Introduction to Automatic Learning

Introduction

Contents

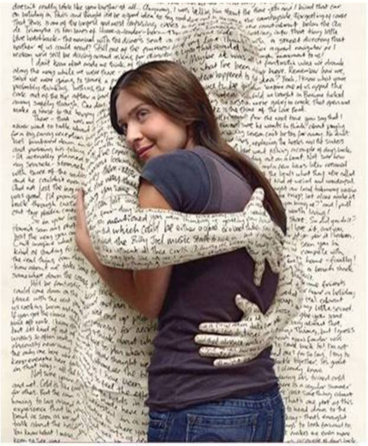
1. Motivation
2. Applications
3. Process
4. Data
5. Learning paradigms
6. Bibliography

Motivation

“We are drowning in data,

but starving for knowledge!”

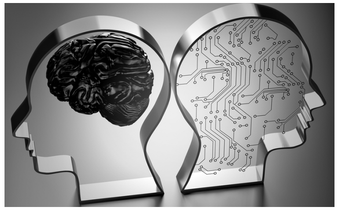
(John Naisbitt)



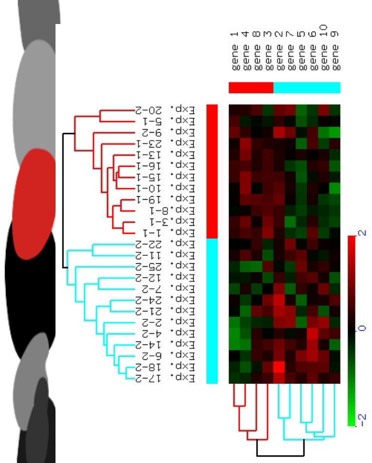
Source: http://kamafig.wordpress.com

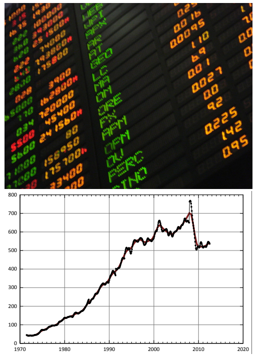
Motivation

Automatic learning: gain knowledge by means of automatic data processing.



Applications





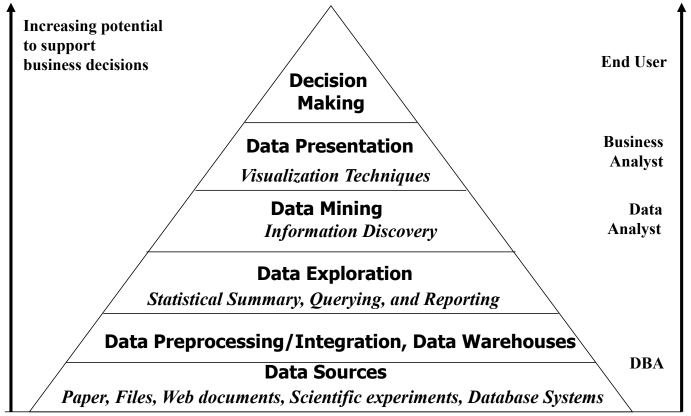
Applications

Applications in Natural language processing and understanding for decision making

- Alzheimer prevention
- Social bots
- Plagiarism detection
- Autism communication bots for education
- Bioinformatics: exploring human genome [Guan et al., 2019]
- Hate crime detection [Nobata et al., 2016]

Process

Data mining



Source: [Han et al., 2011]

Process

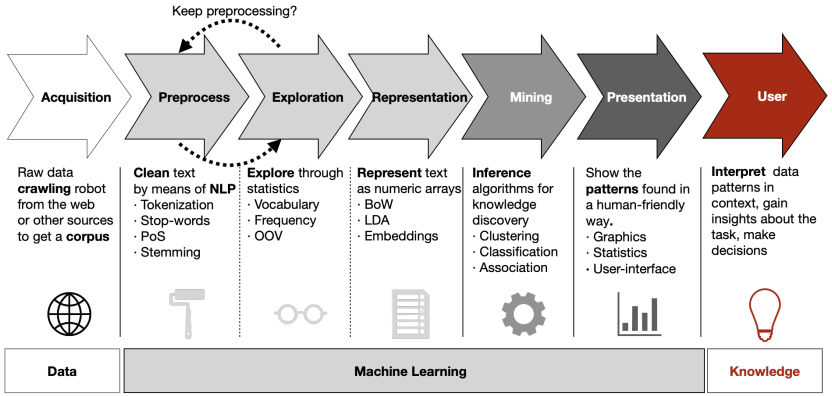
Text mining

What's challenging about text mining? [Weiss et al., 2015, chap 1]

- Numbers vs Text
- Structured vs Unstructured data
- Natural language encloses complex patterns hardly ever regular
  - Pragmatics
  - Psychology
  - . . . Syntax, Semantics, Morphology

Process

Text mining



Keep preprocessing?

