Data Mining on Social Networks

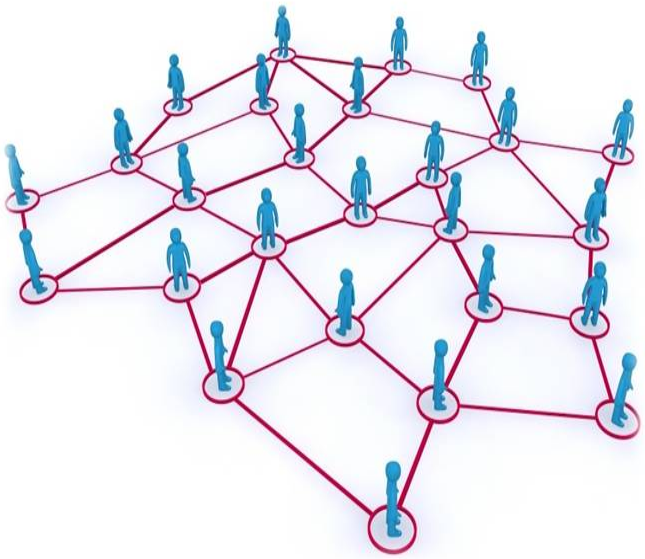
Dionysios Sotiropoulos Ph.D.

Contents

What are Social Media?

A **social network** is defined as a network of interactions or relationships, where the nodes consist of actors, and the edges consist of the relationships or interactions between these actors.

Milgram in the sixties (well before the invention of the internet), hypothesized the likelihood that any pair of actors on the planet are separated by at most six degrees of separation. This is also referred to as the small world phenomenon.



Any web-site or application which provides a social experience in the form of user-interactions can be considered to be a form of social network.

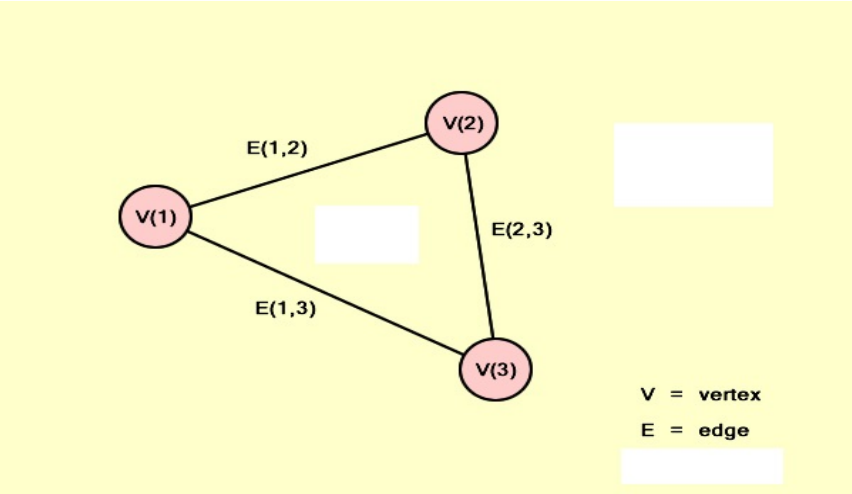
Social networks can be viewed as a structure which enables the dissemination of information.

The analysis of the dynamics of such interaction is a challenging problem in the field of social networks.

Mathematical Representation of Social Networks

Social networks employ the typical **graph notation** $G = (V, E)$ where $\boldsymbol{G}$ stands for the **whole network**, V stands for the **set of all vertices** and E stands for the **set of all edges** .

Example: $G = (V, E)$ where $\mathrm{V} = \{v_1, v_2, v_3\}$ and $E = \{e_{12}, e_{13}, e_{23}\}$



Why Mine Data? Commercial Viewpoint

Computers have become cheaper and more powerful

Lots of data is being collected and warehoused

Competitive Pressure is Strong

Provide better, customized services for an edge (e.g. in Customer Relationship Management)

Why Mine Data? Scientific Viewpoint

Data collected and stored at enormous speeds (GB/hour)

- remote sensors on a satellite
- telescopes scanning the skies
- microarrays generating gene expression data
- scientific simulations generating terabytes of data

