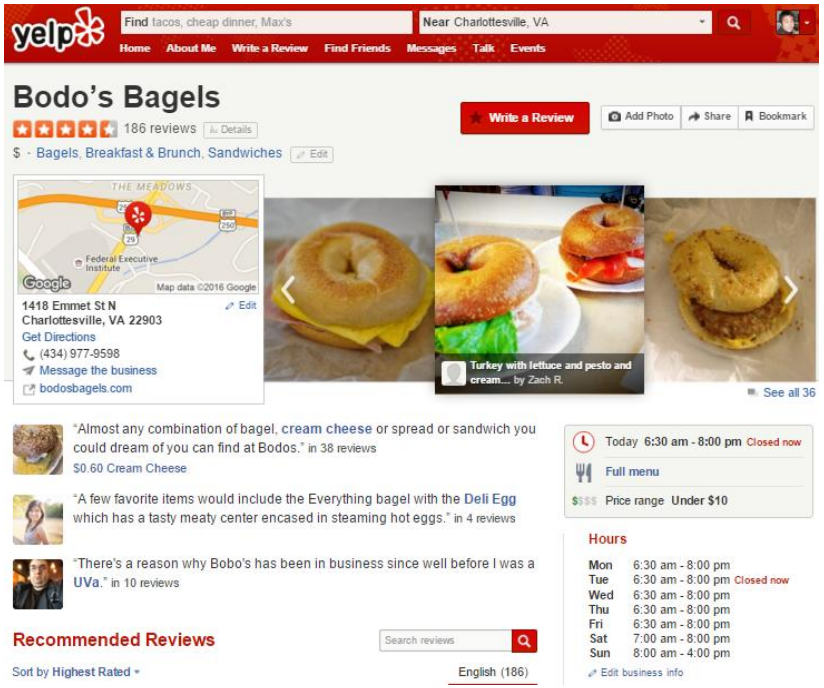
Text Mining is Around Us (3)

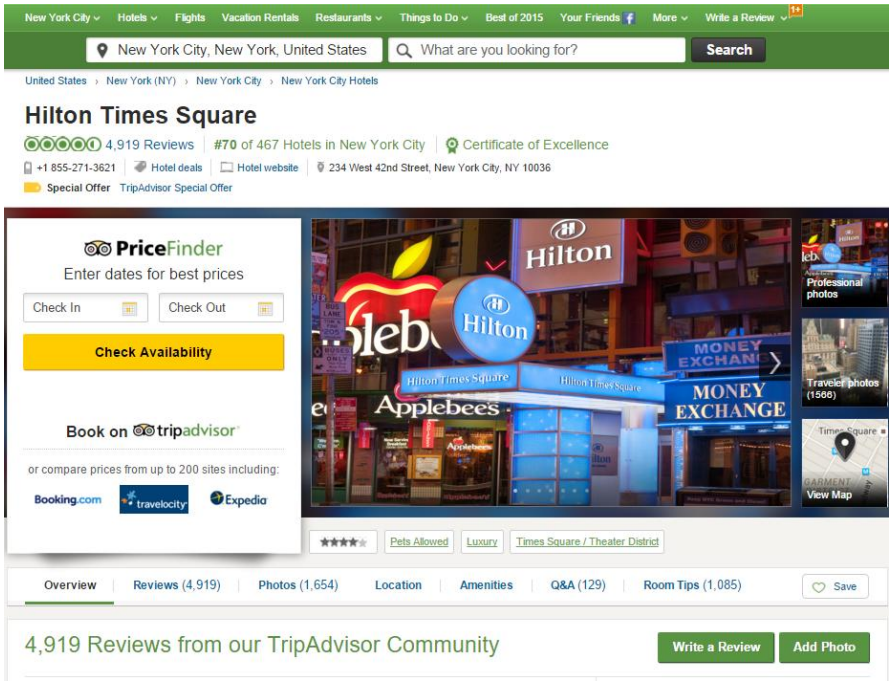
- Document summarization



Text Mining is Around Us (4)