Acquisition → Preprocess → Exploration → Representation → Mining → Presentation → User

Acquisition: Raw data **crawling** robot from the web or other sources to get a **corpus**

Preprocess: **Clean** text by means of **NLP**
· Tokenization
· Stop-words
· PoS
· Stemming

Exploration: **Explore** through statistics
· Vocabulary
· Frequency
· OOV

Representation: **Represent** text as numeric arrays
· BoW
· LDA
· Embeddings

Mining: **Inference** algorithms for knowledge discovery
· Clustering
· Classification
· Association

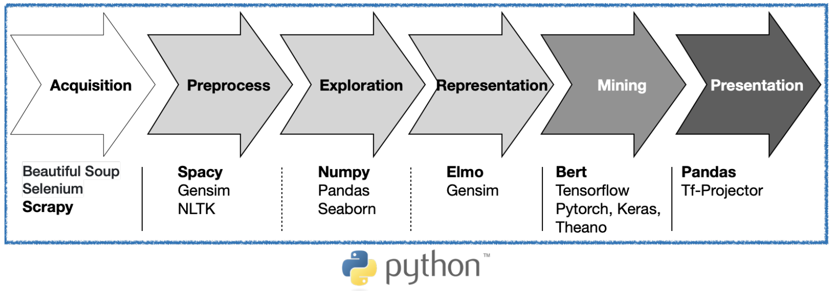
Presentation: Show the **patterns** found in a human-friendly way.
· Graphics
· Statistics
· User-interface

Interpret data patterns in context, gain insights about the task, make decisions

Data | **Machine Learning** | **Knowledge**

Process

Text mining



Acquisition	Preprocess	Exploration	Representation	Mining	Presentation
Beautiful Soup Selenium **Scrapy**	**Spacy** Gensim NLTK	**Numpy** Pandas Seaborn	**Elmo** Gensim	**Bert** Tensorflow Pytorch, Keras, Theano	**Pandas** Tf-Projector

Data

Operative description of the data

Operative description of the data [Witten et al., 2016, chap. 2]

- Data-set: sample, a set of instances (e.g. a collection of e-mails)
- Instance: an individual example in the data-set (e.g. one e-mail)
- Attributes: descriptive features by which we define the instances

Data

Operative description of the data

Data formats: text vs binary

- `.csv`
- `.arff`
- `.xml`
- `.json`
- ...

Data

Operative description of the data

Practice

Hands on training with Weka GUI. . .

- Header vs Data
- Comments
- Number of attributes
- Attribute type
- Missing values
- Enough instances? Enumerating all the possibilities.
- Supervised classification
- Intuition about correlated variables graphically
- Intuition about attribute selection for supervised classification
  - feature X and class show correlation: Good/Bad?
  - feature X and feature Y show correlation: Good/Bad?

Data

Data acquisition: ethical issues

Data acquisition is bound to regulations [Witten et al., 2016, sec. 1.7]

- What is the purpose of the data acquisition?
- Who can access the data?
- Are there caveats in the use of the data?
- Are the resources put to good use?
- Anonymisation / De-identification / Dis-aggregation (a hectic research field within NLP)

Data

Data acquisition: ethical issues

Exercise

- Think about your master thesis select a domain and task
- Find available corpora (text data) in repositories and also in research articles
- Describe the original data-format
- What would be an instance in your task?
- What kind of features would you use?
- Did you get enough data?
- Enumerate 10 conferences/journals to present your methods and results

Learning paradigms

What a machine can “learn” from data [Witten et al., 2016, sec. 1.3, chap. 4]

- Clustering (unsupervised)
  - Descriptive
  - Group instances
  - e.g. author attribution
- Classification or supervised learning
  - Predictive
  - Predict the value of a particular attribute (class)
  - e.g. spam classification
- Association
  - Relational
  - Predict the value of an arbitrary attribute (or combination)
  - e.g. basket analysis
- Semi-supervised learning

Learning paradigms

Practice

Hands on training with Weka GUI. . .

- Clustering
- Association
- Classification
- Attribute selection

Bibliography I

Guan, M., Cho, S., Petro, R., Zhang, W., Pasche, B., and Topaloglu, U. (2019).
Natural language processing and recurrent network models for identifying genomic mutation-associated cancer treatment change from patient progress notes.
JAMIA open, 2(1):139–149.

Han, J., Kamber, M., and Pei, J. (2011).
Data Mining: Concepts and Techniques.
Morgan Kaufmann, 3rd edition.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (2016).
Abusive language detection in online user content.
In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Weiss, S. M., Indurkhya, N., and Zhang, T. (2015).
Fundamentals of predictive text mining.
Springer.

Bibliography II

Witten, I. H., Frank, E., and Hall, M. A. (2016).
Data Mining: Practical Machine Learning Tools and Techniques.
Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, 4 edition.