Text Analysis and Retrieval

1. Introduction

Prof. Jan Šnajder

University of Zagreb
Faculty of Electrical Engineering and Computing (FER)

Academic Year 2022/2023

v3.0

After this lecture, you'll...

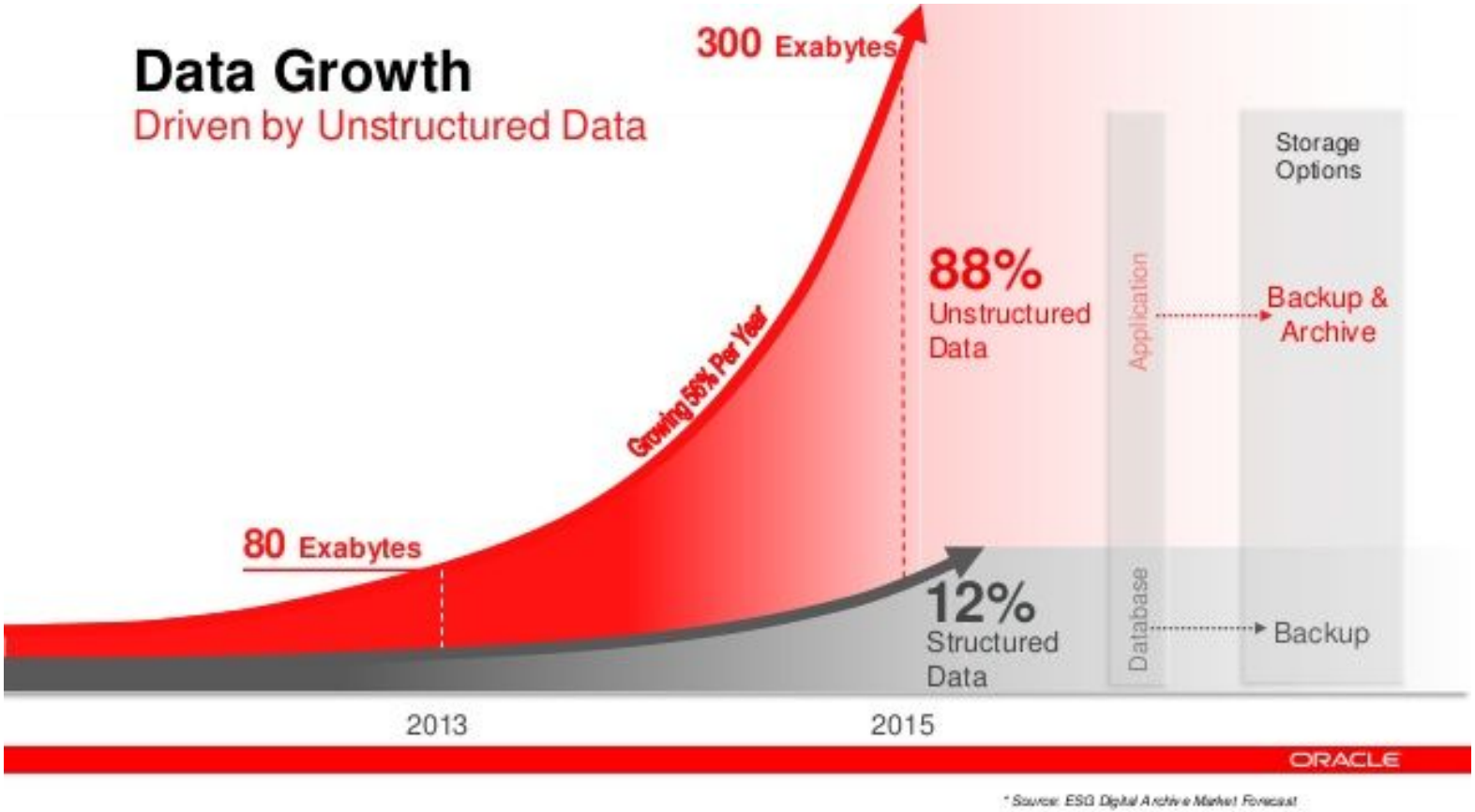
1. Know what this course is about and be glad that you've enrolled it
2. Know what are the topics that we want to cover
3. Know what you need to do to earn your credits

Outline

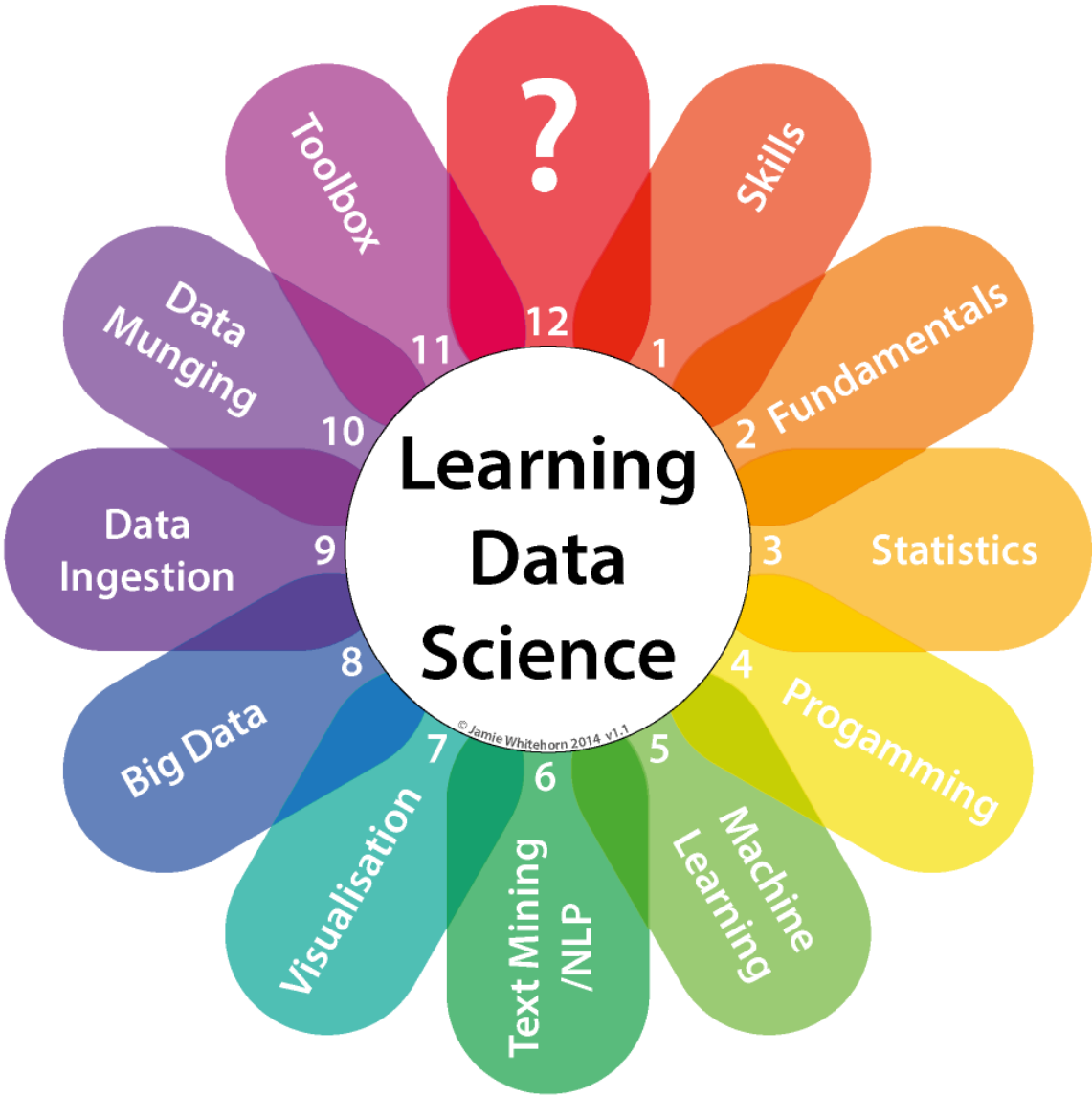
What is this course about?