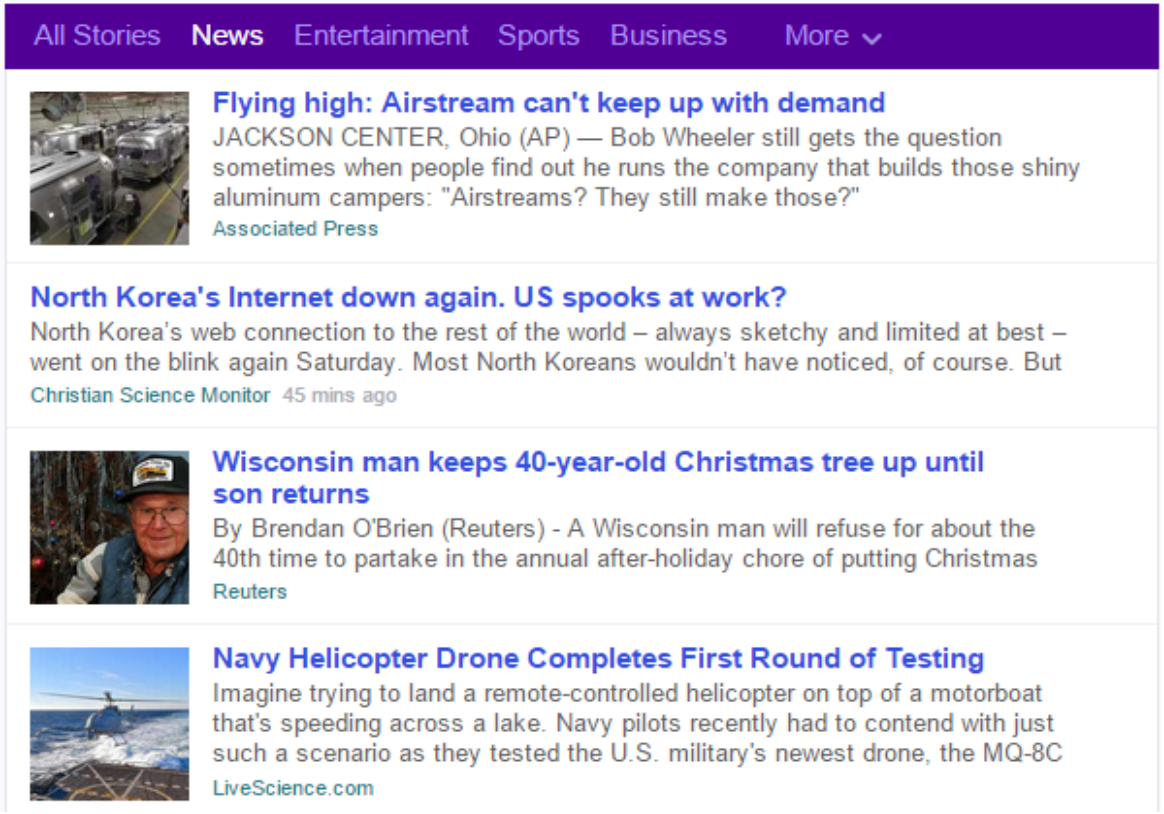
- Restaurant/hotel recommendation





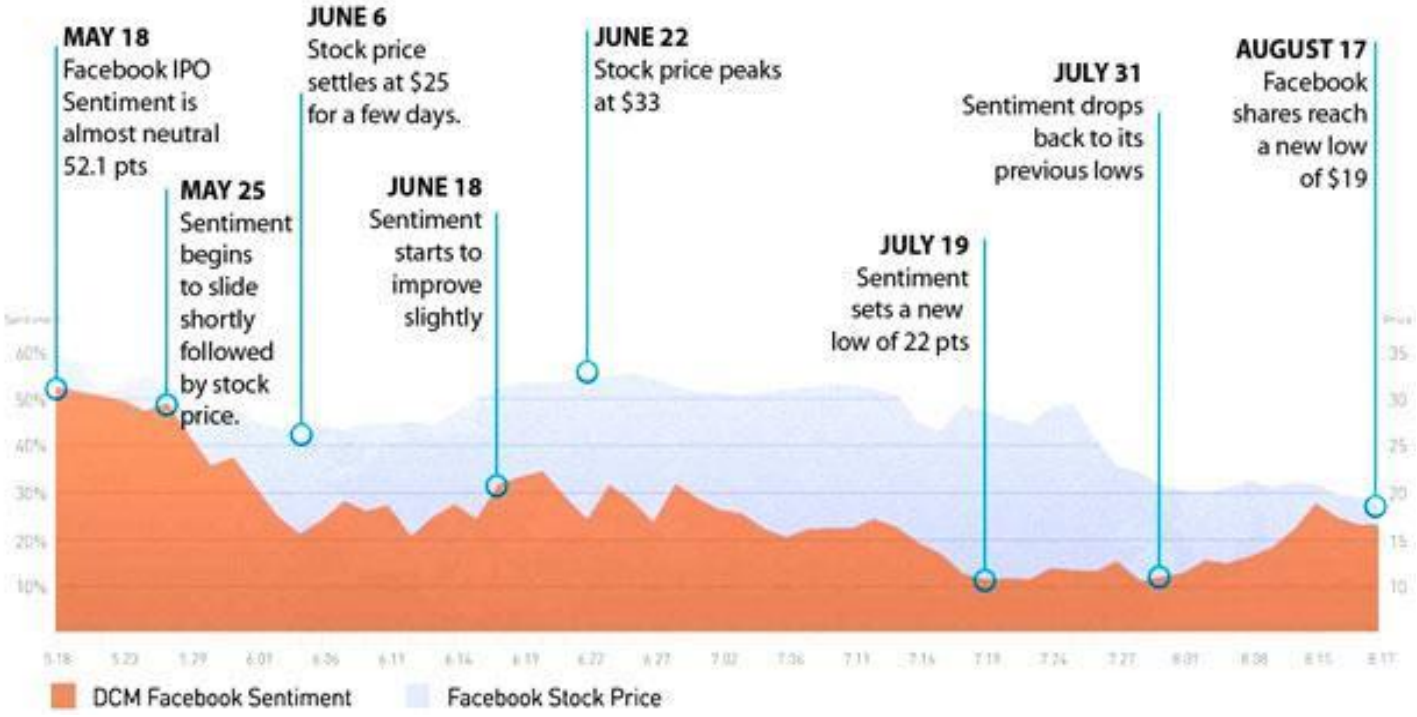
Text Mining is Around Us (5)

