

האוניברסיטה הפתוחה

20595 כריית מידע

כריית מידע

(Data Mining (DM

מרצה: ד"ר מרק לסט

מרכזת הקורס: ד"ר מיה הרמן

יחידה 15: נושאים מתקדמים

תיאור היחידה

כריית טקסט, כריית תוכן באינטרנט, כריית מולטימדיה.

1

יחידה 15

האוניברסיטה הפתוחה

20595 כריית מידע

Lesson 15 – Advanced Topics

• Text Mining

• Web Mining

• Mining Multimedia Data



2

יחידה 15

Text Mining Definition

- ❖ Feldman and Sanger, 2007:
 - **Text mining** can be broadly defined as a knowledge-intensive process in which a user interacts with a document collection over time by using a suite of analysis tools



What Is Unique in Text Mining?

- ❖ Feature extraction.
- ❖ Very large number of features that represent each of the documents.
- ❖ The need for background knowledge.
- ❖ Even patterns supported by small number of document may be significant.
- ❖ Huge number of patterns, hence need for visualization, interactive exploration.

20595 כריית מידע


האוניברסיטה הפתוחה


Slide by Ronen Feldman, BIU

Text Mining

Input

Documents





Output

Patterns

Connections

Profiles

Trends

5

יחידה 15

20595 כריית מידע

האוניברסיטה הפתוחה

Slide by Ronen Feldman, BIU

Document Types

- ❖ Structured documents
 - Output from CGI (Common Gateway Interface)
- ❖ Semi-structured documents
 - Seminar announcements
 - Job listings
 - Ads
- ❖ Free format documents
 - News
 - Scientific papers

6

יחידה 15

20595 כריית מידע

האוניברסיטה הפתוחה

Text Representations

- ❖ Characters
- ❖ n -grams
- ❖ Words
- ❖ Linguistic Phrases
- ❖ Keyphrases
- ❖ Non-consecutive phrases
- ❖ Concepts
- ❖ Frames
- ❖ Parse trees
- ❖ Graphs

7

יחידה 15

20595 כריית מידע

האוניברסיטה הפתוחה

Bag-of-Words Approach

Documents

Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal.

Now we are engaged in a great civil war, testing whether that nation, or ...

Feature Extraction

Token Sets

nation – 5
civil - 1
war – 2
men – 2
died – 4
people – 5
Liberty – 1
God – 1
...

Loses all order-specific information!

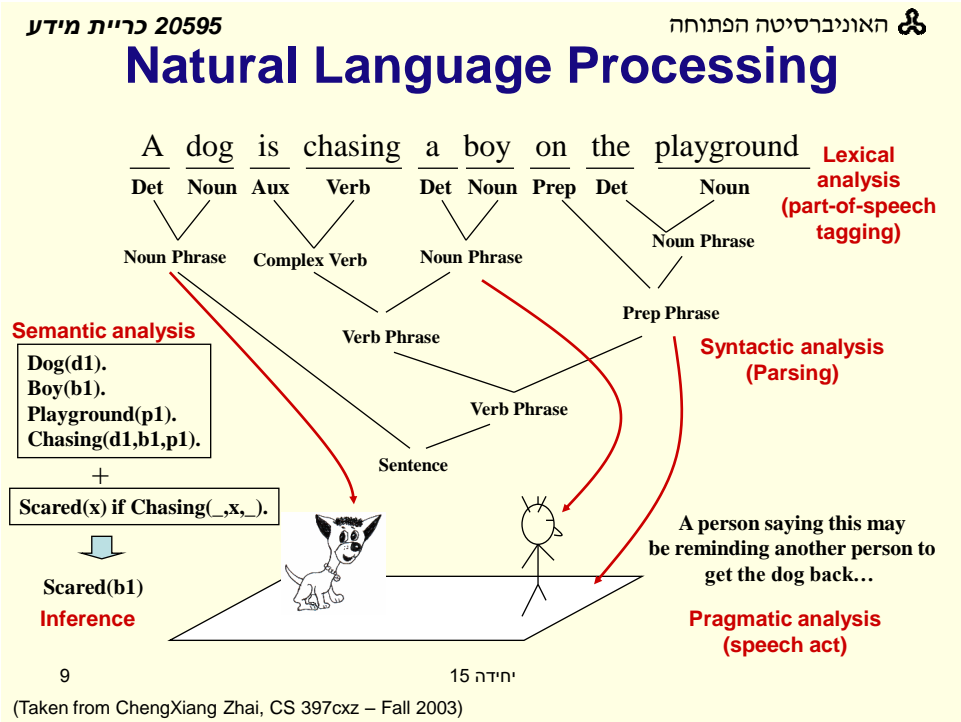
Severely limits context!