- Text is everywhere (books, news, emails, tweets, reviews,. . .)
- Most human knowledge is stored in unstructured, textual form
- The amount of text data is vast and rapidly growing
- We need systems that address diverse information needs of the users and extract information from large volumes of unstructured data

Unstructured data

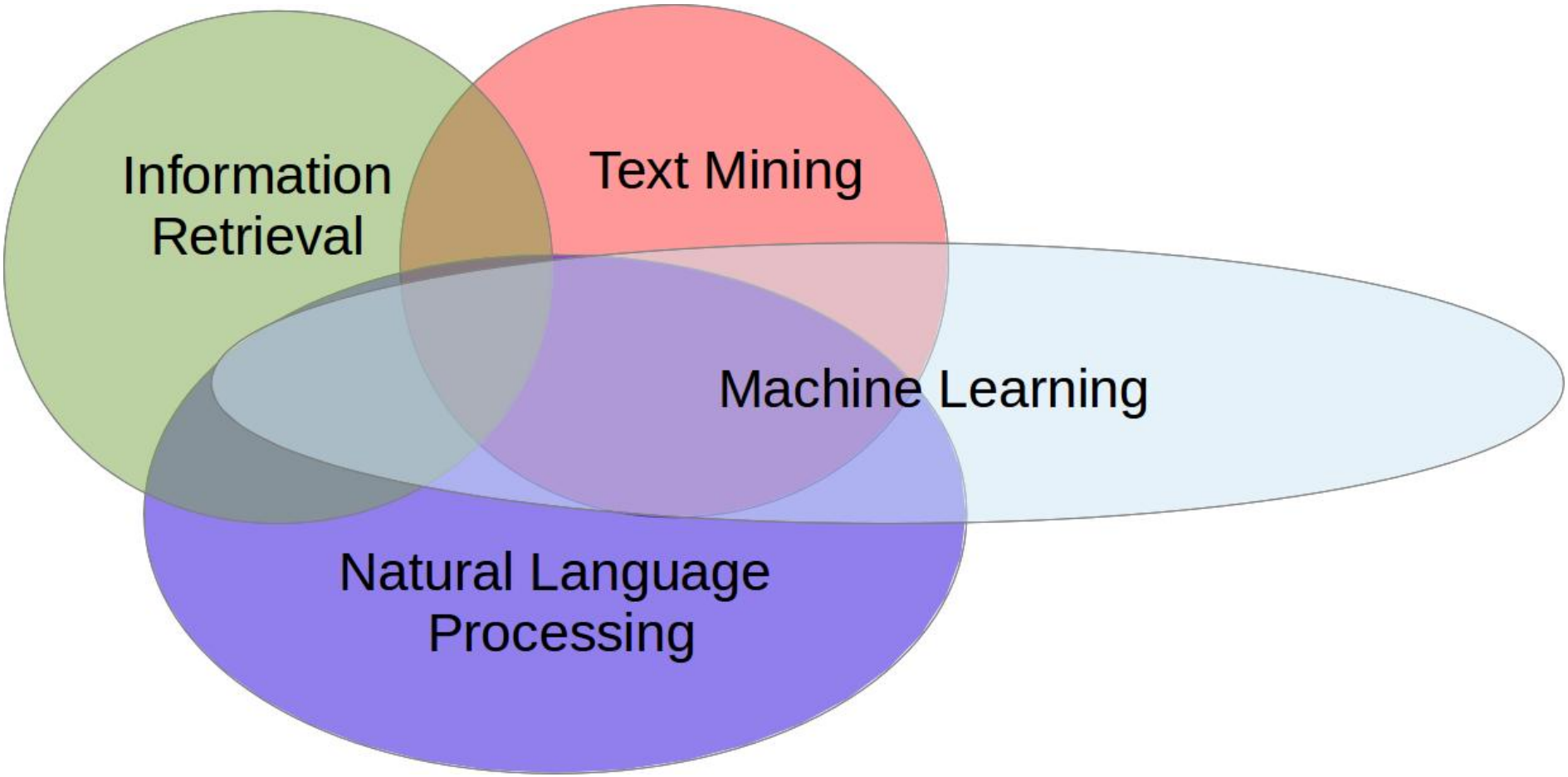


Data Science



http://www.exploringdatascience.com/the-data-science-clock/

What's involved?

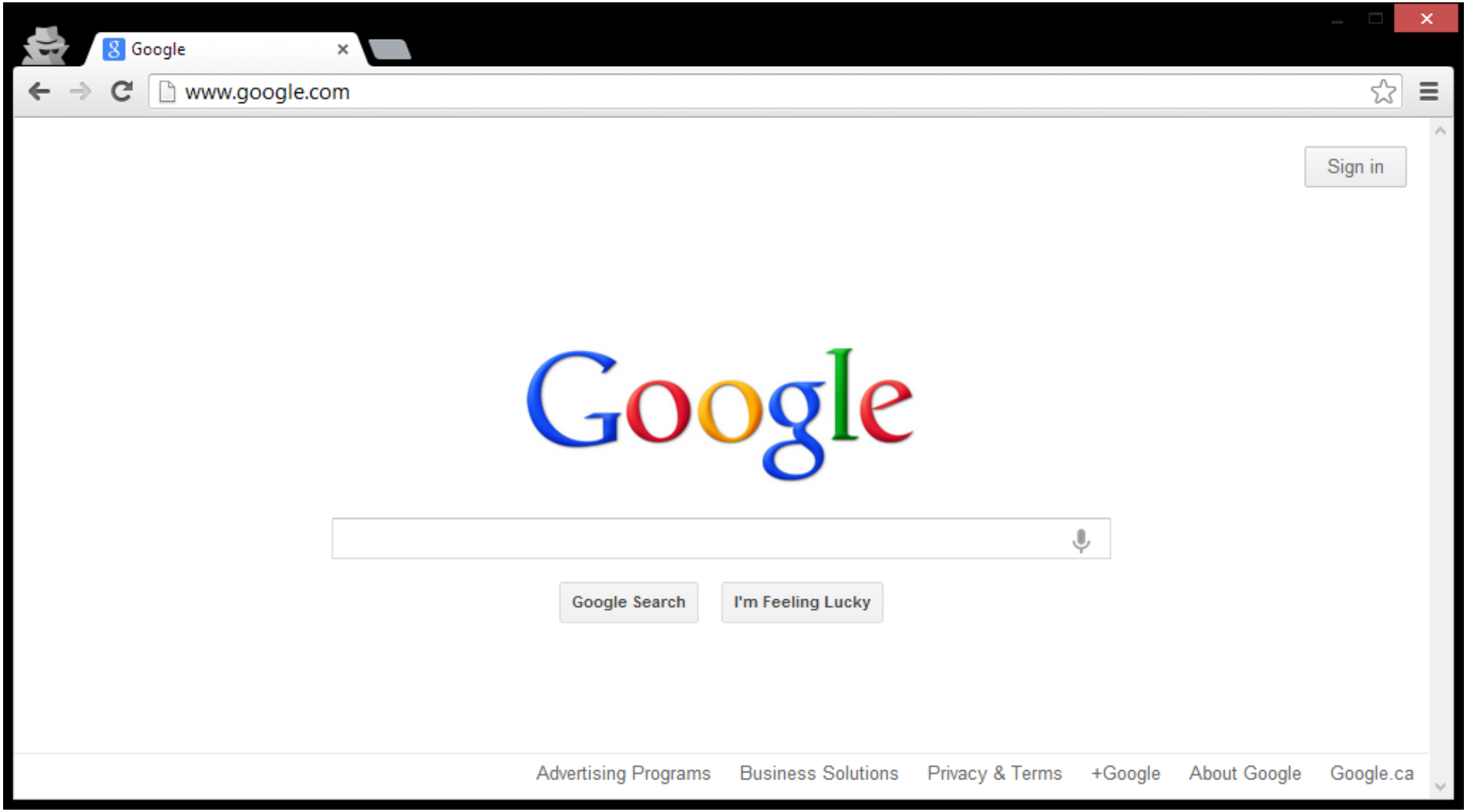


1. Information retrieval

Information Retrieval (Manning & Schütze 2008)

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

Web search


Google
www.google.com
Sign in
Google
Google Search
I'm Feeling Lucky
Advertising Programs
Business Solutions
Privacy & Terms
+Google
About Google
Google.ca

Domain-specific search

ClassifiedAds.com

Log in Help Post an ad (free)

All Categories | small parrot cage | in Earth | SEARCH

Sponsored link related to: **small parrot cage**

Bird Cage Suppliers Accepting Orders in **Small** Volume
small-order.hktdc.com/ Reduce risk in sourcing! Place order in **small** quantity @ hktdc.com

Used Bird and Parrot Cages, good condition
PLS. EMAIL FOR APPT. TO VIEW ALL! Cleaning out my garage of used bird cages ($5-40), iron parrot cages

large parrot cage, 40"wide, 32"deep, 74" high, almost new
green large parrot cage, just bought, too big for house

parrot cage
Parrot cage. Good condition Comes with two bowel feeders and wooden bench.

Parrot Cage With toys and Dishes
Parrot cage with toys and dishes. If interested please text me and I will send you a picture. Or just call me to

Parrot Cage
Black Wrought iron Parrot cage, about 5 ft. tall, fits in corner. Good condition. 561-389-5347

Large Parrot Cage For Sale
LARGE WHITE 5 FOOT PARROT CAGE EXCELLENT CONDITION TOP PERCH WITH LOWER SHELVE FOR STORAGE

LARGE Parrot cage for sale
Large Parrot Cage for sale with top perch and lower storage area comes with ladder for bird to climb back into

wanted parrot cage 3 feet wide black willing to pay 150$
Wanted Parrot cage willing to pay 150 Balck 3 feet wide

Small Plastic cage with Hermit crab
Purchased this summer by my oldest daughter for near $35, (she is onto other activities.) Has large hermit

PARROT CAGE, TOYS, BOOKS, ETC.
ALL YOU NEED NOW IS THE PARROTnbsp; This Dometop cage provides your pet with a new home that is both

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 ··· >

Create your own ad in Earth . It's easy and free!