- News recommendation



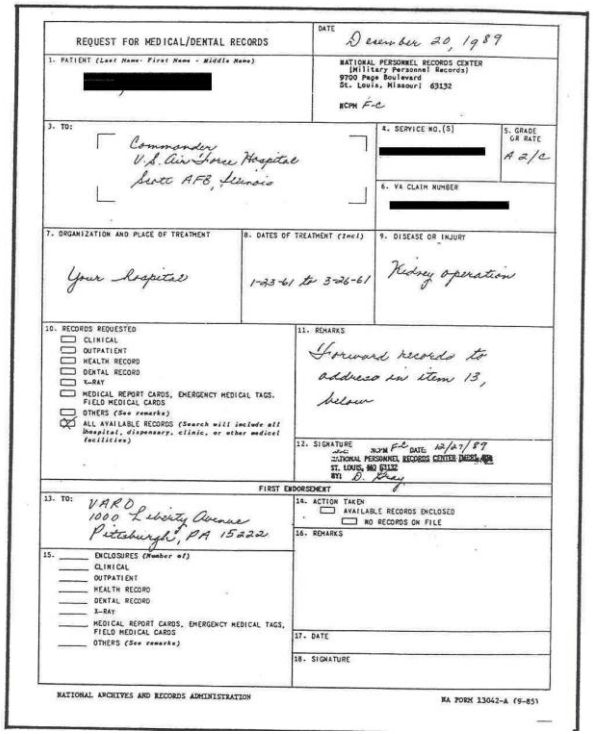
Text Mining is Around Us (6)

- Text analytics in financial services



Text Mining is Around Us (7)

- 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 (MPR), AF
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

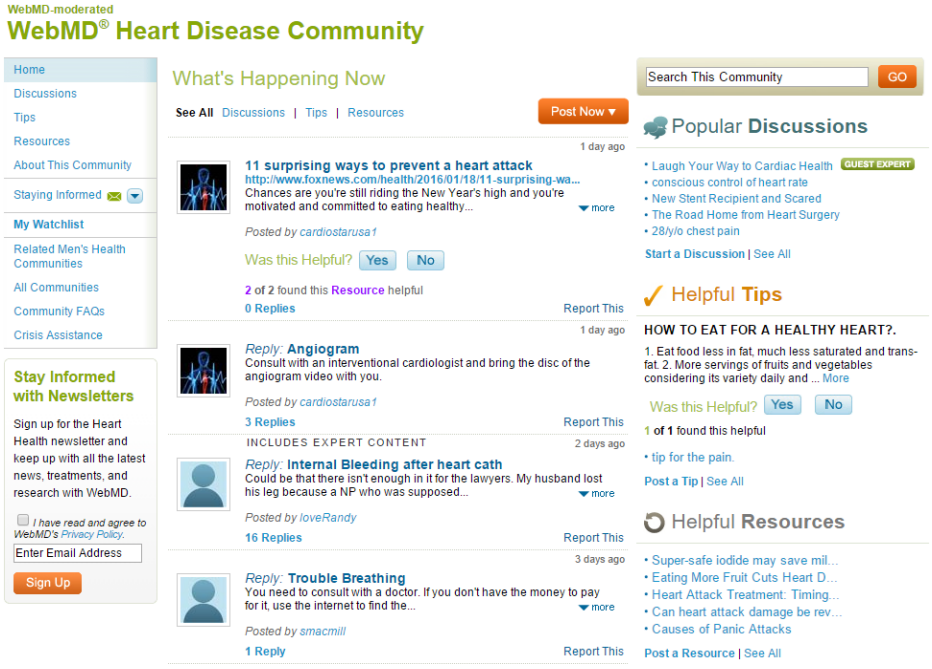
16. REMARKS

15. ____ ENCLOSURES (Number of)
- ____ CLINICAL
- ____ OUTPATIENT
- ____ HEALTH RECORD
- ____ DENTAL RECORD
- ____ X-RAY
- ____ MEDICAL REPORT CARDS, EMERGENCY MEDICAL TAGS, FIELD MEDICAL CARDS
- ____ OTHERS (See 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

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

Post 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

Search This Community GO

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

Text Mining and Data Mining (1)

- **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.

- In **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.

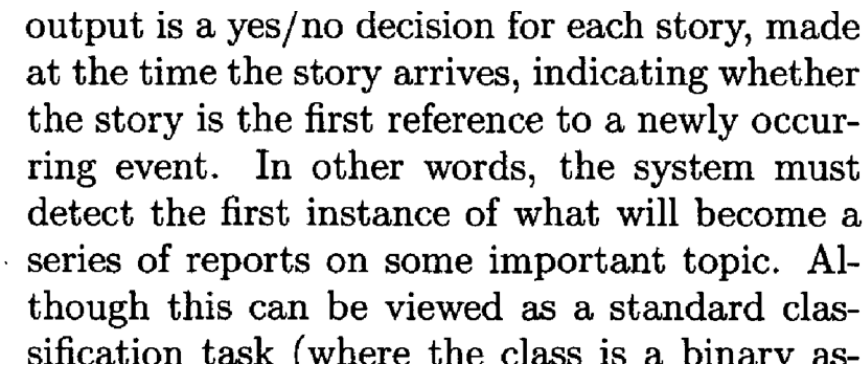
Text Mining and Data Mining (2)

- "*Mining implies extracting precious nuggets of ore from otherwise worthless rock*".