8

יחידה 15

Dr. Mark Last

4



20595 כריית מידע האוניברסיטה הפתוחה

General NLP—Too Difficult!

- ❖ Word-level ambiguity
 - “**design**” can be a noun or a verb (Ambiguous POS)
 - “**root**” has multiple meanings (Ambiguous sense)
- ❖ Syntactic ambiguity
 - “**natural language processing**” (Modification)
 - “**A man saw a boy with a telescope.**” (PP Attachment)
- ❖ Anaphora resolution
 - “**John persuaded Bill to buy a TV for himself.**”
(himself = John or Bill?)
- ❖ Presupposition
 - “**He has quit smoking.**” implies that he smoked before.

**Humans rely on context to interpret (when possible).
This context may extend beyond a given document!**

15 יחידה

(Taken from ChengXiang Zhai, CS 397cxz – Fall 2003)

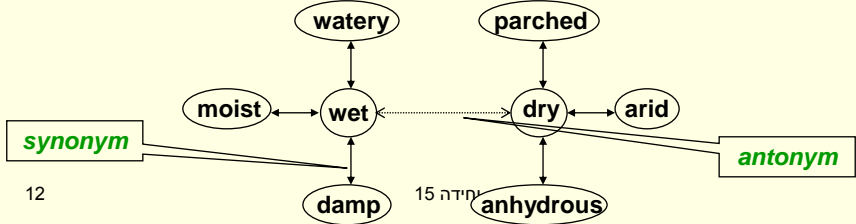
Shallow Linguistics

- Progress on Useful Sub-Goals:
- English Lexicon
 - Part-of-Speech Tagging
 - Word Sense Disambiguation
 - Phrase Detection / Parsing

WordNet

An extensive **lexical network** for the English language

- Contains over **138,838 words**.
- Several graphs, one for each **part-of-speech**.
- **Synsets** (synonym sets), each defining a semantic sense.
- **Relationship** information (antonym, hyponym, meronym ...)
- Downloadable for **free** (UNIX, Windows)
- Expanding to **other languages** (Global WordNet Association)
- Funded **>\$3 million**, mainly government (translation interest)
- Founder **George Miller**, **National Medal of Science**, 1991.





Text Mining Technologies

(based on CACM, Sept. 2006)

- ❖ Information extraction
 - Identifying key phrases and relationships within text
- ❖ Topic tracking
 - A topic-tracking system keeps user profiles and, based on the documents a user views, predicts other documents of interest to the user.
- ❖ Summarization.
 - Text summarization helps users figure out whether a lengthy document meets their needs and is worth reading.



Text Mining Technologies (cont.)

- ❖ Categorization
 - Categorization involves identifying the main themes of a document
- ❖ Clustering
 - Clustering is a technique used to group similar documents, but it differs from categorization in that documents are clustered on the fly instead of through predefined topics.
- ❖ Concept linkage
 - Concept-linkage tools connect related documents by identifying their shared concepts, helping users find information they perhaps wouldn't have found through traditional search methods.