Follow, Like or add us to your Google + circles

Tweet 0 Follow

Like 0

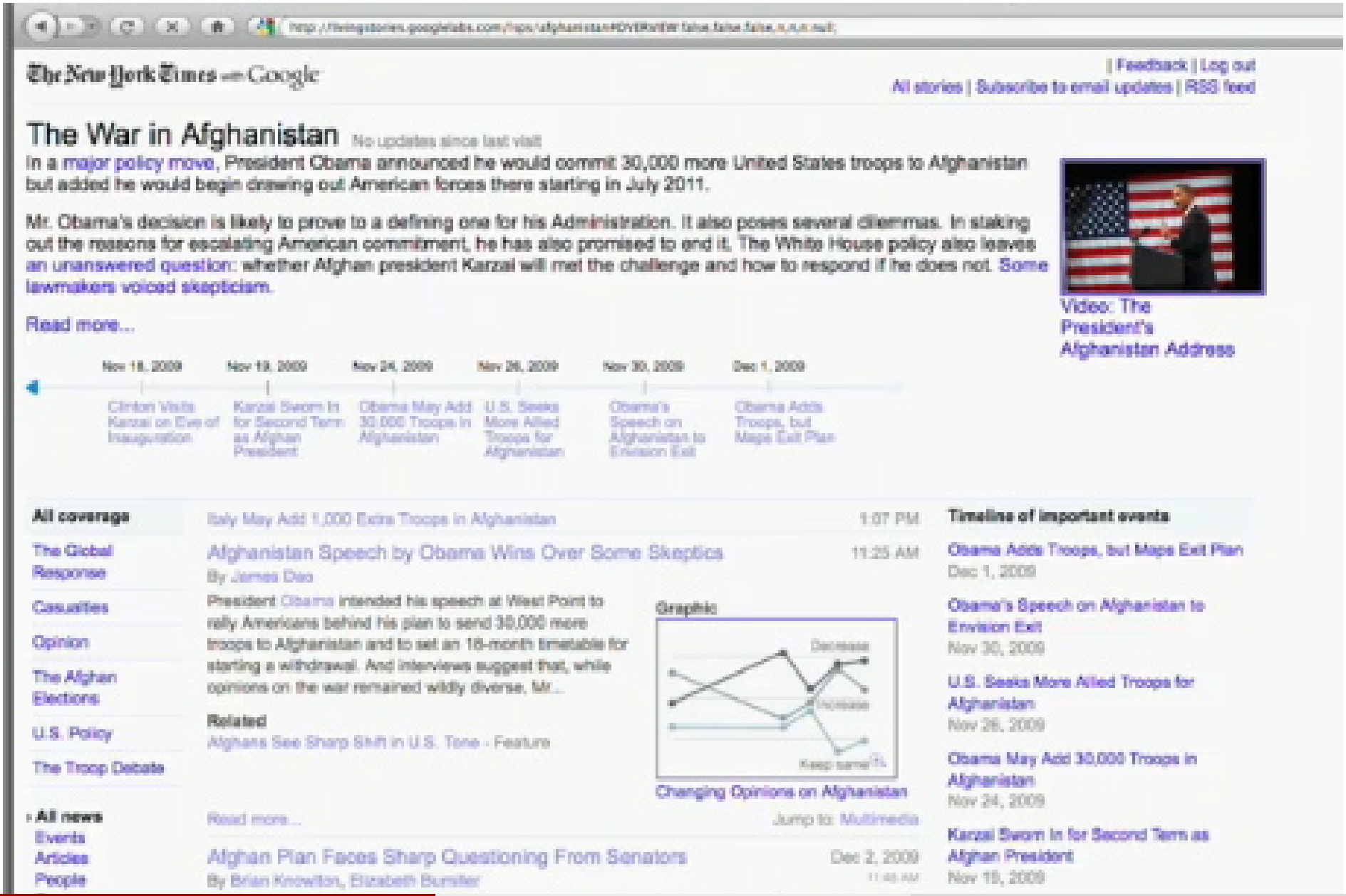
+1 0 Follow 1.2k

2. Text mining (aka text anaytics)

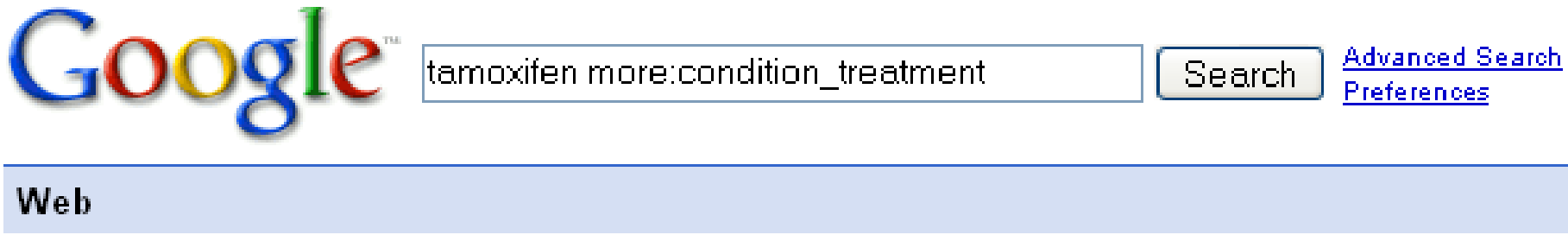
Text mining (Hearst 2003)

Text mining (TM) is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information to form new facts or new hypotheses to be explored further by more conventional means of experimentation.

Text mining


The New York Times on Google
| Feedback | Log out
All stories | Subscribe to email updates | RSS feed
The War in Afghanistan
No updates since last visit
In a major policy move, President Obama announced he would commit 30,000 more United States troops to Afghanistan but added he would begin drawing out American forces there starting in July 2011.
Mr. Obama's decision is likely to prove to a defining one for his Administration. It also poses several dilemmas. In staking out the reasons for escalating American commitment, he has also promised to end it. The White House policy also leaves an unanswered question: whether Afghan president Karzai will met the challenge and how to respond if he does not. Some lawmakers voiced skepticism.
Read more...
Video: The President's Afghanistan Address
Nov 18, 2009
Nov 19, 2009
Nov 24, 2009
Nov 26, 2009
Nov 30, 2009
Dec 1, 2009
Clinton Visits Karzai on Eve of Inauguration
Karzai Sworn In for Second Term as Afghan President
Obama May Add 30,000 Troops in Afghanistan
U.S. Seeks More Allied Troops for Afghanistan
Obama's Speech on Afghanistan to Envision Exit
Obama Adds Troops, but Maps Exit Plan
All coverage
The Global Response
Casualties
Opinion
The Afghan Elections
U.S. Policy
The Troop Debate
All news
Events
Articles
People
Italy May Add 1,000 Extra Troops in Afghanistan
1:07 PM
Afghanistan Speech by Obama Wins Over Some Skeptics
11:25 AM
By James Dao
President Obama intended his speech at West Point to rally Americans behind his plan to send 30,000 more troops to Afghanistan and to set an 18-month timetable for starting a withdrawal. And interviews suggest that, while opinions on the war remained wildly diverse, Mr...
Related
Afghans See Sharp Shift in U.S. Tone - Feature
Graphic
Decrease
Increase
Keep same
Changing Opinions on Afghanistan
Read more...
Jump to: Multimedia
Afghan Plan Faces Sharp Questioning From Senators
Dec 2, 2009
By Brian Knowlton, Elizabeth Bumiller
Timeline of important events
Obama Adds Troops, but Maps Exit Plan
Dec 1, 2009
Obama's Speech on Afghanistan to Envision Exit
Nov 30, 2009
U.S. Seeks More Allied Troops for Afghanistan
Nov 26, 2009
Obama May Add 30,000 Troops in Afghanistan
Nov 24, 2009
Karzai Sworn In for Second Term as Afghan President
Nov 19, 2009