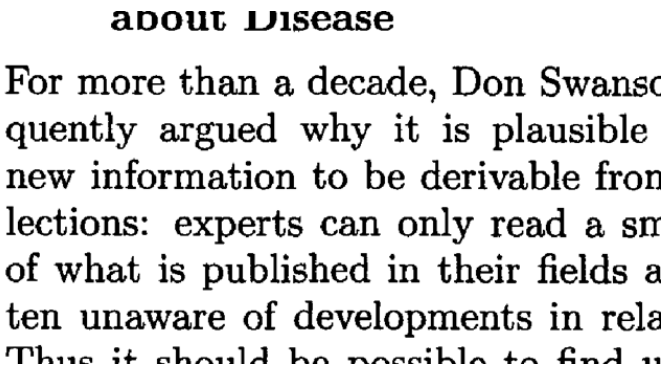
- *If data mining really followed this metaphor, it would mean that people were discovering new factoids within their inventory databases. However, in practice this is not really the case. Instead, data mining applications tend to be (semi)automated discovery of trends and patterns across very large 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 (3)



output is a yes/no decision for each story, made at the time the story arrives, indicating whether the story is the first reference to a newly occurring event. In other words, the system must detect the first instance of what will become a series of reports on some important topic. Although this can be viewed as a standard classification task (where the class is a binary as-



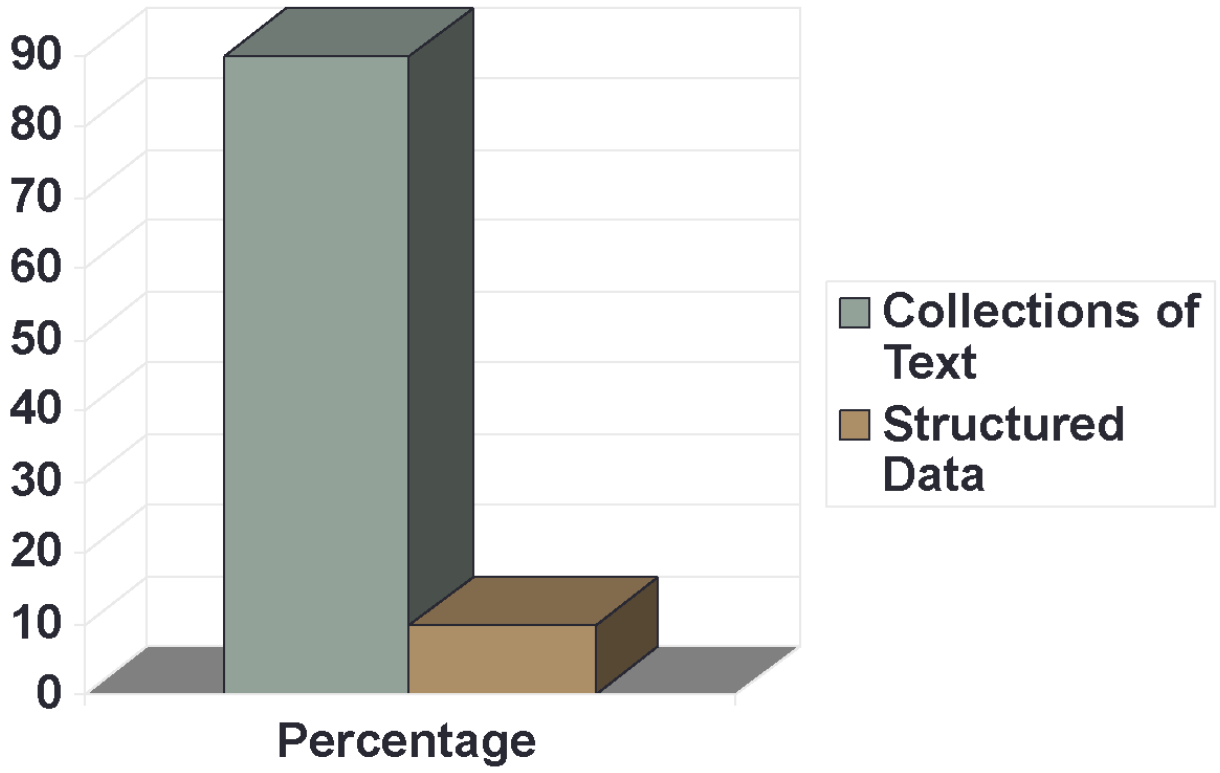
about Disease

For more than a decade, Don Swansc quently argued why it is plausible new information to be derivable fron lections: experts can only read a sm of what is published in their fields a ten unaware of developments in rela Thus it should be possible to find u

Information Retrieval was founded well before the appearence of the Expression Text Mining

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

Interest inText 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 (IR)

- 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 (1)

- **Text Classification** (or text categorization): 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)
- **Text 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 modeling**: identification and tracking of the main topics in a document collection

Tasks Affected by Text Mining (2)