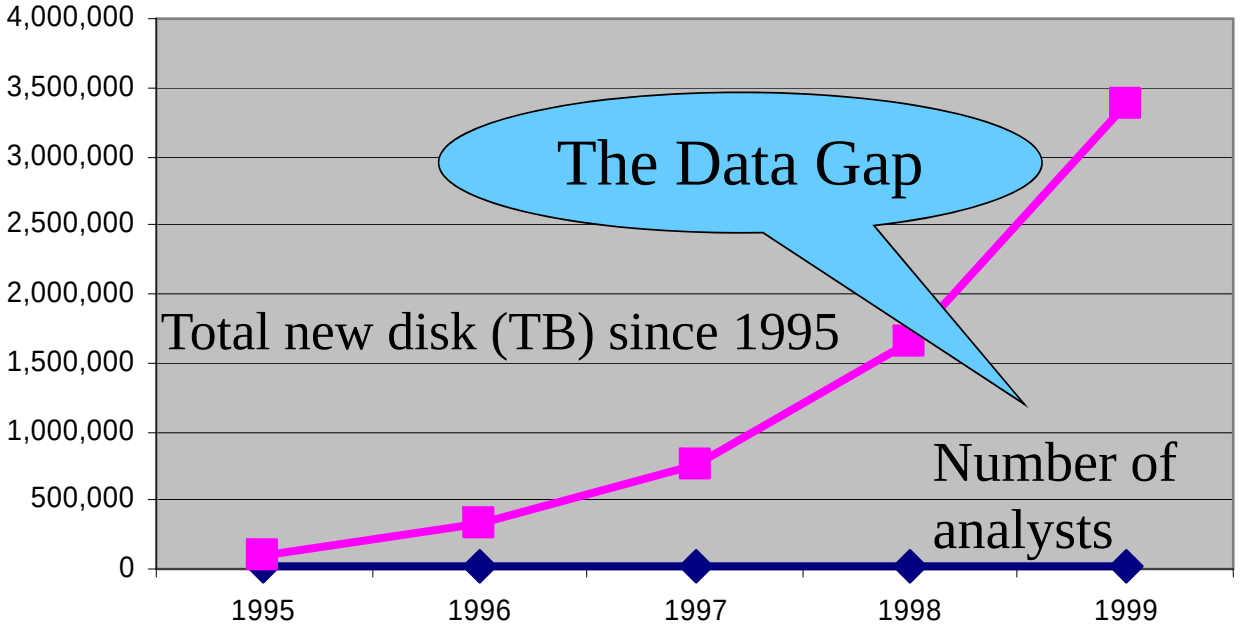
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation

Mining Large Data Sets – Motivation

There is often information "hidden" in the data that is not readily evident

Human analysts may take weeks to discover useful information

Much of the data is never analyzed at all

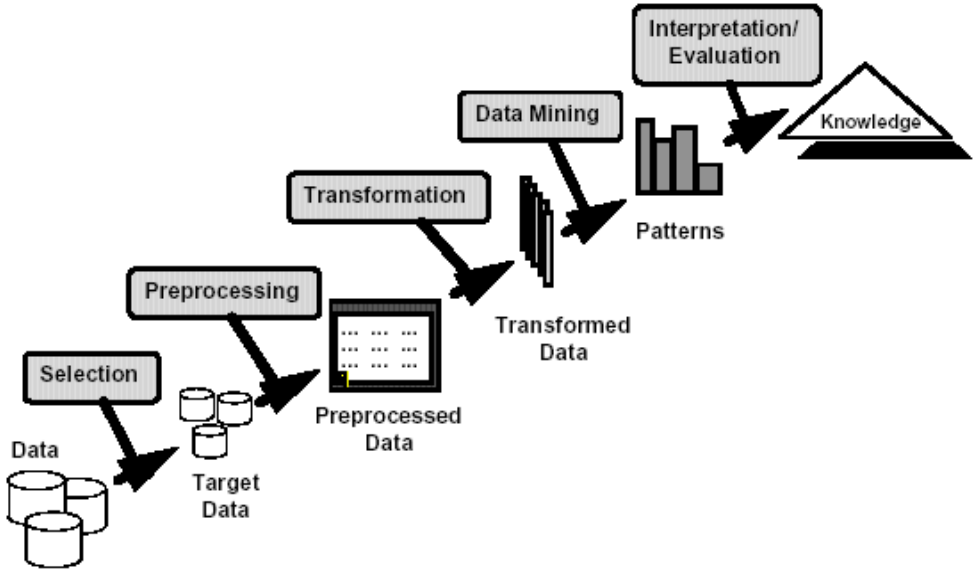


Fundamental Data Mining Concepts IV

What is Data Mining (Many Definitions)

Non-trivial extraction of implicit, previously unknown and potentially useful information from data.

Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns.



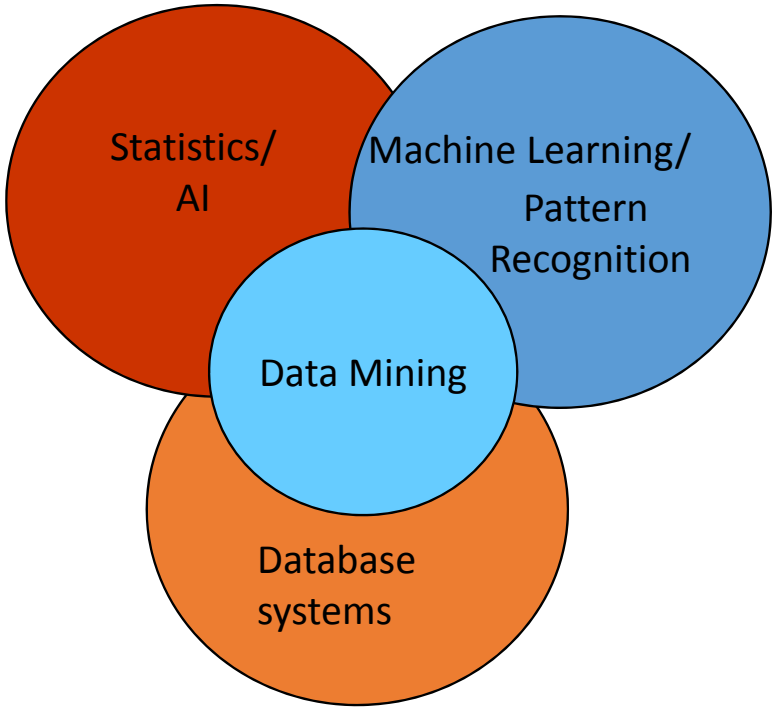
Fundamental Data Mining Concepts V

Origins of Data Mining:

Draws ideas from **Machine Learning** / **AI**, **Pattern Recognition**, **Statistics**, and **Database Systems**

Traditional Techniques may be unsuitable due to:

- **Enormity** of data
- **High dimensionality** of data
- **Heterogeneous**, **distributed nature** of data



Fundamental Data Mining Concepts VI

Prediction Methods

Use some variables to predict unknown or future values of other variables.

Description Methods

Find human-interpretable patterns that describe the data.

Classification [Predictive]

Clustering [Descriptive]

Association Rule Discovery [Descriptive]

Sequential Pattern Discovery [Descriptive]

Regression [Predictive]

Deviation Detection [Predictive]

Fundamental Data Mining Concepts VII

Classification: Definition

Given a collection of records (training set)

Each record contains a **set of attributes**, one of the attributes is the **class**.

Find a model for the class attribute as a function of the values of other attributes.

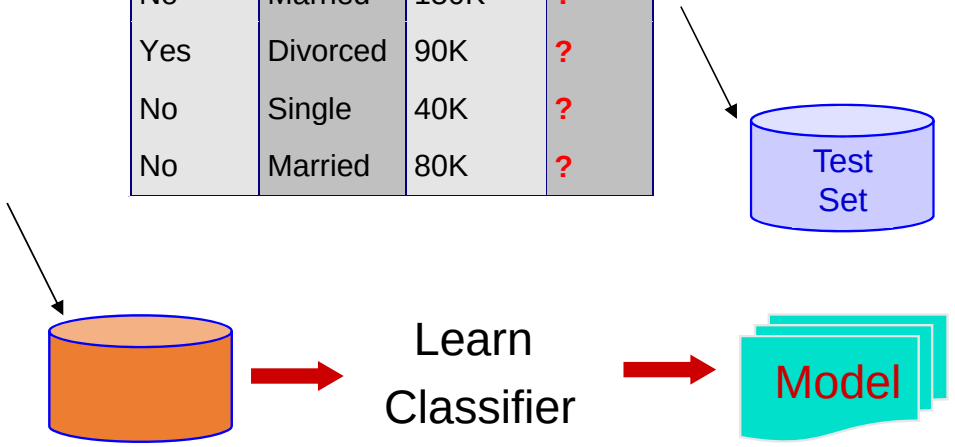
Goal: previously unseen records should be assigned a class as accurately as possible.