Information retrieval + Text mining



3. Natural language processing

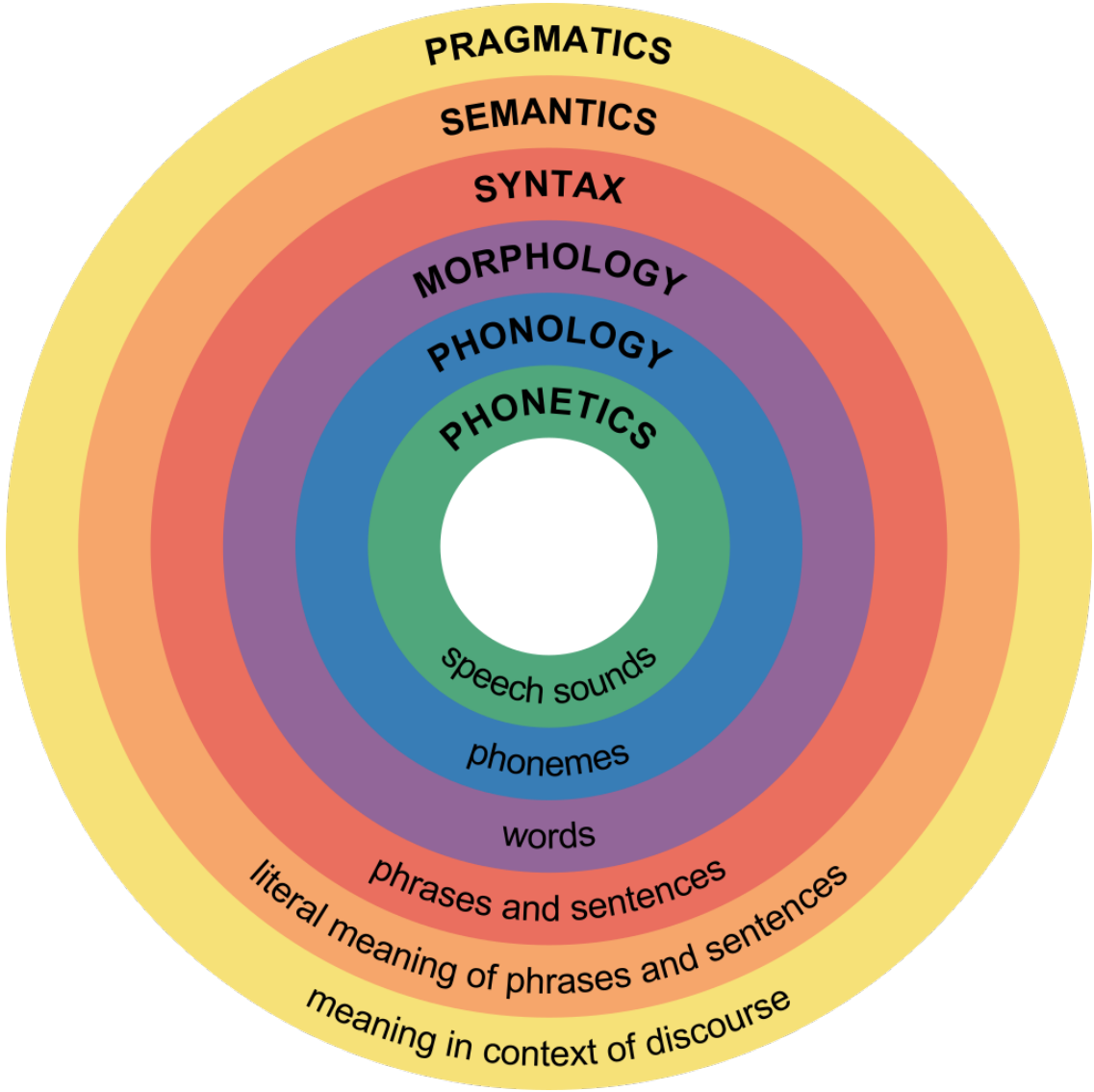
Natural language processing *(Wikipedia)*

Natural language processing (NLP) is a field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human–computer interaction.

Why is TAR a challenge?

- Text is written in natural language. Natural language is a tough nut to crack. It is complex, ambiguous, vague, and relies on commonsense knowledge.
- Full understanding of natural language is an AI-complete problem.
- On top of this: dealing with large amounts of data poses serious technical challenges.

Language complexity



Language vagueness

Rezultati pretrage za *mali stan*

1 2 3 4 5 6 7-12 SLJEDEĆA »

6091 oglas — Sortiraj: Relevantnosti

Vau Vau Njuškalo oglasi

Srdoči, odličan mali stan u novogradnji

Stan u stambenoj zgradi, Prizemlje
Stambena površina: 24.44 m2
Objavljen: 03.03.2014.

46.500 € ~ **355.685 kn**

KVATERNIKOV TRG - UREĐEN mali stan

Stan u stambenoj zgradi, 5 kat
Stambena površina: 40.00 m2
Objavljen: 06.03.2014.

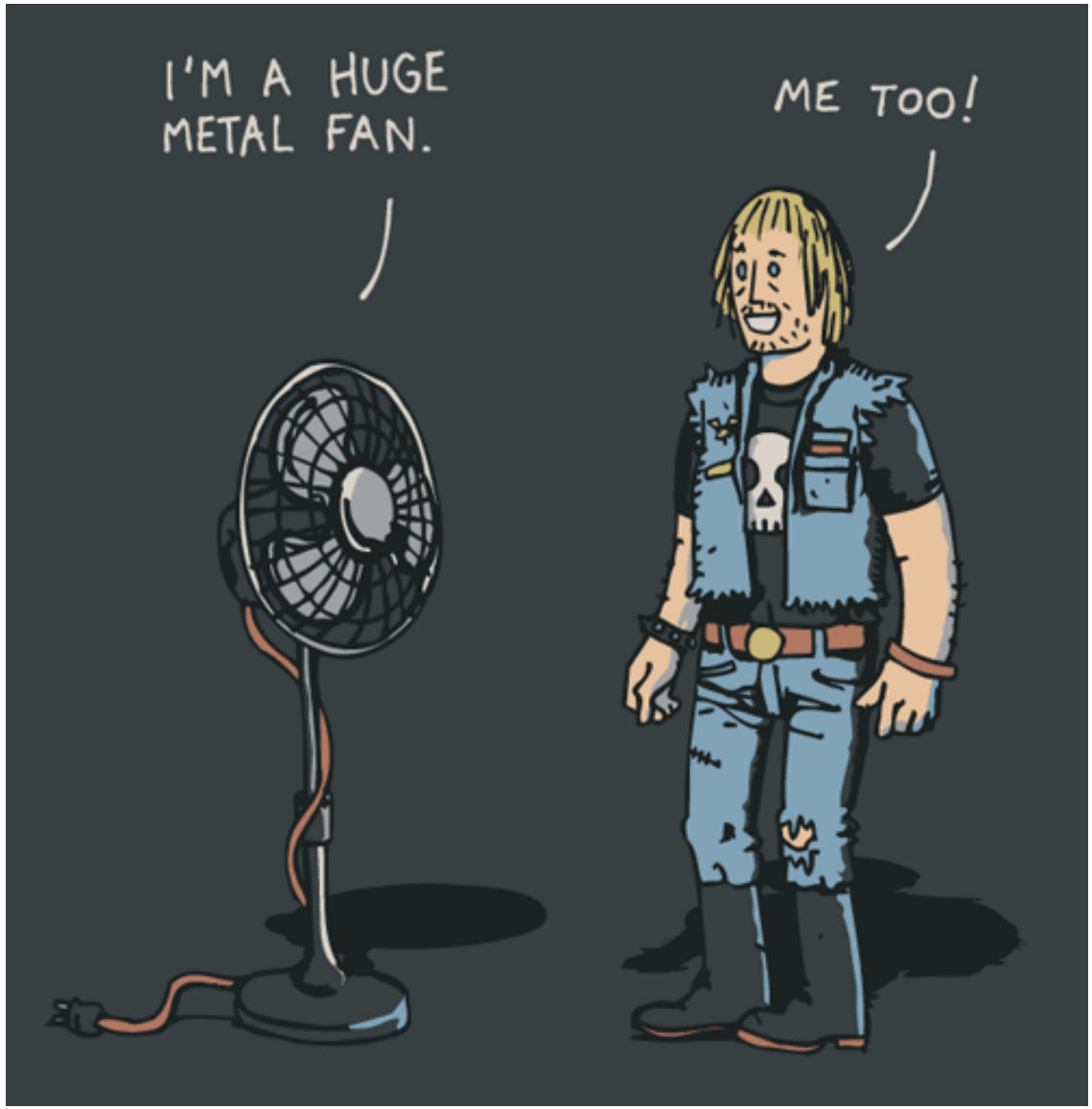
70.000 € ~ **535.441 kn**

MALI STAN U FAŽANI !!