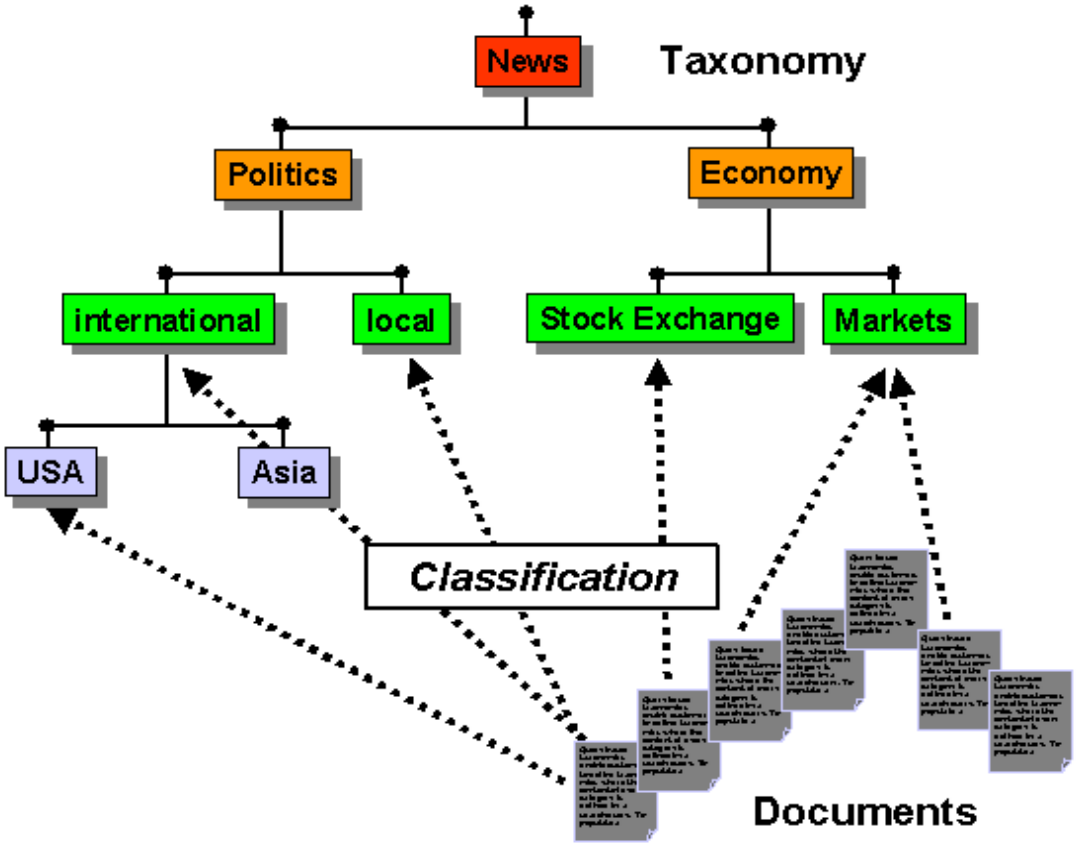
- **Text Summarization**: A text summarizer strives to produce a condensed representation of its input, intended for human consumption