A test set is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it

Classification: Example

Tid	Refund	Marital Status	Taxable Income	Cheat
				(categorical / categorical / continuous / class)
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Fundamental Data Mining Concepts IX

Classification: Example (Direct Marketing)

Goal: Reduce cost of mailing by targeting a set of consumers likely to buy a new cell-phone product.

Approach:

Use the **data** for a **similar product** introduced **before**.

We **know** which **customers decided** to **buy** and which decided **otherwise**. This {**buy**, **don't buy**} decision forms the **class** attribute.

Collect various **demographic**, **lifestyle**, and **company-interaction** related information about all such **customers**.

Use this information as **input attributes** to **learn** a **classifier** model

Clustering: Definition

Given a **set of data points**, each having **a set of attributes**, and a **similarity measure** among them, **find clusters** such that:

- Data points in one cluster are more similar to one another.
- Data points in separate clusters are less similar to one another.

Similarity Measures:

- Euclidean Distance / Cosine Similarity if attributes are continuous.
- Other Problem-specific Measures.

Fundamental Data Mining Concepts XI

Clustering: Example (Market Segmentation)

Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.

Approach:

- Collect different attributes of customers based on their geographical and lifestyle related information.
- Find clusters of similar customers.
- Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Fundamental Data Mining Concepts XII

Regression: Definition

Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.

Greatly studied in statistics, neural network fields.

Regression Examples:

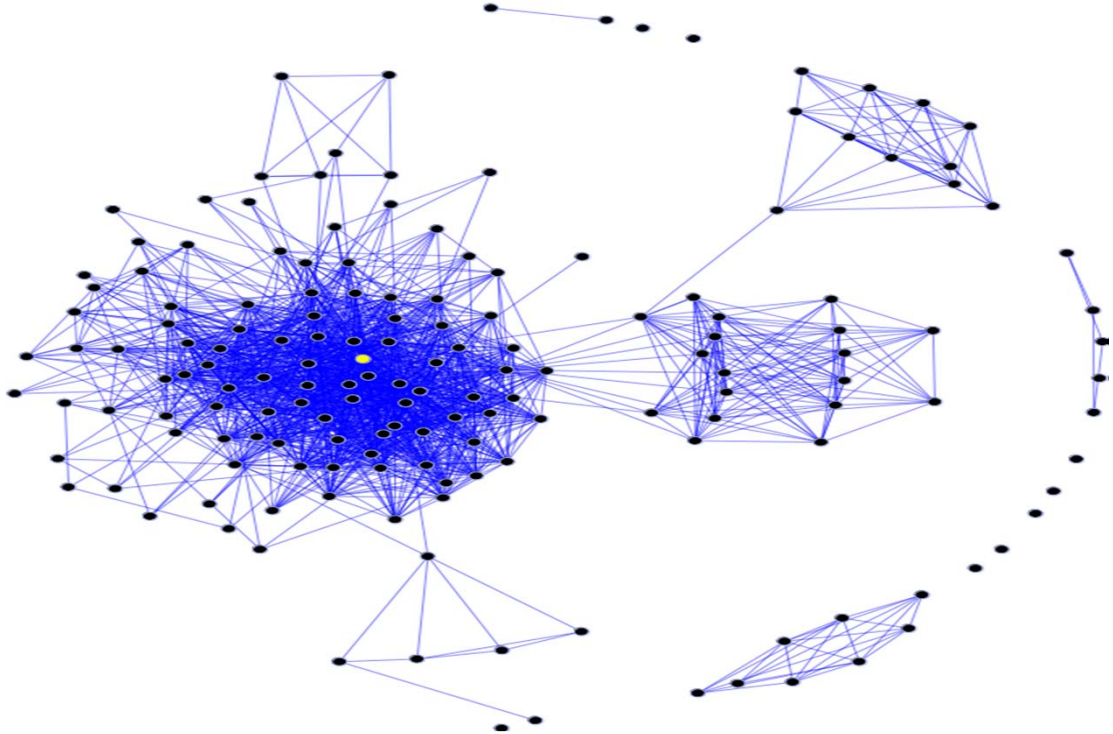
- **Predicting** sales amounts of new product based on advertising expenditure.
- **Predicting** wind velocities as a function of temperature, humidity, air pressure, etc.
- Time series **prediction** of stock market indices

Data Mining Tasks on Digital Social Networks: Community Detection I

The notion of community denotes groups of people with shared interests or activities.