20595 כריית מידע

האוניברסיטה הפתוחה

Information Extraction (IE)

- ❖ IE does not indicate which documents need to be read by a user, it rather extracts pieces of information that are salient to the user's needs.
- ❖ Links between the extracted information and the original documents are maintained to allow the user to reference context.
- ❖ The kinds of information that systems extract vary in detail and reliability.
- ❖ Named entities such as persons and organizations can be extracted with reliability in the 90th percentile range, but do not provide attributes, facts, or events that those entities have or participate in.

15

יחידה 15

20595 כריית מידע

האוניברסיטה הפתוחה

IE Accuracy by Information Type

Information Type	Accuracy
Entities	90-98%
Attributes	80%
Facts	70%
Events	60%

16

יחידה 15

20595 כריית מידע

האוניברסיטה הפתוחה

Slide by Ronen Feldman, BIU

IE Example 1: A Tagged News Document

Ethicon endo-Surgery Acquires Swedish Adjustable Gastric Band

Acquisition

Acquirer: Ethicon Endo-Surgery
Acquired: Obtech Medical

Date

June 27, 2002 2:00pm

Ethicon Endo-Surgery, Inc., a Johnson & Johnson company, has acquired Obtech Medical AG, a privately held **Swiss** company that markets the **Swedish Adjustable Gastric Band** (SAGB), expanding its line of products used in the treatment of **morbid obesity**. Terms of the transaction were not disclosed.

Medical Disorder

"Weight loss surgery for morbid obesity is one of the fastest-growing areas of surgery today," said Nick Valeriani, Company Group Chairman, Johnson & Johnson and Worldwide Franchise Chairman, **Ethicon Endo-Surgery, Inc.**

Employment

Company: Johnson & Johnson
Person: Nick Valeriani
Position: Company Group Chairman

Country

Product

Company

17 יחידה 15

20595 כריית מידע

האוניברסיטה הפתוחה

Slide by Ronen Feldman, BIU

IE Example 2: A Merger Article

In This Document

- Exhibits
- City (1)
- Company (13)
- Date (1)
- Entity (1)
- Money/Amount (2)
- Organization (2)
- Person (1)
- Position (1)
- Province/State (1)
- Stock (1)
- Events & Facts
- Company Location (1)
- Person (1)
- Person Professional (1)

Title: 535875765-32

Date: 2005-06-01

Body: NEW YORK (Reuters) - The proposed merger of US Airways Group Inc. and America West Holdings Corp. has received a new pledge of \$150 million from an investment firm, the Wall Street Journal reported on its Web site on Sunday.

According to the newspaper, which obtained the information from a filing with the U.S. Bankruptcy Court in Alexandria, Virginia, Boston-based Wellington Management Co. LLP pledged \$150 million in new equity to the combined company. Wellington has \$423 billion of funds under management, the Journal said.

The total pledges for the new company now total about \$500 million, the report added.

On May 19, US Airways and America West (business) announced plans to merge in order to compete more aggressively with low-cost rivals. The airline would become the sixth-largest domestic carrier as measured by passenger revenue, and would use the US Airways name. The combined carrier would also fly internationally.

U.S. Bankruptcy Court Judge Stephen Mitchell is expected to grant a motion on Tuesday to set guidelines for other investors to make offers that compete with the merger plan, the newspaper reported.

The companies currently employ 43,000 people and operate roughly 400 aircraft combined.

Key investors to the deal include ACE Aviation Holdings Inc., Air Canada's parent, and Boston - PAK Capital Management. General Electric Co. will take back planes and Airbus will provide a \$250 million loan in exchange for being the A350 launch customer.

America West stock rose 2 cents, to \$5.48 a share, on Friday in trading on the New York Stock Exchange.

18 יחידה 15

Text Categorization (TC)

Basic Definition

❖ TC – task of assigning a Boolean {T, F} value to each pair $\langle d_j, c_i \rangle \in D \times C$

where

$D = (d_1, \dots, d_{|D|})$ is a collection of documents

$C = (c_1, \dots, c_{|C|})$ is a set of pre-defined categories

–Sample categories: “sports”, “entertainment”, “finance”, etc.

