Introduction

[DAT640] Information Retrieval and Text Mining

Krisztian Balog
University of Stavanger

August 19, 2019

About me

- Professor at the University of Stavanger (0.8fte)
- Professor II at NTNU (0.2fte)
- Visiting Staff Faculty Researcher at Google (0.2fte)
- History
    - Spent last year on a sabbatical at Google Research, London, UK
    - Joined UiS in 2013 as Associate Professor, promoted to full professor in 2016
    - PhD from the University of Amsterdam (2008)
- More: `https://krisztianbalog.com/`

About the course

- “Text data access”
- **Information retrieval** (search engines)
  - Analysis, organization, storage, and retrieval of information
- **Text mining** (text analytics)
  - Deriving high-quality information from textual data by analyzing trends and patterns

Prerequisites/requirements

- No formal prerequisites, **but** you are expected to know
    - Programming (in Python)
    - Databases (basic concepts)
    - A bit of statistics
- **Hands-on – Bring your own device** (laptop)
    - Python 3.6+ (Anaconda distribution)
    - GitHub user and git client (e.g., GitHub Desktop)

Course organization

- Course runs from week 34 to week 46
- Regular schedule
  - Monday and Tuesday are lectures (with me)
  - Wednesdays are labs (with student assistant)
- **There will be exceptions to the regular schedule!**
  (will be announced on Canvas)

Assignments

- Three main assignments, with the first two divided into sub-assignments
  - That is, 5 assignments in total: 1A, 1B, 2A, 2B, 3
- Can work alone or in pairs
  - But you cannot work with the same person on more than one main assignment
  - A sign-up form will be made available for each assignment
    - I prefer to work with either (1st choice), (2nd choice), or (3rd choice)
    - I prefer to be teamed up with a random person
    - I prefer to work alone
- 2-3 weeks per assignment
- Deadlines are strict, no extensions, no exceptions!
- Assignments account for **40% of the grade**

Exam

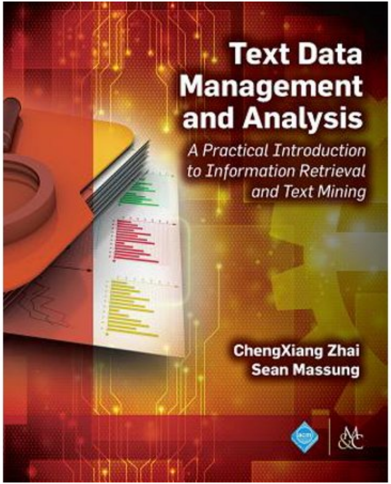
- Digital exam (Inspera)
- Mixture of exercises, multiple choice, and essay questions
- **60% of the grade**
- In previous years, it was open book
- Trial exam at the last week of the course

Resources

- One-stop shop: `http://bit.ly/uis-dat640`
  - Course schedule and curriculum
  - Lecture slides (will be made available before each lecture)
  - Example code
  - Assignments
- Canvas is (only) used for announcements

Textbook

Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining (Zhai and Massung), ACM and Morgan & Claypool Publishers, 2016.



Getting help

- Wednesday labs are for working on the assignments. This is **the** time to get help!
- Lecture breaks
- **Office hours**: (most) Wednesdays, 14:00-15:00 KE-433
- No drop-ins unannounced! Make an appointment via email.
- For all course-related matters, the **primary contact email** is **`dat640help@gmail.com`**

FAQ

- **Is it obligatory to attend the lectures?**
  No. However, it is highly recommended in order to get the full learning experience. Also, everything that is covered in the lectures may be asked back at the exam.
- **Is it obligatory to attend the labs?**
  No. However, this is the time and place to work and get help on the assignments.

MSc projects (DAT620) and thesis

- I have a couple of DAT620 projects available
- I also get overwhelmed by requests for supervising MSc theses
- Both of these require that you take this course (and get a good grade)