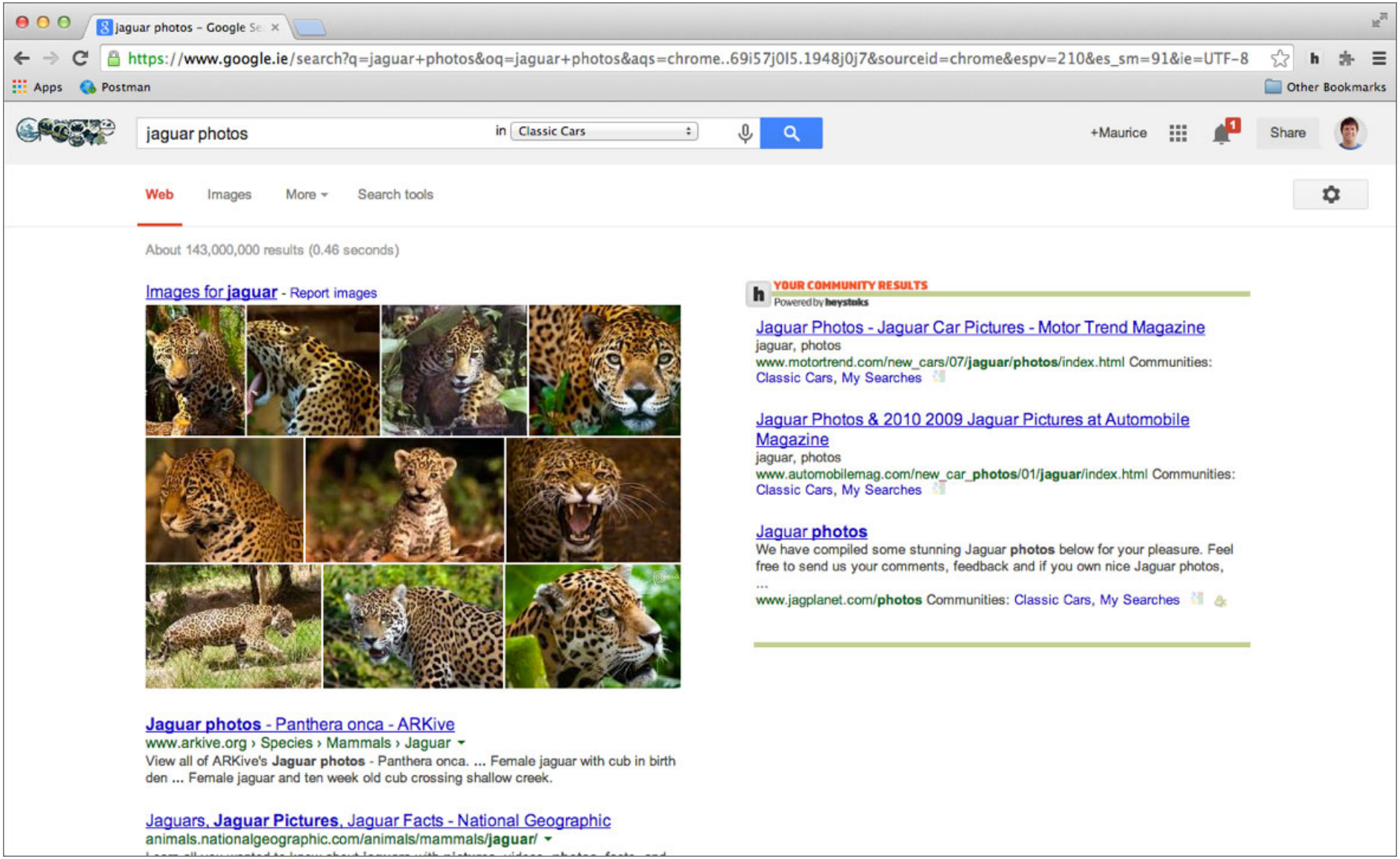
Stan u stambenoj zgradi, ---
Stambena površina: 35.00 m2
Objavljen: 05.03.2014.

48.500 € ~ **370.984 kn**

Language ambiguity

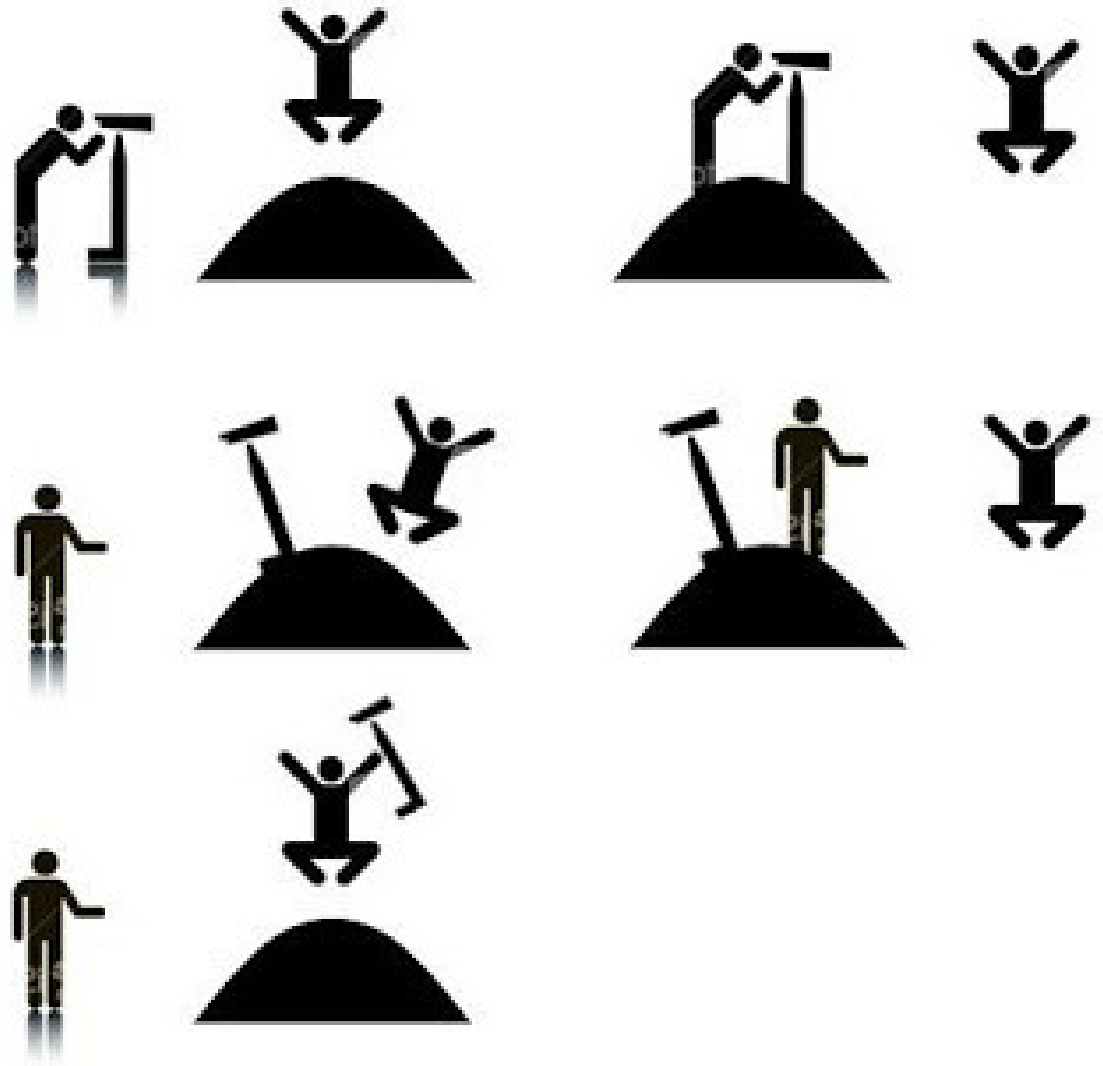


Word sense ambiguity



Structural ambiguity

"I saw the man on the hill with a telescope"



Outline

Course aims and target audience

Aims

Provide a systematic overview of both traditional and advanced methods for text analysis and retrieval. The course is divided into two parts. The first part covers the fundamentals: NLP, IR, and ML. The second part focuses on concrete applications in information extraction and text analytics, with an emphasis on methods based on statistical natural language processing and machine learning.

Audience

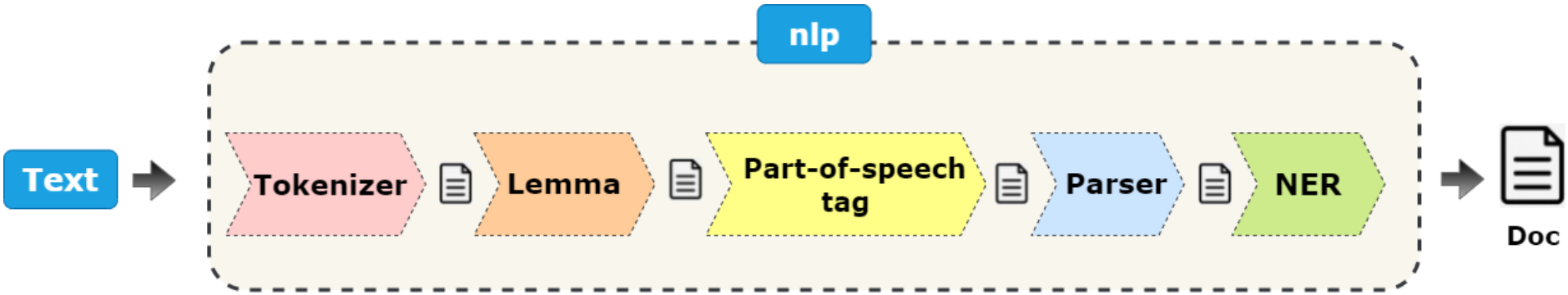
Students interested in gaining an understanding of and practical experience with basic information retrieval and text analysis techniques. Knowledge of at least basic machine learning is required. No previous knowledge of natural language processing is required.