- political biases
- voting patterns
- consuming habits

Community Detection (Definition):
is the task of discovering inherent community structures or clusters within online social networks.



Structural Definition:

Communities as cliques

Data Mining Tasks on Digital Social Networks: Community Detection II

Core Methods :

- Quality Functions
- The Kernighan-Lin(KL) algorithm
- Agglomerative/Divisive Algorithms
- Spectral Algorithms
- Multi-level Graph Partitioning
- Markov Clustering

Challenges :

- Community Discovery in Dynamic Networks
- Community Discovery in Heterogeneous Networks
- Coupling Content and Relationship Information for Community Discovery

Data Mining Tasks on Digital Social Networks: Evolution Analysis

"Informally, evolution refers to a change that manifests itself across the time axis."

Communities can be perceived as clusters built at each time-point: analysis of community evolution involves tracing the same community/cluster at consecutive time-points and identifying changes.

Communities can also be perceived as smoothly evolving constellations: community monitoring then involves learning models that adapt smoothly from one time-point to the next.

Data Mining Tasks on Digital Social Networks: Evolution Analysis II

Challenges :

- Incremental Mining for Community Tracing
- Tracing Smoothly Evolving Communities
- Verifying that the communities discovered by a learning algorithm are indeed the real ones
- Verifying that the community evolution patterns detected are realistic, i.e. they conform to prior knowledge or can be verified by inspection of the underlying data

Data Mining Tasks on Digital Social Networks: Social Influence Analysis I

Social influence is an **intuitive** and **well-accepted phenomenon** in social networks[D. Easley and J. Kleinberg].

The **strength** of **social influence** depends on **many factors** such as:

- the **strength** of **relationships between people** in the **networks**
- the **network distance between users**
- **temporal effects**
- **characteristics** of **networks** and **individuals** in the **network**.

"A central problem for social influence is to understand the interplay between similarity and social ties" [D. Easley and J. Kleinberg]

Data Mining Tasks on Digital Social Networks: Social Influence Analysis II

Homophily is one of the most **fundamental characteristics** of **social networks**.

This suggests that an actor in the social network tends to be similar to their connected neighbors or "friends".

The **phenomenon** of **homophily** can **originate** from many **different mechanisms**:

Social influence: This indicates that **people tend** to **follow** the **behaviors** of their **friends**.

Selection: This indicates that **people tend** to **create relationships** with **other people** who are **already similar** to them;

Confounding variables: Other **unknown variables** exist, which **may cause friends to behave similarly with one another**.

Challenges :

Influence Maximization

Data Mining Tasks on Digital Social Networks: Link Inference / Prediction I

Examine :

Node Neighborhood based Features.

In simple words, it means that in social networks if vertex x is connected to vertex z and vertex y is connected to vertex z, then there is a heightened probability that vertex x will also be connected to vertex y.

Shortest Path Distance.

The fact that the friends of a friend can become a friend suggests that the path distance between two nodes in a social network can influence the formation of a link between them. The shorter the distance, the higher the chance that it could happen.

Examine :

Hitting Time.

The concept of hitting time comes from random walks on a graph. For two vertices, x and y in a graph, the hitting time, $H(x, y)$ defines the expected number of steps required for a random walk starting at x to reach y.

Data Mining Tasks on Digital Social Networks: Machine Learning

The **fundamental assumption** of **independence** is **not valid** when **learning** within the **context** of **Social Networks**.

Social network data sets are often called **"relational"** since the **relations among entities** are **central**, e.g. :

- **friendship** or **following ties** between members of the social network.
- **interactions** such as **wall posts**, **private messages**, **re-tweeting** or **tagging a photo**.

