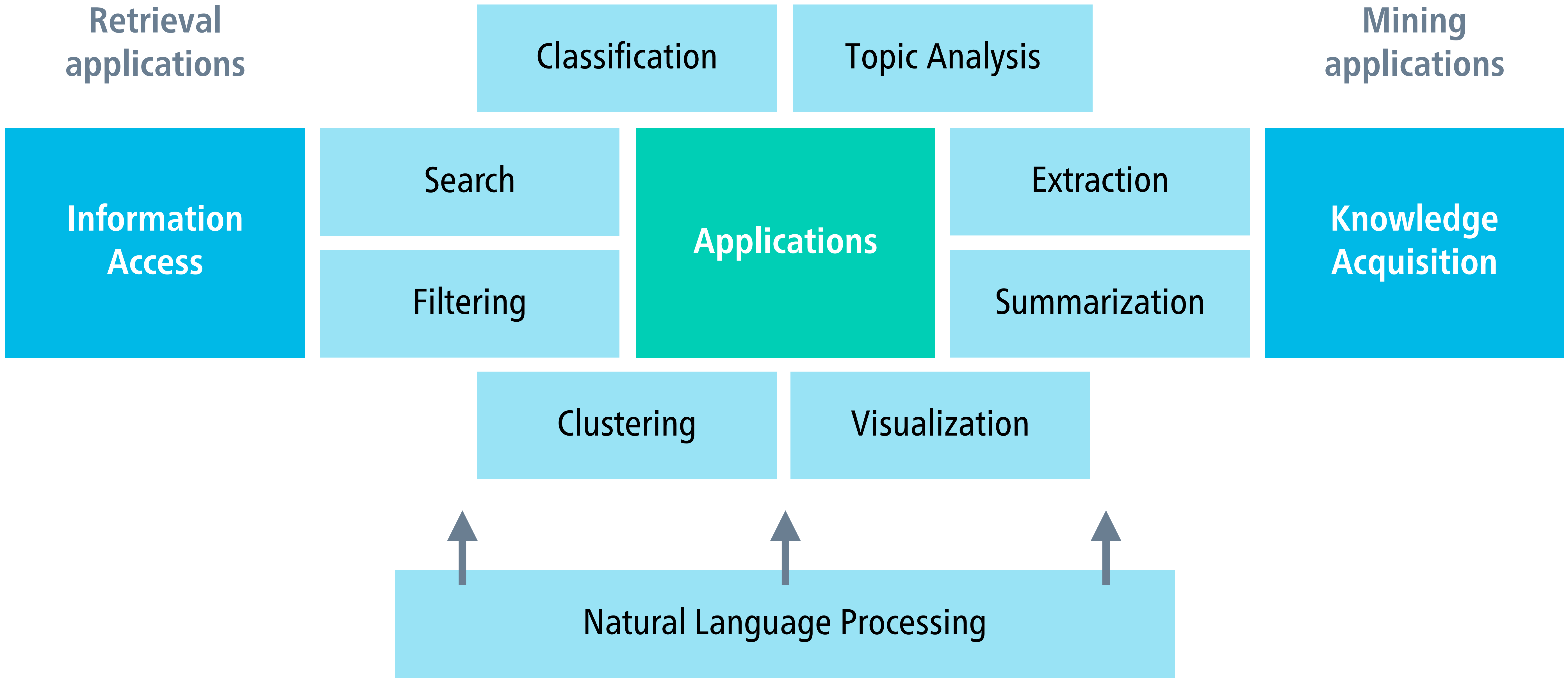
732A81/732A92/TDDE16 Text Mining (2022)

Project kick-off

Marco Kuhlmann
Department of Computer and Information Science

LINKÖPING UNIVERSITY

Conceptual framework for text mining



Adapted from Zhai and Massung (2016)

	Monday	Tuesday	Wednesday	Friday
W44		LEC Information Retrieval	LAB Information Retrieval	LAB Information Retrieval
W45	LEC Text Classification	Individual Supervision	LAB Text Classification	LAB Text Classification
W46	Individual Supervision	LEC Clustering & Topic Modelling	LAB Clustering & Topic Modelling	LAB Clustering & Topic Modelling
W47	LEC Word Embeddings	Individual Supervision	LAB Word Embeddings	LAB Word Embeddings
W48	Individual Supervision	LEC Information Extraction	LAB Information Extraction	LAB Information Extraction
W49	Individual Supervision	LEC Project kick-off	Individual Supervision	Individual Supervision
W50	Individual Supervision	Individual Supervision	Individual Supervision	Individual Supervision
W51	Individual Supervision	Individual Supervision	Individual Supervision	
W02	Individual Supervision	Individual Supervision	Individual Supervision	Individual Supervision

Examination of the project component

Examination

	Computer labs	Project
ECTS credits	3 credits	3 credits
To be done	in pairs	individually
Deliverables	notebooks + diagnostic test	written project report
Grading	Pass/Fail	ECTS, U345

Knowledge requirements for the project component

- You identify and formulate a *substantial* text mining problem *with some help from a teacher*.