Topics

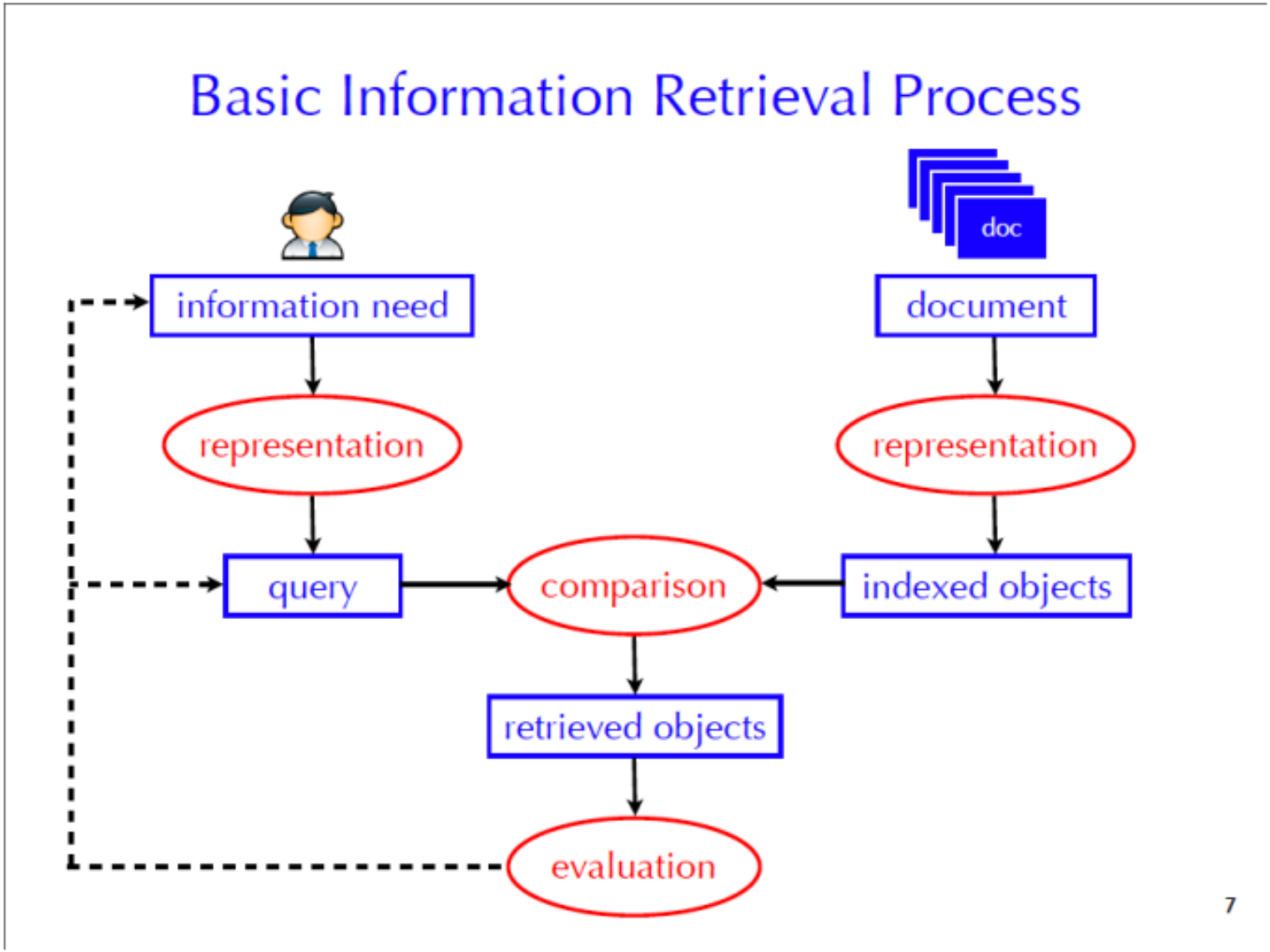
1. Introduction
2. Basics of NLP
3. Basics IR
4. Machine learning for NLP
5. Semantics
6. Neural NLP: Recurrent models
7. Neural NLP: Transformers

   (two weeks break)
8. Information extraction
9. Question answering
10. Summarization
11. Natural language inference
12. Sentiment analysis
13. Author profiling; Wrap-up