Inductive text classification / categorization

- ❖ The Goal
 - Infer a classification model from a representative sample of labeled training documents
- ❖ Requirements in the Web Domain
 - High accuracy
 - The correct category/ categories of each document should be identified as accurately as possible
 - Interpretability
 - An automatically induced model should be subject to scrutiny by a human expert
 - Speed
 - The model should be capable to process massive streams of web documents in minimal time
 - Multilinguality
 - The model induction methods should maintain a high performance level over web content in multiple languages

20595 כריית מידע

האוניברסיטה הפתוחה

❖ Binary TC – two non-overlapping categories only

- Example: “academic” vs. “non-academic”

❖ Multi-Class TC – more than two non-overlapping categories

- Example: “sports” or “politics” or “computing”
- A multi-class problem can be reduced into multiple binary tasks (*one-against-the-rest* strategy)

❖ Multi-Label TC – overlapping categories are allowed

- Example: an “academic” document on “computing”
- A multi-label task can be split into a set of binary classification tasks

❖ Ranking categorization

- *Category ranking*: which categories match a given document best?
- *Document ranking*: which documents match a given category best?

21

יחידה 15

20595 כריית מידע

האוניברסיטה הפתוחה

Document Clustering

❖ Motivation

- Automatically group related documents based on their contents
- No predetermined training sets or taxonomies
- Generate a taxonomy at runtime

❖ Clustering Process

- Data preprocessing: remove stop words, stem, feature extraction, lexical analysis, etc.
- Hierarchical clustering: compute similarities applying clustering algorithms.
- Model-Based clustering (Neural Network Approach): clusters are represented by “exemplars”. (e.g.: SOM)

22

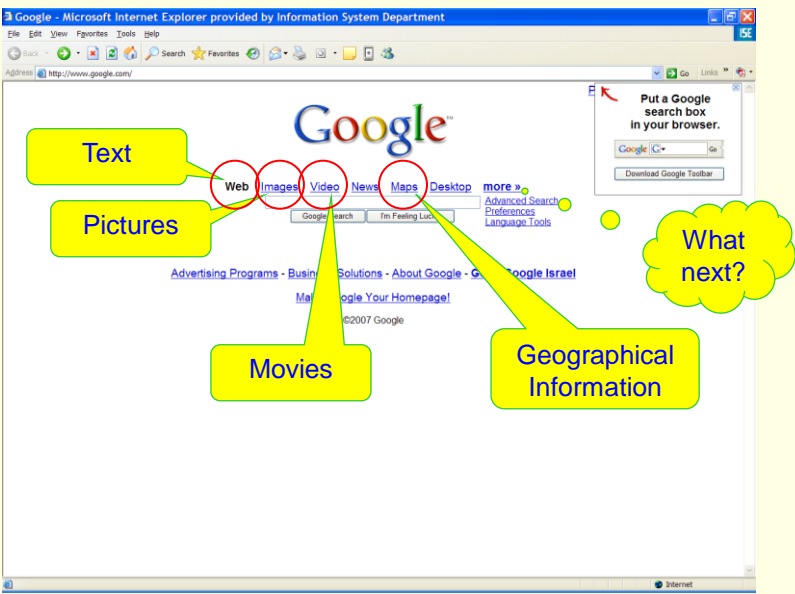
יחידה 15

Dr. Mark Last

11

Lesson 15 – Advanced Topics

- Text Mining
- Web Mining ←
- Mining Multimedia Data



20595 כריית מידע

האוניברסיטה הפתוחה

Web Mining Tasks

❖ Web mining - the application of data mining techniques to extract knowledge from Web content, structure, and usage.