- You implement and apply *suitable* text mining methods, analyse experimental results *with appropriate evaluation methods*, and summarise them *with well-developed judgements*.

- You clearly present and discuss the conclusions of your work.

Form of the examination

- The project component is examined by a written report.
- Detailed instructions for the written report and information about its assessment are available on the course website.

  Instructions for the project report

Formal requirements – highlights

- length: 4–8 pages of content + unlimited references

  standard template

- standard conventions of academic writing

  polished language, references, use of mathematics where appropriate

- due date: 2023-01-14 (plus usual extension)

  additional examination dates: 2023-03-17, 2023-08-27

Example projects from previous years

What people like and dislike about the Paperwhite

- Many companies are interested in finding out about what their customers think about their products.

  sentiment analysis

- What do Text Mining methods tell us about what people like and dislike about the Amazon Kindle Paperwhite?

- Collect a data set, train and compare different types of classifiers, identify the most informative features.

Quantifying text emotiveness

- The notion of emotiveness refers to how emotionally engaged a writer or speaker was while producing a text.
- There are psycholinguistic theories about how emotiveness can be measured in text.

  Trager coefficient, aggressiveness coefficient, readiness to action
- Part-of-speech tag the inaugural speech corpus, analyse the emotiveness of the speeches over time, explain the results.

Sentiment analysis of Twitter data

- Can we use text classification to predict the sentiment of a tweet in relation to a given topic?

- Build a ‘silver standard’ based on the hypothesis that :) indicates a positive tweet while :(indicates a negative tweet.

  noisy labels

- Collect data using the Twitter API, preprocess the data, train different text classifiers, identify most informative features.

  Adele, Adidas, Burger King, Ryanair, Taco Bell, …

Job market analysis for statistics and data mining

- Which areas can one work in as a data miner? Which personal traits and qualifications are sought in each area?

  technical, bank, insurance, academic work, business

- Collect a data set consisting of job ads, preprocess the data, train a topic model, analyse the results (subjectively).

  How can one make an informed choice regarding the number of topics?

Answering multiple choice questions

- Build a system for automatic answering of multiple choice questions based on information retrieval.
- Collect data from a school textbook (8th grade) and Wikipedia and build a knowledge base of documents.
- Find the k most relevant documents for the question and the k most relevant documents for every possible answer.
- The score of a potential answer is the sum of the tf–idf similarities of the most relevant documents.

Predicting drug interactions

- Build a binary classifier that can warn doctors when two drugs interact, e.g. whether there is an adverse effect.
- Collect data from official drug descriptions, which list adverse effects on the substance (but not the drug) level.
- Explore both supervised and unsupervised learning.
- Evaluate using a manually constructed gold standard, constructed in consultation with a doctor.

Family tree extraction for Tolkien’s world

- Uses the Lord of the Rings Wikia to automatically extract family trees for the characters in Tolkien’s world.

- Evaluate the results of the extraction procedure using the infoboxes section of each character page.

- Low precision and recall – this should work much better!

Tips and tricks

Tips and tricks

- Many of you will have started the project by looking for data sets you find interesting and want to know more about.

- Now it is time to spend some time to actually look at the data and related work. Based on that, you may want to switch data set!

  ACL Anthology

- Be incremental. Collect ‘small’ results. Once you feel that you have enough, try to integrate them into a big picture.

  Examples: replicate previous work, validate your models

How to get data?

- Ready-made datasets from shared tasks, data science competitions, public providers

  RepEval 2017 Shared Task, Kaggle, Riksdagens öppna data

- Data from companies made available via APIs

  Twitter, Musixmatch

How to process data?

- Use existing software libraries

  pandas, spaCy, NLTK, scikit-learn, Gensim

- Use R (or whatever ecosystem you are most comfortable with) if you find that it's easier for you!

  No requirement on the programming language.

How to validate?

- intrinsic evaluation using easy-to-calculate measures such as accuracy, precision, recall, topic coherence, perplexity, …
- extrinsic evaluation, for example by embedding the component into a larger system or doing a user study
- theory-based evaluation: do the results confirm the hypotheses; how well do the results fit the facts

How to get help?

- Pitch your project idea to us!
- We will be offering one-to-one feedback opportunities throughout the rest of the course.

  minus Christmas break
- You can also send us an email, but note that we will be prioritising personal contact.