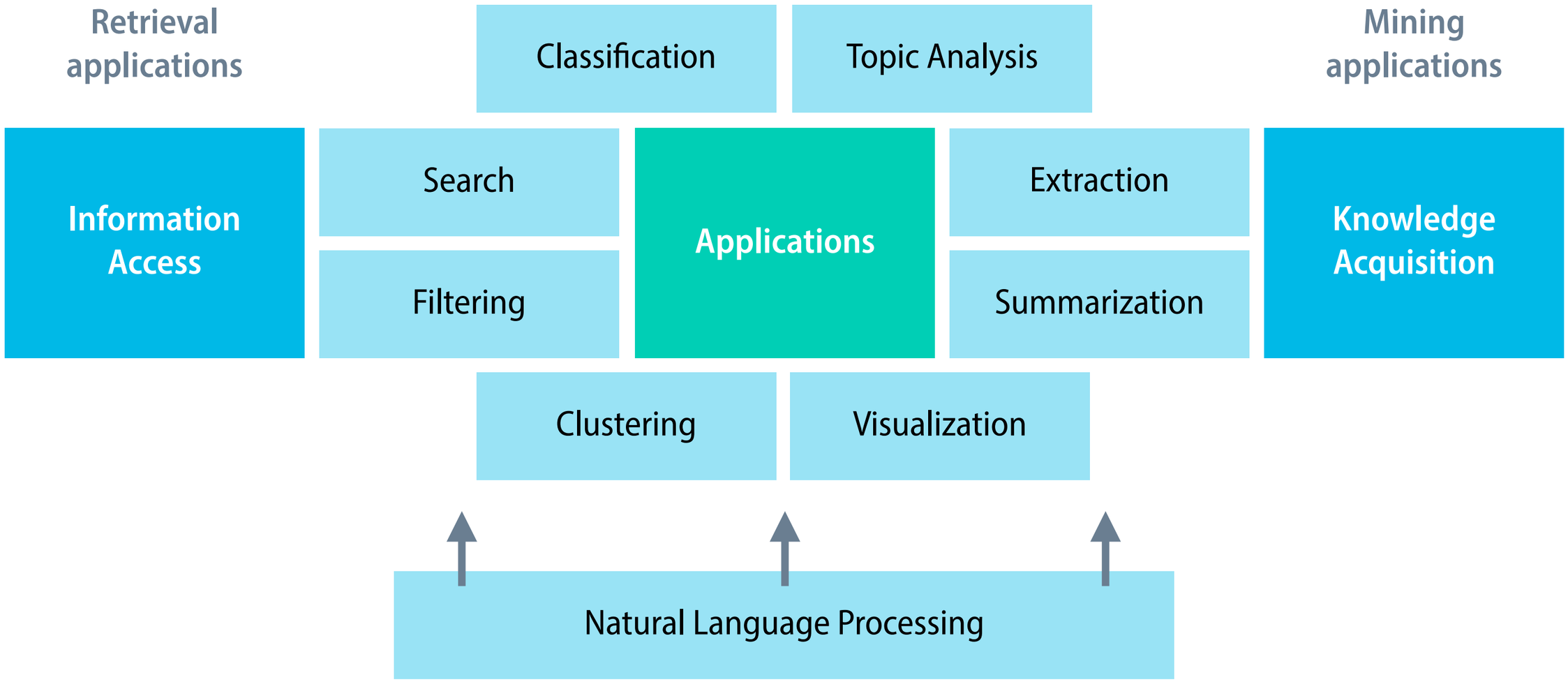
732A92/TDDE16 Text Mining (2020)

Project kick-off

Marco Kuhlmann
Department of Computer and Information Science

Conceptual framework for text mining



Adapted from Zhai and Massung (2016)

	Monday	Tuesday	Wednesday	Friday
W45	LEC Course Introduction	LEC Information Retrieval	LAB Information Retrieval	LAB Information Retrieval
W46	Individual Supervision	LEC Text Classification	LAB Text Classification	LAB Text Classification
W47	Individual Supervision	LEC Clustering and Topic Modelling	LAB Clustering and Topic Modelling	LAB Clustering and Topic Modelling
W48	Individual Supervision	LEC Word Embeddings	LAB Word Embeddings	LAB Word Embeddings
W49	Individual Supervision	LEC Information Extraction	LAB Information Extraction	LAB Information Extraction
W50	LEC Project kick-off	Individual Supervision	Individual Supervision	Individual Supervision
W51	Individual Supervision	Individual Supervision	Individual Supervision	Individual Supervision
W01				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
Form of hand-in	notebooks	written project report
Grading	Pass/Fail	ECTS, U345

Knowledge requirements for the project component

- You apply the text mining techniques covered in the course to *self-defined* problems.

  specific task or research question

- You evaluate the performance of methods and systems with *appropriate* validation methods.

  need to make an active choice about how you want to validate your results

- You interpret your results with *well-developed* judgements.

  nuanced arguments based on relevant material, examples, clearly written

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

- between 2,000 and 4,000 words (main text)

  4–8 pages of single-spaced 12 pt Times New Roman, 2.5 cm margins

- standard conventions of academic writing

  polished language, references, use of mathematics where appropriate

- due date: 2021-01-15 (plus usual extension)

  additional examination dates: 2021-03-18, 2021-08-28

Formal requirements – highlights

- title page with descriptive title, full name, and LiU-ID

  Do not use a generic title such as ‘Text Mining Project Report’!