25

יחידה 15

20595 כריית מידע

האוניברסיטה הפתוחה

Web Content Mining: Definition

- ❖ “Web Content Mining is the process of extracting useful information from the contents of Web documents. It may consist of text, images, audio, video, or structured records such as lists and tables.”
- ❖ “Web Content mining refers to the overall process of discovering potentially useful and previously unknown information or knowledge from the Web data.”

26

יחידה 15

The Web: Opportunities & Challenges

- ❖ Web offers an unprecedented opportunity and challenge to data mining
 - The amount of information on the Web is huge and easily accessible.
 - The coverage of Web information is very wide and diverse. One can find information about almost anything.
 - Information/data of almost all types exist on the Web, e.g., structured tables, texts, multimedia data, etc.

More Opportunities & Challenges

- ❖ Much of the Web information is semi-structured due to the nested structure of HTML code.
- ❖ Much of the Web information is linked. There are hyperlinks among pages within a site, and across different sites.
- ❖ Much of the Web information is redundant. The same piece of information or its variants may appear in many pages.
- ❖ The Web is noisy. A Web page typically contains a mixture of many kinds of information, e.g., main contents, advertisements, navigation panels, copyright notices, etc.

Even More Opportunities & Challenges

- ❖ The Web consists of surface Web and deep Web.
 - Surface Web: pages that can be browsed using a browser.
 - Deep Web: databases that can only be accessed through parameterized query interfaces.
- ❖ The Web is also about services. Many Web sites and pages enable people to perform operations with input parameters, i.e., they provide services.
- ❖ The Web is dynamic. Information on the Web changes constantly. Keeping up with the changes and monitoring the changes are important issues.
- ❖ Above all, the Web is a virtual society. It is not only about data, information and services, but also about interactions among people, organizations and automatic systems, i.e., communities.

Lesson 15 – Advanced Topics

- Text Mining
- Web Mining
- Mining Multimedia Data



Similarity Search in Multimedia Data

- ❖ Description-based retrieval systems
 - Build indices and perform object retrieval based on image descriptions, such as keywords, captions, size, and time of creation
 - Labor-intensive if performed manually
 - Results are typically of poor quality if automated
- ❖ Content-based retrieval systems
 - Support retrieval based on the image content, such as color histogram, texture, shape, objects, and wavelet transforms

Approaches Based on Image Signature

- ❖ Color histogram-based signature
 - The signature includes color histograms based on color composition of an image regardless of its scale or orientation
 - No information about shape, location, or texture
 - Two images with similar color composition may contain very different shapes or textures, and thus could be completely unrelated in semantics
- ❖ Multifeature composed signature
 - Define different distance functions for color, shape, location, and texture, and subsequently combine them to derive the overall result

Mining Multimedia Databases

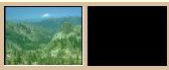
Refining or combining searches



Search for “blue sky”
(top layout grid is blue)