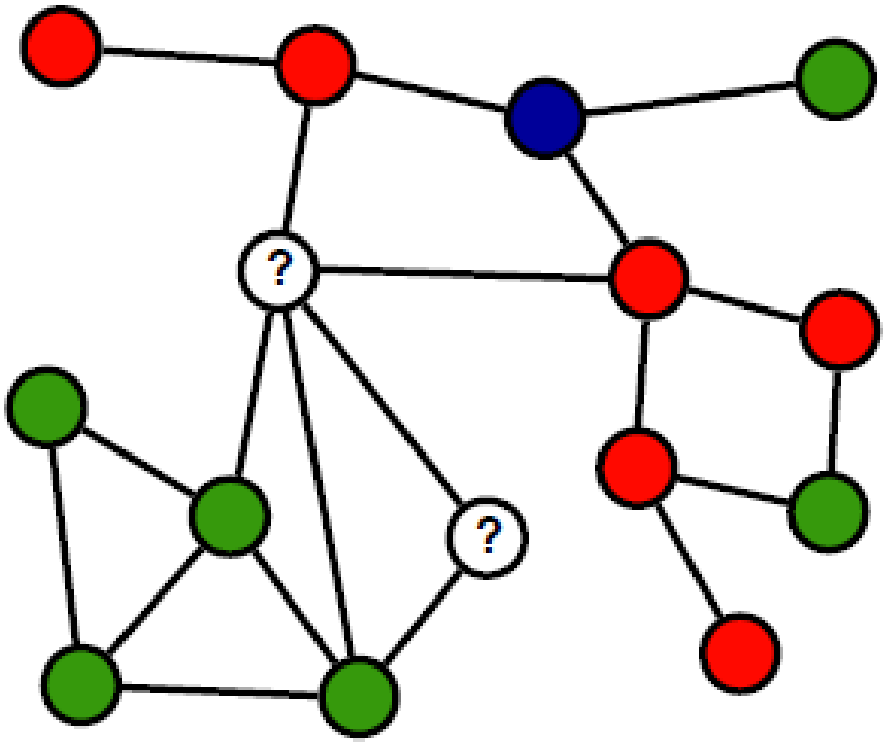
Traditional machine learning algorithms are not directly applicable on relational data sets.

Data Mining Tasks on Digital Social Networks: Node Classification

When dealing with large graphs, such as those arising in the context of social networks, a subset of nodes may be labeled.

Existing labels can indicate demographic values, interest or other characteristics of the nodes (users).

A core problem is to use this information in order to extend the labeling so that all nodes are assigned with a label.



Data Mining Tasks on Digital Social Networks: Text Mining I

Text mining: seeks to extract useful information from data sources through the identification and exploration of interesting patterns.

In the case of text mining, however, the data sources are document collections, and interesting patterns are found not among formalized database records but in the unstructured textual data in the documents in these collections.

Document Features:

Text mining preprocessing operations attempt to transform a natural language document from an irregular and implicitly structured representation into an explicitly structure representation.

An essential task for most text mining systems is the identification of a simplified subset of document features that can be used to represent a particular document as a whole.

This set of features constitutes the representational model of a document .

Data Mining Tasks on Digital Social Networks: Natural Language Processing

Natural Language Processing is an essential preprocessing step for both Topic Modeling and Sentiment Analysis.

NLP Preprocessing Operations:

- Tokenization
- Stop-word Removal
- Stemming
- Lemmatization
- Translation
- Special Characters Replacement
- Repeating Characters Removal
- Spelling Correction
- Synonyms Replacement
- Replacing Negations with Antonyms

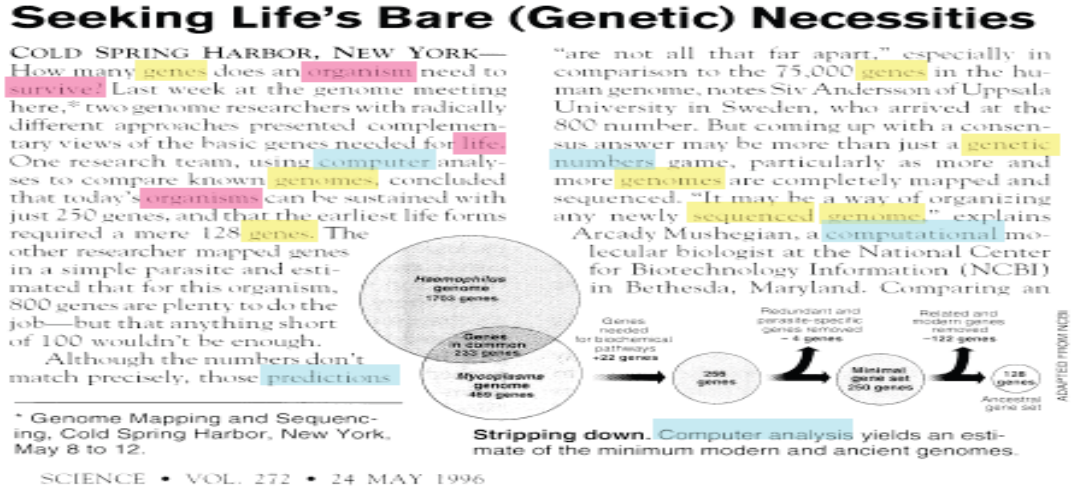
Data Mining Tasks on Digital Social Networks: Topic Modeling I