- **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

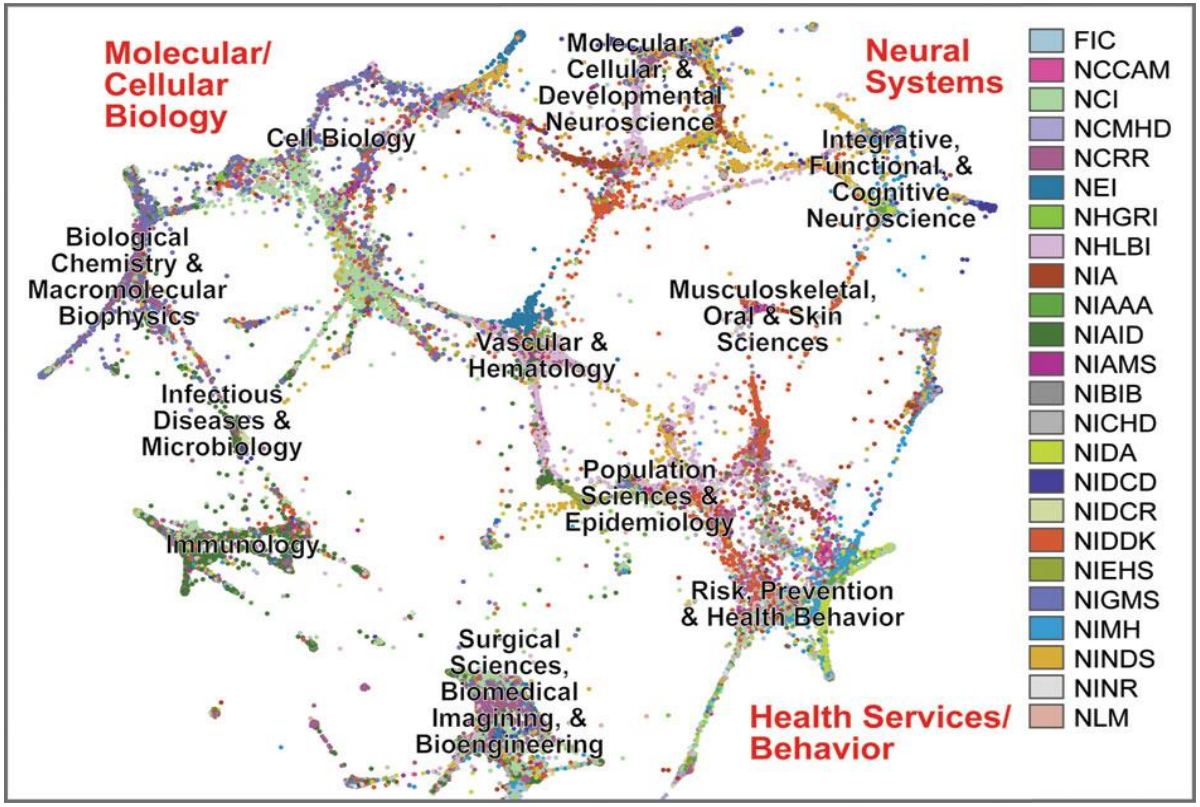
Text Classification

- Possible application: adding structure to the text corpus



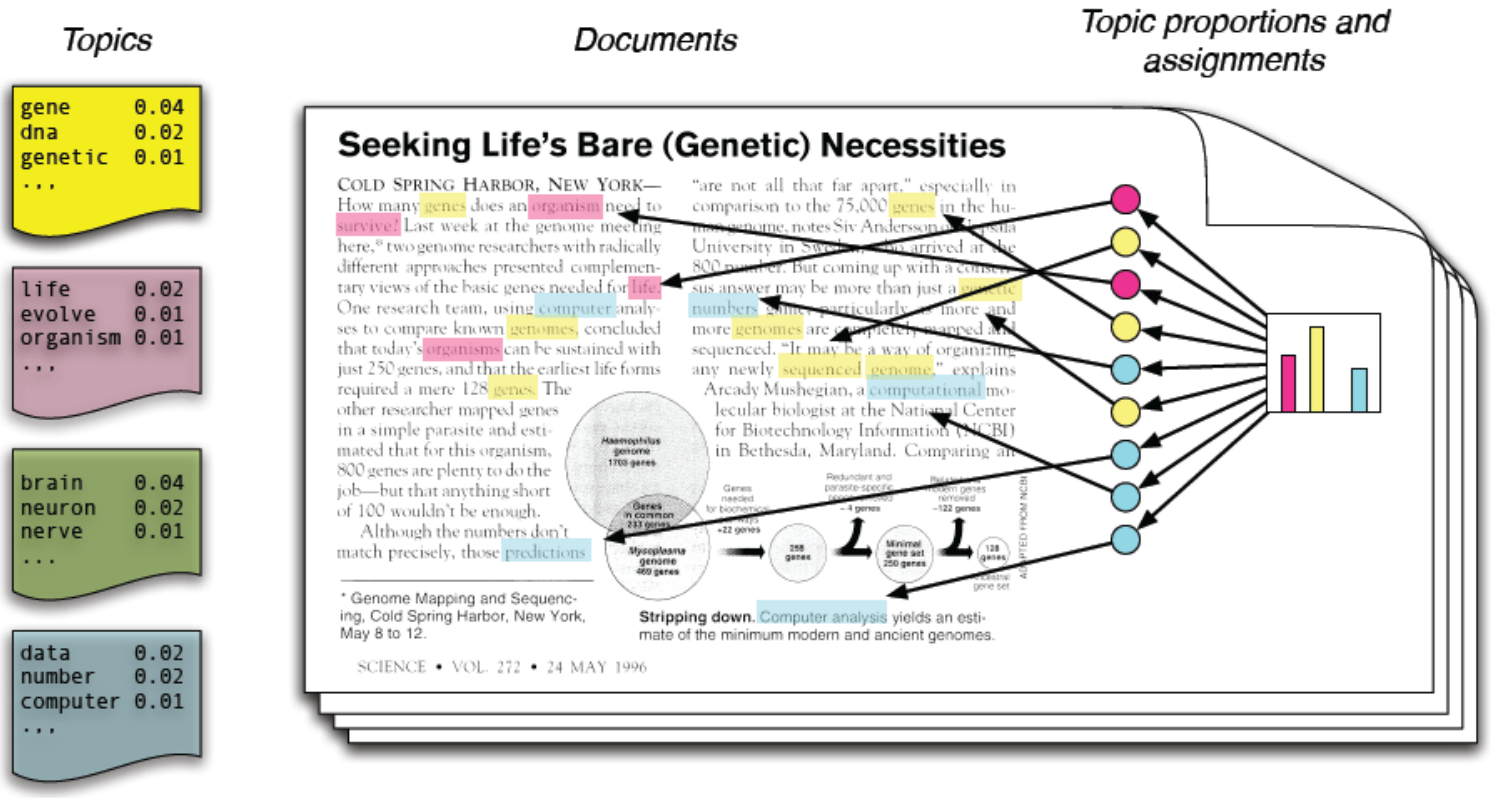
Text Clustering

- Possible application: identifying structures in a text corpus



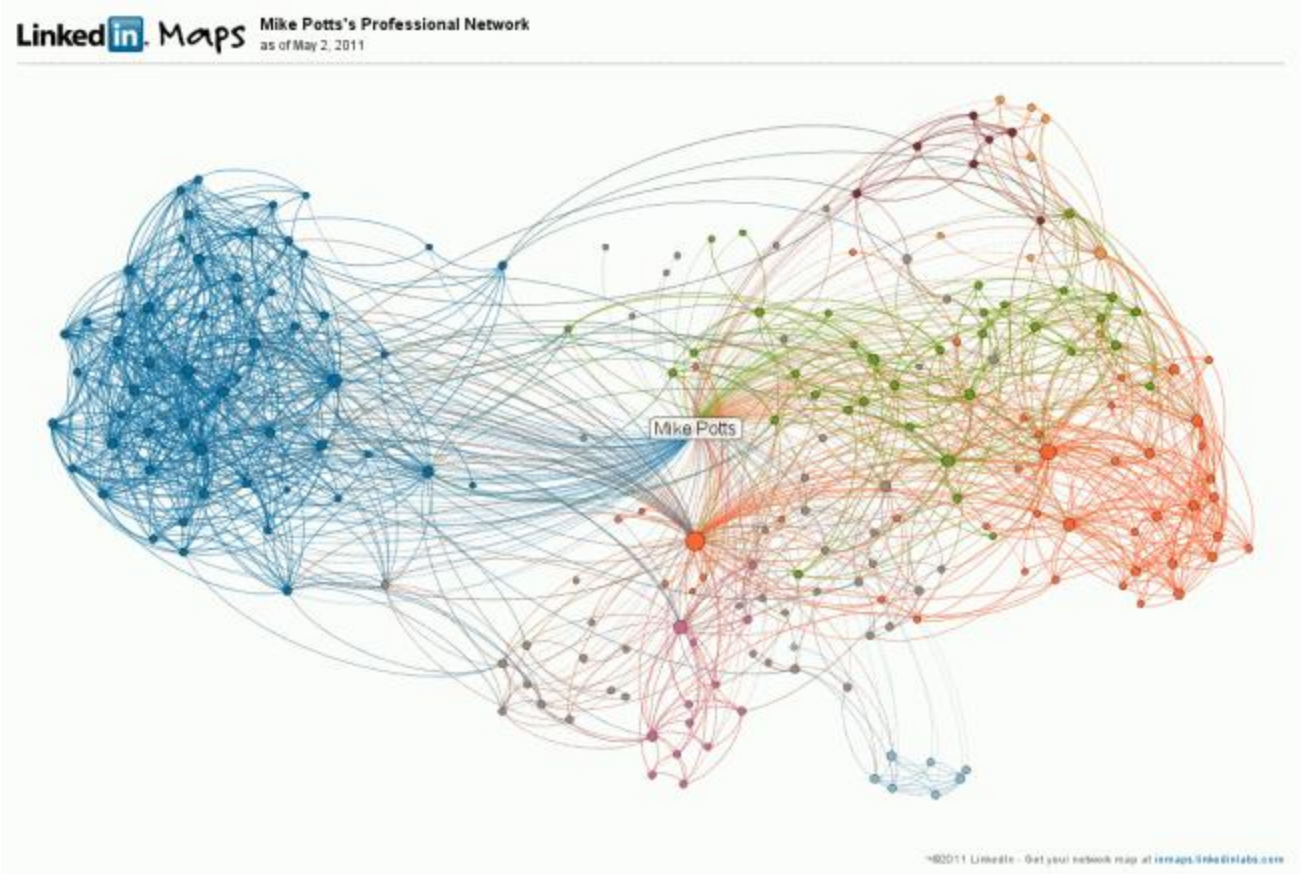
Topic Modeling

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



Social Media Analytics

- Exploring additional structure in the text corpus



Texts: Mining Structured Information

- **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/Exploratory Analysis of Texts

- **Predictive Analysis of Texts**
  - developing computer programs that automatically recognize or detect *a particular concept* within a span of text.

