INFO319:
Organisation of the course

Background

- INFO310 in the past:
  - semantic technologies, some “big data” sources
- INFO319 in the past:
  - Big Data for Emergency Management (BDEM)
  - an INTPART project (2018-2021)
  - old INFO310 morphed into INFO319
    - big data for emergency management
    - critique: more emergency management than big data
- Revised INFO319:
  - more focussed on big data
  - news production as main application domain

Goal and contents

«The course provides the theoretical and technical foundations for managing and rapidly exploiting big data sets, for example data originating from public sources, social media/crowdsourcing, and the Internet/Cloud of Things. It covers general theories and technologies for big data, including sourcing, analysis, curation, processing, use, and evaluation. The course focusses on the exploitation of big data in a selected domain, such as media content production. It gives the academic background for supervised research on uses of big data in that domain. The course involves development work using selected technologies and standards for big data. Examples of research and research methods in the area will also be presented and discussed.»

Theme: big data for news purposes (but not exclusively). Textual and graph data (not images/video). Based the News Angler project.

Knowledge and skills

- Knowledge:
  - «Upon completion of the course the candidate shall understand central concepts, standards, and technologies for big data know about current research and industry trends in big data understand how big data can be exploited in the selected domain know about current research and industry trends in big data in the selected domain know the relevant research methods for the area»
- Skills:
  - «Upon completion of the course the candidate shall be able to use central technologies and tools for managing and using big data exploit big data in the selected domain prepare and evaluate uses of big data in practice»

Organisation

- Sessions
  - 8 bi-weekly 6-hour sessions
  - lectures and discussions, work with assignments, presentations
  - physical presence, no streaming/recording
  - participation at 80% of course seminars is mandatory
- Exercises between the sessions
- Group programming project (1-3 persons, 3 recommended)
- Individual essay
- *Compulsory requirements are only valid the same semester*

Readings and themes

- “Theory track”:
  - Rob Kitchin (2021):
    The Data Revolution: A Critical Analysis of Big Data, Open Data and Data Infrastructures. Sage.
  - architecture, privacy and the GDPR
    (EU’s General Data Protection Regulation)
- Technology track:
  - Bill Chambers and Matei Zaharia (2018):
    Spark: The Definitive Guide – Big Data Processing Made Simple. O'Riley.
  - selected cloud and big data techniques and tools:

Sessions

1 Introduction to big data. Big-data processing. Spark

2 More about Spark. Data sources. Twitter
   - guest presentation on Twitter-based projects at MediaFutures

3 Streaming Spark. Kafka
   - guest presentation about deep image analysis

4 Big data architecture. Cloud, NREC and Openstack
   - guest presentation about News Hunter and big-data architectures

5 Cloud management. Terraform and Ansible. Docker

6 Societal issues. Privacy. GDPR
   - guest presentation about big-data quality

7 Essay presentations

8 Project demonstrations

Technology

- Linux (Ubuntu) and Python centric
- Virtual instances in NREC (Norwegian Research and Education Cloud)
  - http://nrec.no
- Starts with a quota of 20 instances in various sizes
  - expandable on demand
  - GPU instances should be available

Assessment

- Portfolio (55%)
  - practical assignment in groups (group programming project)
    - can be an extension of the exercises
    - proposals by October 12th 1500
  - individual, theoretical essay
    - can be about the group programming project
    - proposals by October 12th 1500
- Oral presentations (15%)
    - essay presentations: November 24th
    - project presentations: December 8th
- Written exam (30%)
  - 3 hours, December 19th 0900

The Programming Project

Group programming project

- The project shall develop an application that uses semantic technologies. Development and run-time platform is free choice, as is programming language. The project should be carried out in groups of three and not more. Working individually or in pairs is not recommended.
- Framed as a *design science research* project
  - ...more in later sessions
- Proposal: Wednesday October 12th 1500
- Presentation: Thursday December 8th

Example project types

- Detecting trends (topics, hashtags, …)
  - in tweets and/or other sources, in real time
- Detecting viral tweets
  - other platforms *most* welcome: e.g., Facebook, ...
- Mining social connections from tweets and other sources
- Making social media content more findable
- Dark-entity detection (named entities that are not yet described in encyclopedia like Wikidata, DBpedia, GeoNames, …)
- Detecting tweets and/or other items related to a story in progress
- Detecting situations that stir emotions (or conflicting emotions)

The Essays

Individual essays

- The essay shall present and discuss selected theory, technology and tools related to semantic technologies, backed by scholarly and other references
- Can be a reflection of the programming project, e.g.:
  - the idea, its potential impact, risks
  - privacy, data quality, security issues
- Proposal: Wednesday October 12th 1500
- Presentations: Thursday November 24th