Topic modeling provides methods for automatically organizing, understanding, searching, and summarizing large electronic archives.

Discover the hidden themes that pervade the collection.

Annotate the documents according to those themes.

Use annotations to organize, summarize, and search the texts.



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

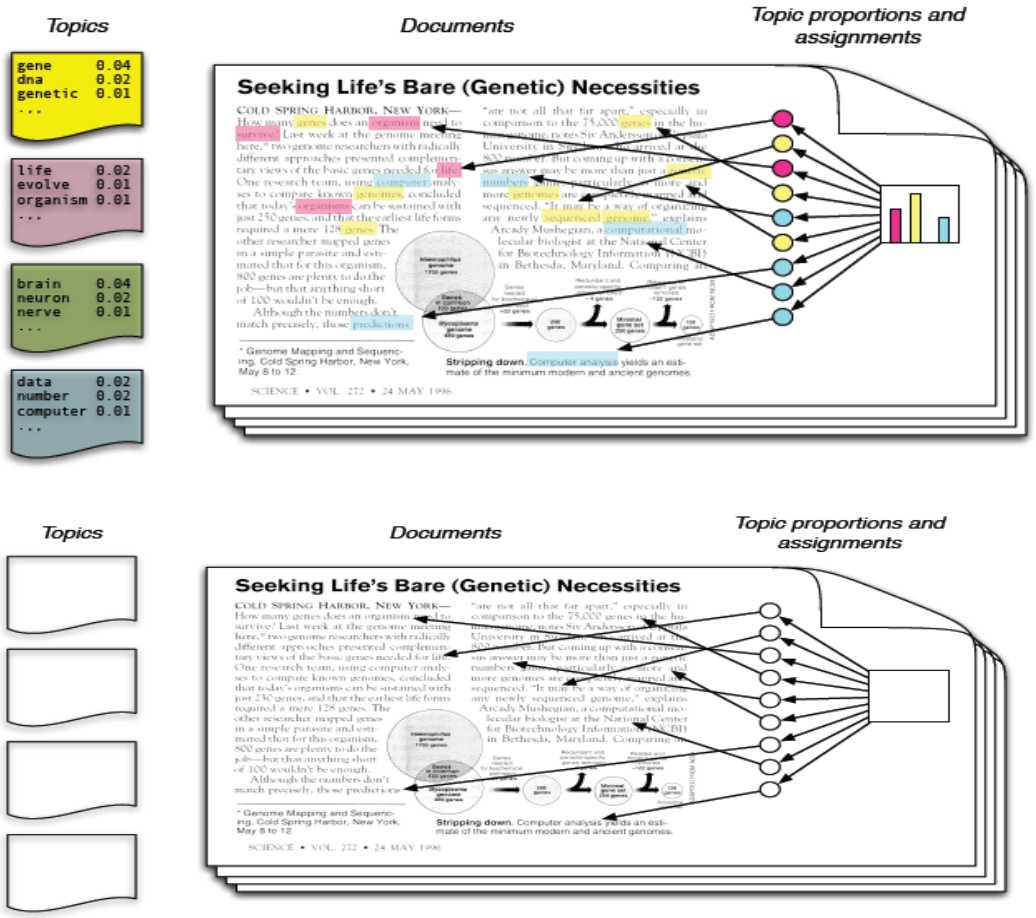
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

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

SCIENCE • VOL. 272 • 24 MAY 1996

Simple Intuition: Documents exhibit multiple topics.