- **Exploratory Analysis of Texts**
  - developing computer programs that automatically discover *interesting and useful patterns or trends* in text collections.

Predictive Analysis of Texts: Examples

- *Opinion Mining*/*Sentiment Analysis*
  - automatically detecting whether a span of opinionated text expresses a positive or negative opinion about the item being judged

- *Emotion Detection (incl. Affective Computing)*
  - 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." ???

Emotion Detection: Social Media

“[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)

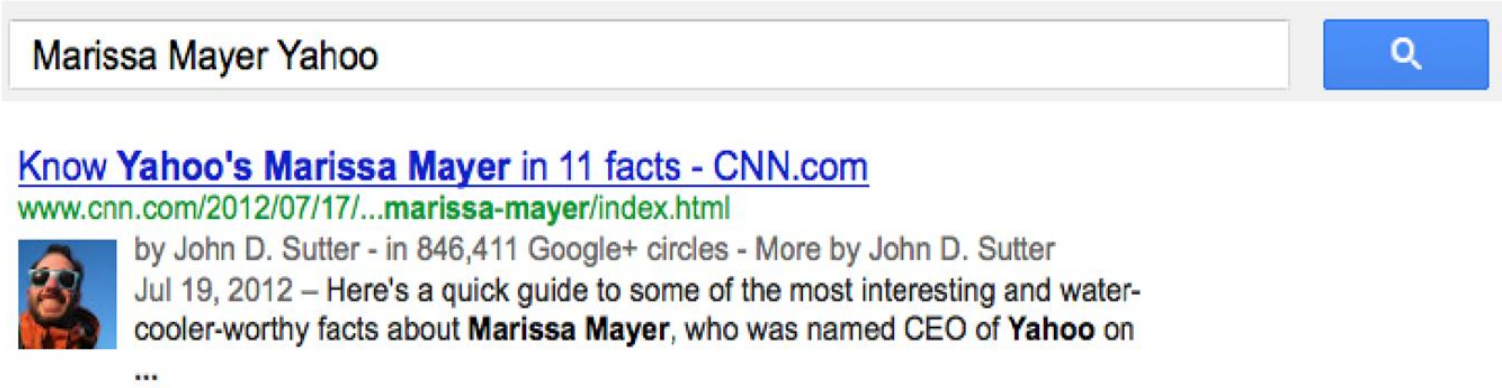
Predictive Analysis of Texts: Examples

- *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>)