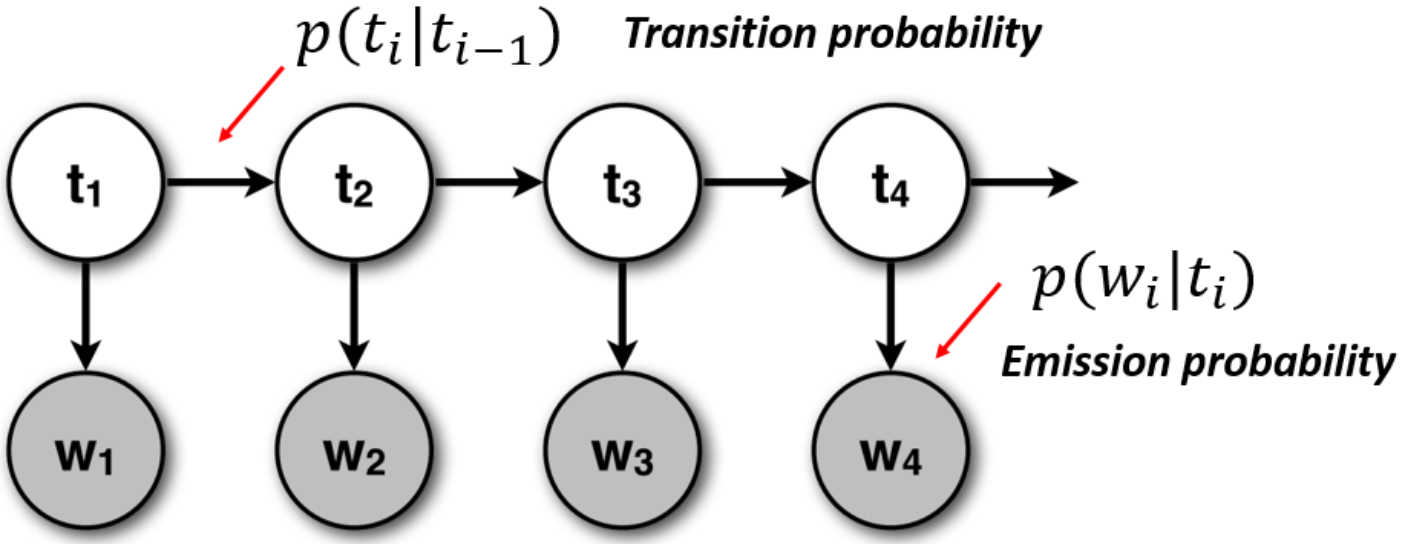
Basic of NLP



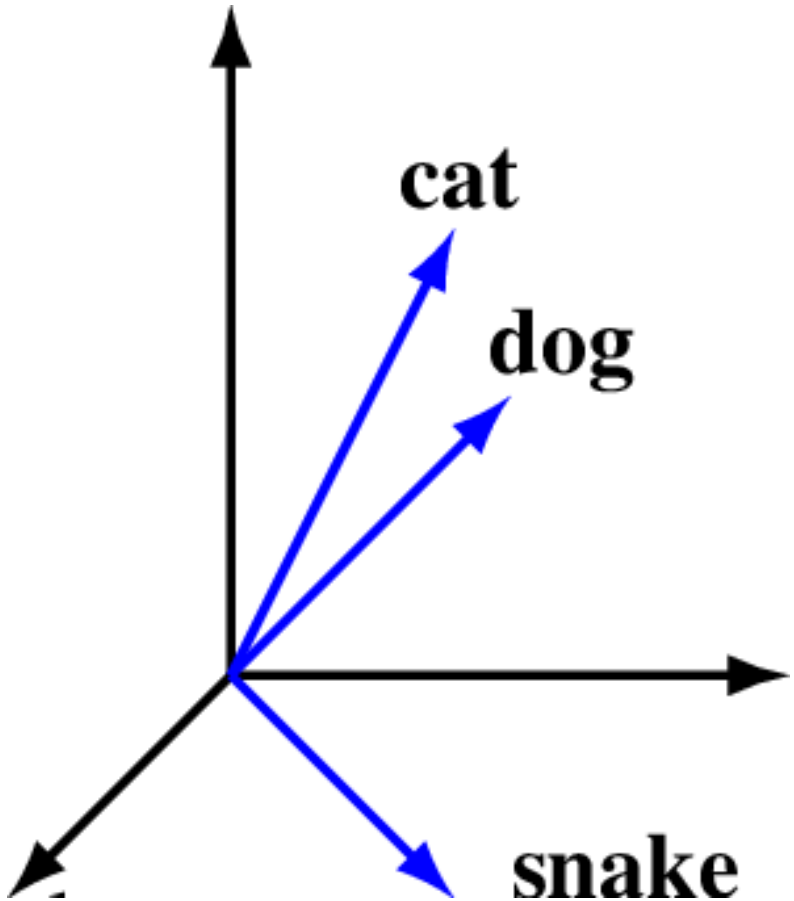
Information retrieval



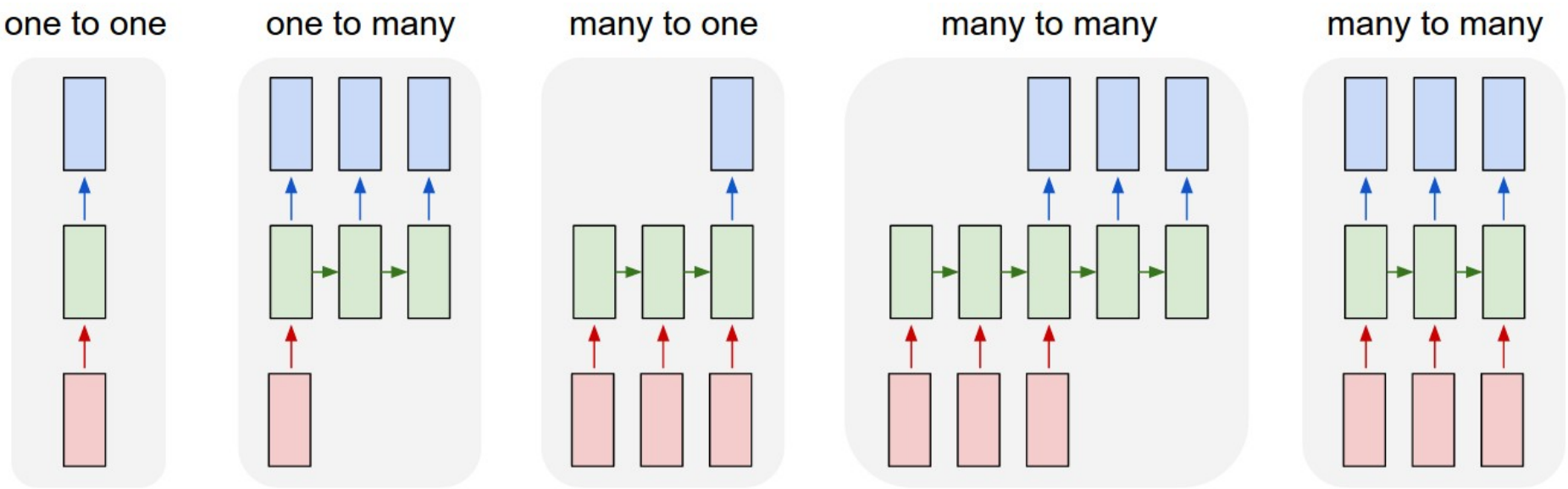
Machine learning for NLP



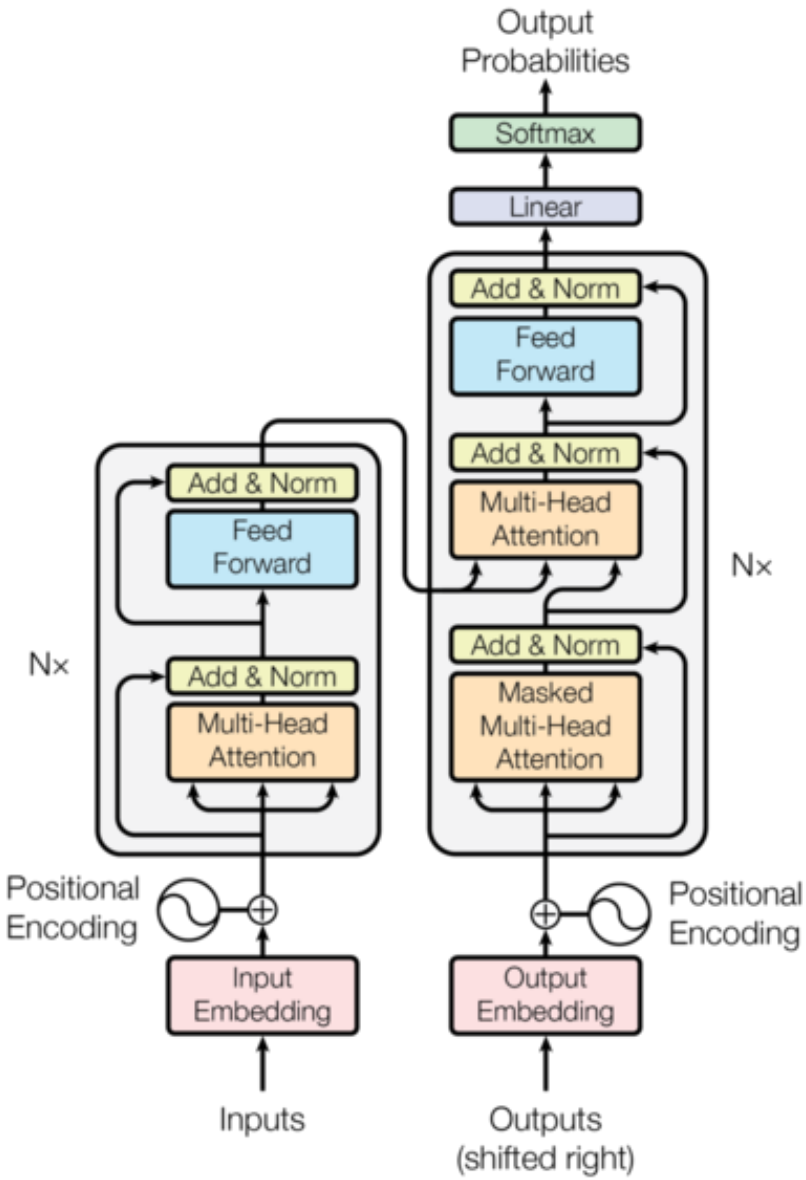
Semantics



Neural NLP: Recurrent models



Neural NLP: Transformers



Outline

Teaching staff

Lecturers:

- Prof. Jan Šnajder

TAs:

- Josip Jukić

Student TAs:

- Tin Ferković, Ivan Martinović, Luka Pavlović, Marko Rajnović, Matea Vasilj, Janko Vidaković

Course web site:
`http://www.fer.hr/predmet/apt`

Powered by...

Text Analysis and Knowledge Engineering Lab

`http://takelab.fer.hr`


TL
TakeLab

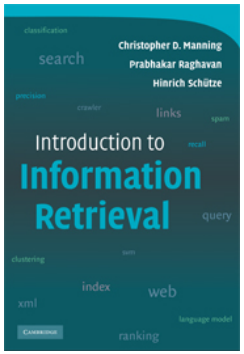
Learning outcomes

On successful completion of the course, you will be able to:

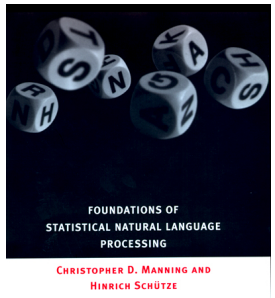
1. Summarize the application areas, trends, and challenges in text analysis and retrieval
2. Describe the fundamental techniques of text analysis and retrieval
3. Use linguistic preprocessing tools
4. Design and implement a text analysis/retrieval system
5. Apply machine learning algorithms to text analysis tasks
6. Evaluate a text analysis/retrieval system
7. Organize and formulate a system description paper
8. Describe, review, analyze, and criticize the main text analysis methods present in scientific papers

Textbooks

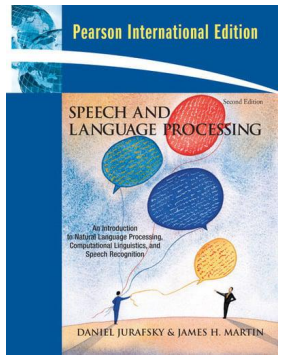
- C. D. Manning, P. Raghavan, H. Schütze. *Introduction to Information Retreval*, CUP, 2008



- C. D. Manning, H. Schütze: *Foundations of Statistical Natural Language Processing*, MIT Press, 1999

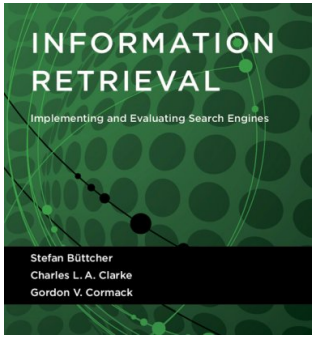


- D. Jurafsky, J. H. Martin. *Speech and Language Processing*, 3rd ed., 2023.