- abstract with a concise summary of the project (max. 200 words)

  The abstracts of all projects will be published on the course website.

- references to all sources (text books, articles, code)

  You may choose whatever style you are comfortable with.

- link to GitHub/GitLab repository with your code

Suggested structure for the report

1. **Introduction.** Introduce the problem that you have addressed in your project. What did you do? Why did you do it?

2. **Theory.** Present relevant theoretical background, and in particular the machine learning models that you have used.

3. **Data.** Present your data. How does it look like? Where did you get it from? What pre-processing did you do, if any?

4. **Method.** Explain how you carried out your study. Aim to be detailed enough for others to reproduce your results.

Suggested structure for the report

5. **Results.** Present your results in an objective way. Use tables and charts, but do not forget to summarise your results text form.

6. **Discussion.** Analyse your results and discuss the possibilities and limitations of your approach. Compare your study to related work.

7. **Conclusion.** Based on your results and their analysis, what new knowledge do you take away from the project?

8. **References.** Present a complete list of references. Choose a bibliographic style and stick to it.

Assessment of the report

- Five criteria, adapted from the review form of the *Transactions of the* ACL, a major journal in text mining and NLP.

- For each criterion, I will assign an integer score between 0 (does not meet expectations) and 9 (exceeds expectations).

  descriptors available for scores 0, 1 (meets expectation), and 9

- To get a passing grade, your score *for each criterion* must be at least 1 (meets expectation).

Grading

Your grade is calculated from two component scores:

- **Component 1** is the score for a single criterion: *Soundness and correctness*.
- **Component 2** is the median of the scores for the remaining four criteria: *Clarity*, *Related work*, *Creativeness*, and *Substance*.

The grade is calculated according to the table on the next slide.

Component 1	Component 2	Grade 732A92	Grade TDDE16
1	1	E	3
1	5	D	3
1	9	C	4
5	1	D	3
5	5	C	4
5	9	B	5
9	1	C	4
9	5	B	5
9	9	A	5

Criterion 1: Soundness and correctness

Is the technical approach sound and well-chosen? Are the claims made in the report supported by proper experiments, and are the results of these experiments correctly interpreted?

o Troublesome. There are some ideas worth salvaging here, but the work should really have been done or evaluated differently.

1 Fairly reasonable work. The approach is not bad, the evaluation is appropriate, and at least the main claims are probably correct.

5 Generally solid work, although there are some aspects of the approach or the evaluation that I am not sure about.

9 The approach is fully apt and all claims are convincingly supported. The report discusses the limitations of the work.

Criterion 2: Clarity

Is it clear what was done in this project, why it was done, and how it was done? Is the report well-written and well-structured?

0 There are some important questions about the method, results, or analysis that even expert readers are not able to resolve.

1 Any student who has successfully completed the course should understand what was done in this project, why it was done, and how.

9 The report is well-polished. In regard to clarity of presentation, it would be acceptable for an academic conference.

Criterion 3: Related work

Does the report show awareness and understanding of related work documented in scientific sources? Is it clear where the work done in the project sits with respect to that related work?

o The report shows little awareness and understanding of related work. References to scientific sources are missing or incomplete. There is no account of how the work done in the project compares to the related work.

1 The report shows some awareness and understanding of related work. Scientific sources are adequately referenced. The relation between the work done in the project and the work documented in the scientific sources is clear.

9 The report features a precise and enlightening comparison with related work. References are complete and consistently formatted. The majority of scientific sources are peer-reviewed research articles.

Criterion 4: Creativeness

How creative is the project? For example: Does the project target a new problem? Does it contribute a new data set? Does it use any machine learning models that were not covered in the course?

0 There are no creative elements in this project. The project is essentially a repetition of one of the lab assignments.

1 The project contains creative elements.

9 There are many creative elements in this project. The project goes significantly beyond what has been covered in the course.

Criterion 5: Substance

Based on the report, does this project have enough substance, or would there have been room for more ideas, results, or analysis? (The expected amount of work for the project module is 88 hours.)

0 Seems thin. I (the examiner) would have expected significantly more ideas, results, or analysis for a project with this timeframe.

1 Represents an appropriate amount of work for a project in this course.

9 Contains significantly more ideas, experiments, and analysis than what I (the examiner) would have expected for a project with this timeframe.

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 kinds 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.

Detecting hate speech in tweets

- Build a classifier that can detect whether a tweet contains hateful speech or is offensive in some other way.
- Use several document representations, including tf–idf (with several n-gram sizes) and word2vec.
- Explores a wide range of classification techniques, including logistic regression, XGBoost, and CNN.

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

- A good way to start the project is to pick a data set that you find interesting and want to know more about.
- Spend some time to actually look at the data. What have others done with it? What could you do with it?
- Be incremental. Collect ‘small’ results. Once you feel that you have enough, try to integrate them into a big picture.

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

- Scrape data using web scraping tools such as Scrapy

  may require preprocessing, manual annotation – licenses?

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
- subjective evaluation of how easy it is to explain the results, how well the results fit the facts, how well they fit a theory

How to get help?

- Pitch your project idea to me!
- I will be offering one-to-one feedback opportunities throughout the rest of the course.
- You can also send me an email, but note that I will be prioritising personal contact.