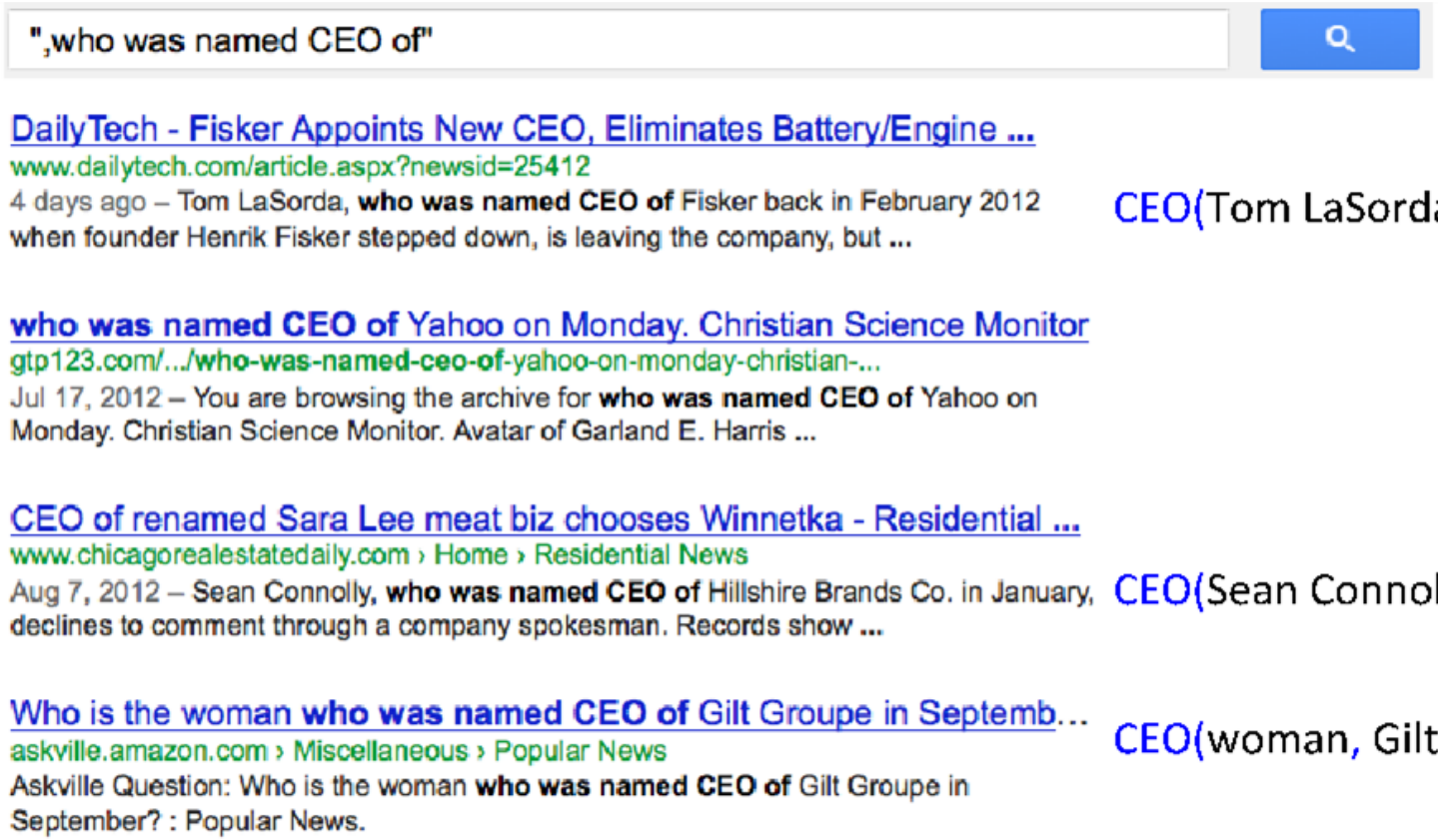
    …

Relation Learning: CEO(<person>,<company>) (1)



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

Relation Learning: CEO(<person>,<company>) (2)



CEO(Tom LaSorda, Fisker)

CEO(Sean Connolly, Hillshire Brands)

CEO(woman, Gilt Groupe)

Predictive Analysis of Texts: Examples

- 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 Texts

- Text clustering
  - The document collection can be constituted of distinct documents talking about distinct research areas
- Topic modeling
  - Each document can be characterized by one or more topics, which are extracted via topic modeling and represented as sets of words associated with the topic
- …

Mining Structured Text

- Several Web resources have a structure: for example Web pages are written is 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 requisite information.

Welcome to the Text Mining Course!