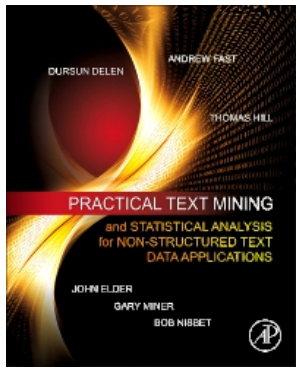


Textbooks

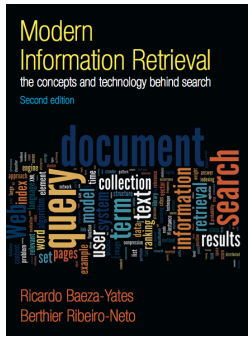
- S. Buettcher, C. L. A. Clarke, G. V. Cormack. *Information Retrieval: Implementing and Evaluating Search Engines*, MIT Press, 2010



- G. Miner, J. Elder IV, T. Hill, R. Nisbet, D. Delen, A. Fast. *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*, Academic Press, 2012

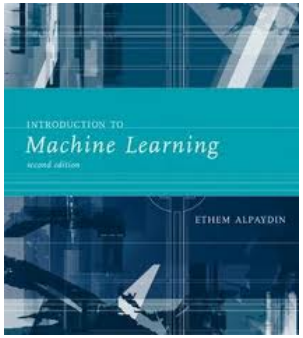


- R. Baeza-Yates, B. Ribeiro-Neto. *Modern Information Retrieval: The Concepts and Technology Behind Search*, ACM Press, 2011

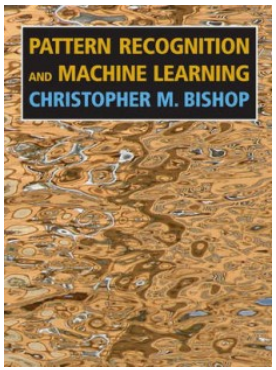


Textbooks – Machine learning

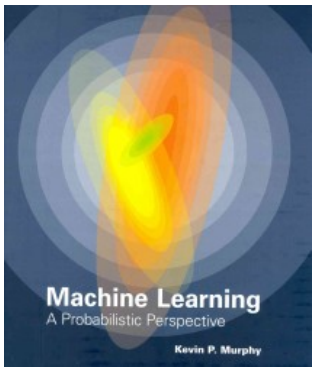
- Ethem Alpaydin: *Introduction to Machine Learning*, MIT Press, 2009



- Christopher Bishop: *Pattern Recognition and Machine Learning*, Springer, 2007



- Kevin P. Murphy: *Machine Learning: A Probabilistic Perspective*, MIT Press, 2012



Exams

- No exams

Theory: study assignments and in-class quizzes

- Together, in class:
  - 45 minute introduction to a topic
  - study assignment handed out
- You, at home:
  - do the study assignment
  - prepare for the in-class quiz and self-check the learning outcomes
- Together, in the next class:
  - 15 minute QA session to resolve unclarities
  - 15 minute in-class quiz (6 points)
  - 15 minute discussion session to dive deeper (1 bonus point)

Applications: paper reading and discussions

- Together, in class:
  - 15 minute introduction to a topic
  - pointer to the paper
- You, at home:
  - read the paper
  - submit the paper summary and questions (1 point)
- Together, in the next class:
  - 15 minute paper comprehension quiz (5 points)
  - 30 minutes going through the paper together
  - 15 minutes paper scoring
  - 15 discussion session (1 bonus point)

Lab assignments

- Hands-on experience of using the latest NLP tools in Python
- Complements the lectures and prepares you for the project
- You need to solve and submit your assignments – no demonstrations
- 3 assigments for 3 points each

Lab schedule

- Assignment 1: Preprocessing (week 3)
- Assignment 2: Classification and sequence labeling (week 5)
- Assignment 3: Neural NLP (week 7)

Project assignments

- Groups of 3 students work together on a focused TAR topic:
  1. Study related work
  2. Prepare the data (collect, clean, annotate)
  3. Implement the system
  4. Evaluate the system
  5. Write up a project report
  6. Give a 10-minute presentation + demonstration
- Expected workload: 44 hours per person over 11 weeks
- Score based on the results, project report, and presentation
  - Group score, to be distributed among and by the group members
  - Maximum differential between members' points is $2\times$ (e.g., if one member has 50, the others must have at least 25)
- Project assignments $\Rightarrow$ soon available at the course website

Project report

- In a form of a scientific paper
- Min. 3 pages and max. 4 pages + 1 page for references
- Structure:
  1. Intro: explain the problem, the motivation, and what you did
  2. Related work: short task overview, similar algorithms/systems
  3. Description of your system
  4. Evaluation of your system
  5. Conclusion
  6. References
- Preferably in English (but we won't insist ☺)
- To be typeset in LaTeX (we provide a template)
- Reviewed by teaching staff + TAs

Project report

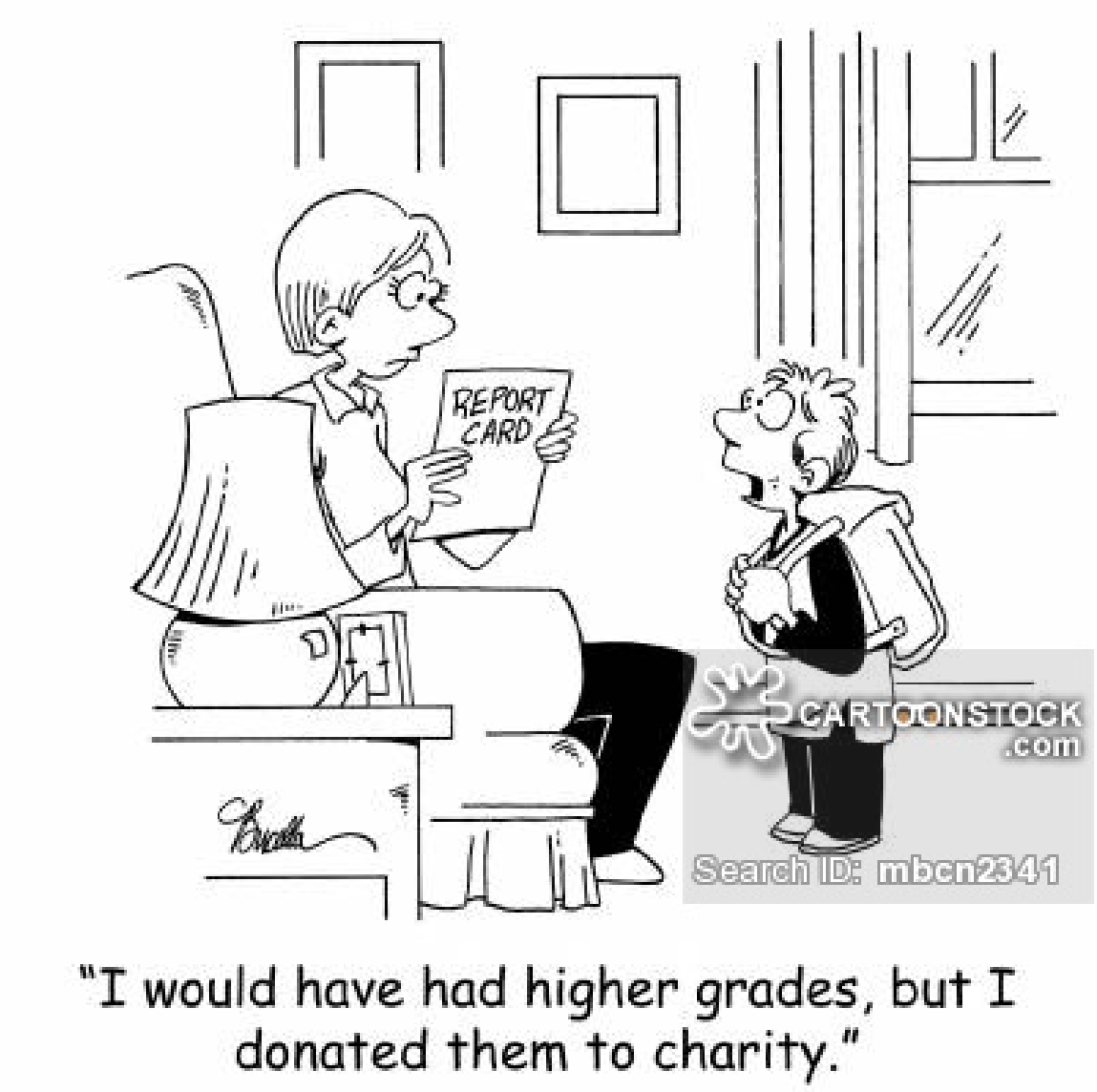
- Published on-line in the now renowned "TAR Course Project Report" series (with your consent)
  - Take a look at TAR 2021 PR, TAR 2020 PR, TAR 2019 PR, TAR 2018 PR, TAR 2017 PR, TAR 2016 PR, and TAR 2014 PR
- We highly encourage you to write your report in English and let us publish it on-line
  $\Rightarrow$ gives great EXPOSURE to your work
  $\Rightarrow$ future employers will FIGHT over you! ☺

Grade breakdown

	Continuous		Exam	
	Threshold	% grade	Threshold	% grade
Theory	50%	18%		
Labs	50%	9%		
Applications	50%	18%		
Project		55%		

If you fail to pass theory or reading threshold, you may retake the quiz (but you may retake at most 2 quizzes).

Grading



Excellent (5)	**89**
Very good (4)	**76**
Good (3)	**63**
Acceptable (2)	**50**

Project work schedule

- March 4: Topics announced
- March 4–13: Bidding for topics (in teams)
- March 14: Topics assigned
- Apr: Project progress checkpoint "alpha"
- May: Project progress checkpoint "beta"
- June: Project submission deadline (no extension!)
- June: Project presentations

ECTS breakdown

- This course is worth 5 ECTS credits = 150 hours of work
- 44 hrs for the project
- 26 hrs for paper reading
- 26 hrs for the lectures
- 36 hrs for theory self-study
- 10 hrs for lab sessions
- 8 hrs for project presentations

You now...

1. Know what this course is about and be glad that you've enrolled it
2. Know what are the topics that we want to cover
3. Know what you need to do to earn your credits