IN 3120/4120 - Search Technology

Group 2
TA: Markus S. Heiervang

Today's plan

- Introduction
- Assignments & info
- Curriculum overview
- (advanced) Python coverage
- Getting started with the github repository
- Assignment pre-code walkthrough
- Assignment A: Inverted index and Postings merger
    - brief coverage of inverted index
- Q & A
- Bonus: Simple search engine

A little about me

- First year masters student: “Informatics: Language technology”
- Took IN3120 last year (got B)
- Former TA in the introuctory course: IN1000 - Introduction to Object-oriented programming

Contact details:

- markuhei@ifi.uio.no

You'll have a head start if you have taken these courses:

- +++IN2110 - Methods in language technology
- ++ IN3050/IN4050 - Introduction to Artificial intelligence and Machine Learning
- ++IN1140 - Introduction to language technology
- +IN3110/IN4110 - Problem solving in high level programming languages
- … (other language technology courses)

It is also recommended to have taken (or take next to IN3120)

- IN2010 - Algorithms and data structures / IN2040 - Functional programming

As we will be considering complexity of some algorithms

Some notes about this course

- If you're not already a python expert, you can expect to get better at python after having done the assignments
- You might think this course includes crawling the web. It does not
- Some of the assignments might have you stuck. Don't be afraid to ask for help from your peers. Don't be afraid to ask me for help either

- The curriculum in this course is really broad, and the assignments only cover a fraction of it. I would strongly advise to work on the old exams, as many of the questions there are likely to reappear in a different form

Curriculum

- Algorithms and data structures for strings and search
  - PageRank
  - Inverted index
  - TF-idf, cosine similarity
  - Suffix array
  - Tries og Aho-corasick
  - Heaps law, Zipf's law
  - Document vectors, (word embeddings?)
  - Edit distance, jaccard similarity, DCG...
- Machine learning
  - Validation
  - Classification: Rocchio, K nearest neighbors, SVM
  - Shallow knowledge on neural networks
  - Naive bayes
- Compression
  - Rice coding, VB coding, gamma and delta coding, PFOR, etc.
- Etc
  - Bloom filter

Math in this course

- Basic linear algebra: knowledge of vectors and matrices, dot product and vector distance
- Exponential equations
- Sigma notation and formulas for e.g. ml
- Partial derivatives ???

Additional note

- Some formulas (e.g. on lecture slides) might look complicated and very scary for people who have not yet understood them, though in reality, they are usually much simpler than they look.

Assignments

- Assignment A: Inverted index
- Assignment B: Suffix array and aho corasick
- Assignment C: Simple search engine
- Assignment D: Ranking
- Assignment E: Naive bayes classification

5 assignments, 2 weeks between each

All of them are on the github page

NLP lingo

- Document
  - A text, a webpage, a string
  - Can be separated in multiple fields (e.g. if)
- Corpus (plural: corpora)
  - A collection of documents
- Tokenization
  - To split a text into tokens, i.e. isolate words, punctuation, etc…
- Normalization
  - e.g. casting text to lowercase

Getting started

- https://github.com/aohrn/in3120-2020