Data Mining Tasks on Digital Social Networks: Topic Modeling II



Fundamental Assumptions:

Each topic is a distribution over words.

Each document is a mixture of corpus-wide topics.

Each word is drawn from one of those topics.

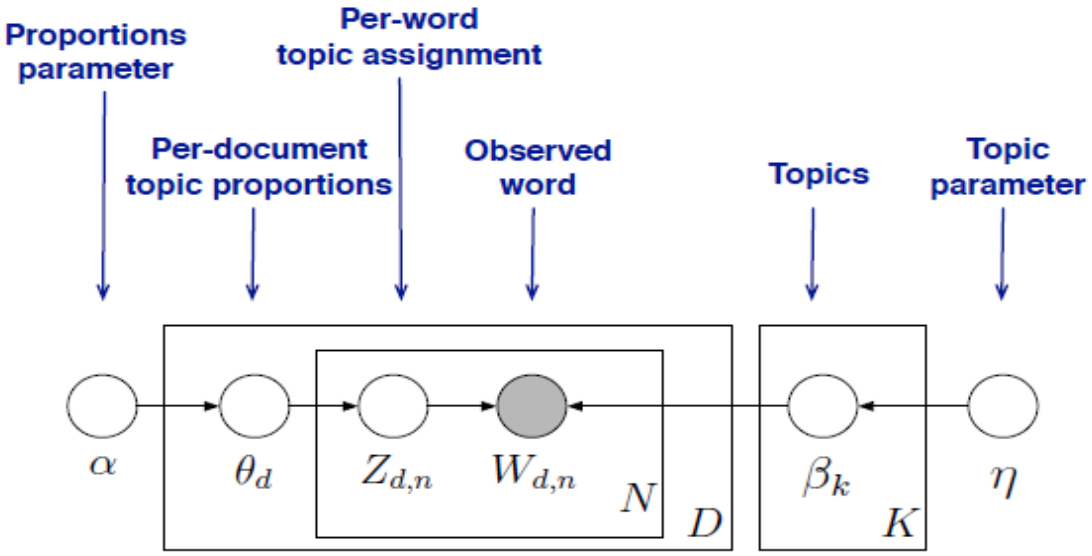
In reality we observe the documents.

The other structure are hidden variables.

Data Mining Tasks on Digital Social Networks: Topic Modeling III

Probabilistic Topic Modeling: the primary objective is to **infer** the **hidden variables**, that is **compute** their **distribution conditioned** on the **documents**.

$$p(topics, proportions, assignments|documents)$$



Nodes are random variables.
Shaded nodes are observed.
Plates indicate replicated variables.

Data Mining Tasks on Digital Social Networks: Sentiment Analysis I

Sentiment Analysis: the use of natural language processing (NLP) and computational techniques to automate the extraction or classification of sentiment from typically unstructured text.

Motivation:

- **Consumer information**
  - Product reviews.
- **Marketing**
  - Consumer attitudes
  - Trends
- **Politics**
  - Politicians want to know voters' views
  - Voters want to know politicians' stances and who else supports them
- **Social**
  - Find like-minded individuals or communities

Data Mining Tasks on Digital Social Networks: Sentiment Analysis II

Related Problems:

- Which features to use?
  - Words (unigrams)
  - Phrases/n-grams
  - Sentences
- How to interpret features for sentiment detection?
  - Bag of words (IR)
  - Annotated lexicons (WordNet, SentiWordNet)
  - Syntactic patterns
  - Paragraph structure

Data Mining Tasks on Digital Social Networks: Sentiment Analysis III

Challenges:

- Harder than topical classification, with which bag of words features perform well
- Must consider other features due to...
  - Subtlety of sentiment expression
  - Irony
  - expression of sentiment using neutral words
- Domain/context dependence
  - words/phrases can mean different things in different contexts and domains
- Effect of syntax on semantics

Data Mining Tasks on Digital Social Networks: Sentiment Analysis IV

Approaches:

- Supervised Methods
  - Naïve Bayes
  - Maximum Entropy Classifier
  - SVM
  - Markov Blanket Classifier
    - Accounts for conditional feature dependencies
    - Allowed reduction of discriminating features from thousands of words to about 20 (movie review domain)
- Unsupervised methods
  - Use lexicons

Assume pairwise independent features