IN2110: Språkteknologiske metoder

Introduksjon

Erik Velldal

Språkteknologigruppen (LTG)

19. Januar, 2022

UNIVERSITAS OSLOENSIS

MDCCCXI

Getting introduced

`https://nettskjema.no/a/242463`

Topics for today

- AI, NLP, ML — What are they?
  - Definitions
  - Applications
  - Historical review
- Outline of lectures
- Practical details
  - Syllabus
  - Obligatory assignments
  - Programming
  - Communication

Screencasting

- First two lectures will be via Zoom.
- Awaiting clarification wrt physical lectures for the remainder of the semester.
- Will screencast all lectures regardless.
- The schedule will link to the videos.
- Both physical + digital group sessions.

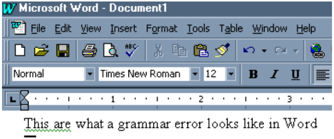
What is Natural Language Processing?

- Making computers ‘understand’ human language
- Aka language technology or computational linguistics
- Young and interdisciplinary field:
- Computer science + linguistics
- (+ cognitive science, statistics, machine learning . . .)
- Sub-field of AI.

NLP applications

NLP applications

- Grammar and/or spell checkers, auto-completion
- Machine translation
- Intelligent information extraction
- Summarization
- Sentiment analysis
- Q&A systems, dialog systems, and chatbots
- (Speech recognition and synthesis)
- Any application requiring an understanding of language. . .


Microsoft Word - Document1
File Edit View Insert Format Tools Table Window Help
Normal
Times New Roman
12
B I U

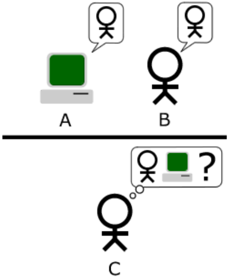
This are what a grammar error looks like in Word

What is AI?

- The term ‘AI’ coined by John McCarthy (Dartmouth Workshop, 1956).
  - *The science and engineering of making intelligent machines.*
  - *Every aspect of learning or any other feature of intelligence can be so precisely described that a machine can be made to simulate it.*

What is AI?

- The term ‘AI’ coined by John McCarthy (Dartmouth Workshop, 1956).
  - *The science and engineering of making intelligent machines.*
  - *Every aspect of learning or any other feature of intelligence can be so precisely described that a machine can be made to simulate it.*
- Alan Turing, 1950:
  - *I propose to consider the question, ’Can machines think?’*
- The Turing Test, based on the imitation game.
- Language understanding has always been central to AI.



AI in this course

- For our purposes: AI is a toolkit of methods for representation and problem solving, a bag of tricks.

Paradigm shifts in NLP (and AI at large)

- 50s–80s: mostly rule-based (symbolic / rationalist) approaches.
- Hand-crafted formal rules and manually encoded knowledge.
- (Though some AI research on neural networks in the 40s and 50s).

Paradigm shifts in NLP (and AI at large)

- 50s–80s: mostly rule-based (symbolic / rationalist) approaches.
- Hand-crafted formal rules and manually encoded knowledge.
- (Though some AI research on neural networks in the 40s and 50s).
- Late 80s: success with statistical ('empirical') methods in the fields of speech recognition and machine translation.
- Late 90s: NLP (and AI at large) sees a massive shift towards statistical methods and machine-learning.
- Based on automatically inferring statistical patterns from data.
- 00s: Machine-learning methods dominant.

Paradigm shifts in NLP (and AI at large)

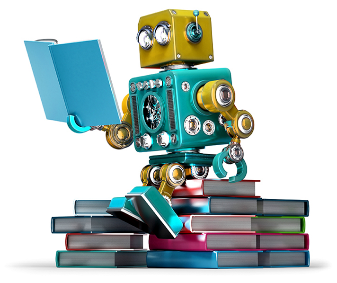
- 50s–80s: mostly rule-based (symbolic / rationalist) approaches.
- Hand-crafted formal rules and manually encoded knowledge.
- (Though some AI research on neural networks in the 40s and 50s).
- Late 80s: success with statistical ('empirical') methods in the fields of speech recognition and machine translation.
- Late 90s: NLP (and AI at large) sees a massive shift towards statistical methods and machine-learning.
- Based on automatically inferring statistical patterns from data.
- 00s: Machine-learning methods dominant.
- 2010–: neural methods and deep learning.
- Today, in the popular media, AI is mostly synonymous with ML.