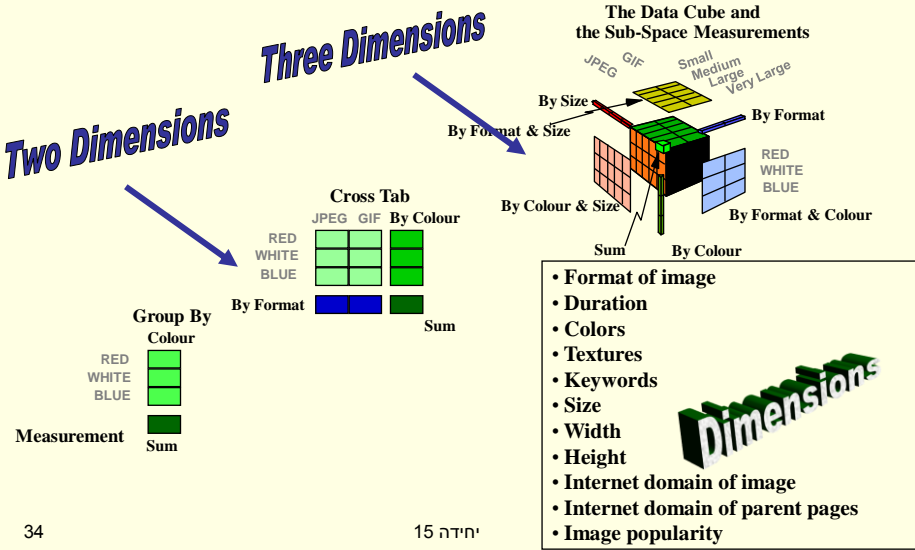


Search for “airplane in blue sky”
(top layout grid is blue and
keyword = “airplane”)



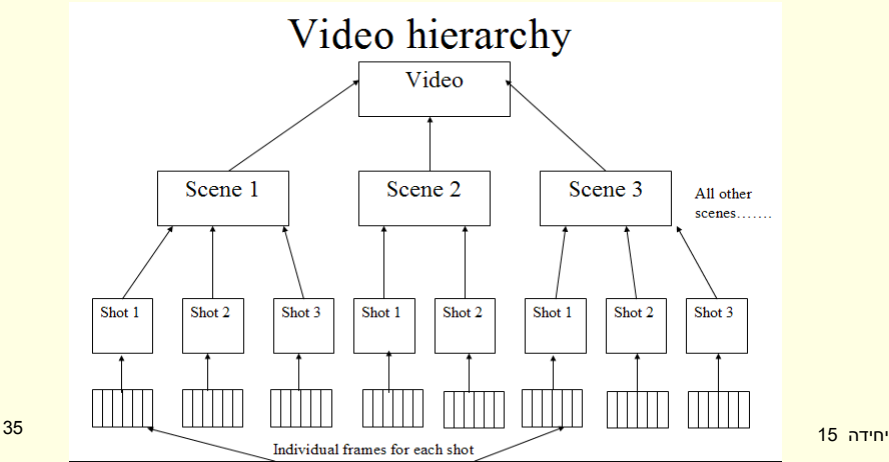
Search for “blue sky and
green meadows”
(top layout grid is blue
and bottom is green)

Mining Multimedia Databases



Video Segmentation

Videos sequences can be segmented in a hierarchical fashion into **scenes** (specific story sequence), **shots** (single camera sequence), and **frames** (individual images extracted from each shot).



Object-Based Video Segmentation

- ❖ Each frame in a given video sequence (i.e., shot) is segmented into objects based on a pre-specified criteria (such as color, texture, motion, etc.)



Frame 90



Frame 90 Segmented Regions

Object Based Video Segmentation (cont.)

- ❖ All objects are maintained (foreground/background) for the entire sequence
- ❖ Corresponding objects are matched and assigned same labels
- ❖ Valuable when implementing
 - MPEG-4 Video Compression
 - MPEG-7 Video Retrieval
 - Video Databases and Video Data Mining Applications

Video Mining – Challenges and Research Directions

- | | |
|--|---|
| <ul style="list-style-type: none">❖ Video clustering and categorization❖ Video based object recognition❖ Video segmentation and summarization❖ Video feature extraction and representation❖ Video indexing and retrieval❖ Video search engines❖ Video editing and browsing systems❖ Visual event and activity detection | <ul style="list-style-type: none">❖ Statistical techniques for video analysis❖ Semantic video content analysis❖ Video processing for HCI❖ Video surveillance (person identification, abnormal activity labeling ...)❖ Consumer video applications (sports highlight detection, commercial message extraction ...) |
|--|---|

Summary

- ❖ The feature extraction stage in text mining is the most challenging and labor intensive
- ❖ Text mining tasks include information extraction, document categorization, document clustering, and many others
- ❖ Web mining is the application of data mining techniques to extract knowledge from Web content, structure, and usage
- ❖ Mining multimedia, especially video data is an emerging and still under-explored field