The basis of empirical methods

Machine Learning

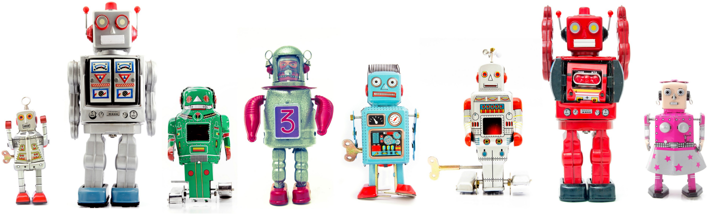
- *the study of computer algorithms that improve automatically through experience* (Tom Mitchell 1997).

- Similar to statistical data analysis, but the models are applied to solve a practical tasks rather than to describe the data.
- Goal: to learn from data.
- Not interested in simply learning by rote; want to generalize.
- A core element in the emerging field of *data science*.



Machine learning

- Learning = advanced counting of observations.
- Many different algorithms, but two main approaches:
- supervised (*veiledet*) and unsupervised (*ikke-veiledet*) learning.



Supervised learning (*Veiledet læring*)

- Requires training data; pre-defined examples of what we want the algorithm to learn.
- Learning from labeled data.

Supervised learning (*Veiledet læring*)

- Requires training data; pre-defined examples of what we want the algorithm to learn.
- Learning from labeled data.

Unsupervised learning (*Ikke-veiledet læring*)

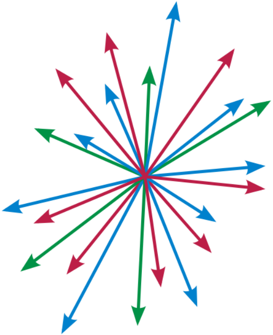
- Learning from unlabeled data: no pre-defined examples.
- The algorithm attempts to find structure in the data on its own.

Supervised vs unsupervised learning

- Currently we get most precise results with supervised learning.
- Typically requires manually labeled training data (= costly).
- A lot of research directed at making better use of unsupervised methods; we have much more unlabeled data available.
- A lot of fuzz about Big Data: great for training unsupervised methods or when applying a pre-trained supervised model.
- ML is no free lunch:
- The data is often more important than the algorithm.
- And related to this; how we choose to represent the data.

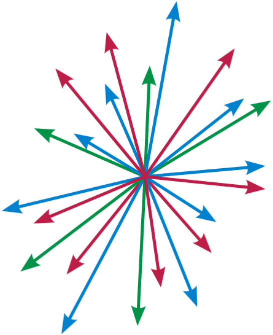
Overview of lectures

- `https://www.uio.no/studier/emner/matnat/ifi/IN2110/v22/`



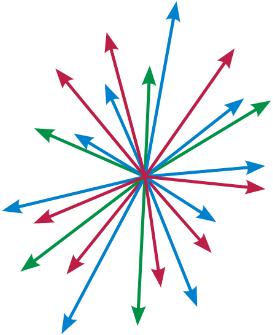
Overview of lectures

- https://www.uio.no/studier/emner/matnat/ifi/IN2110/v22/
- Vector space models (non-probabilistic ML)
- Representing documents
- Representing word meaning



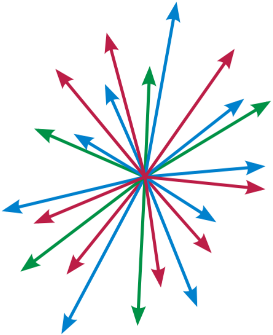
Overview of lectures

- `https://www.uio.no/studier/emner/matnat/ifi/IN2110/v22/`
- Vector space models (non-probabilistic ML)
- Representing documents
- Representing word meaning
- Classification (supervised learning)
- Clustering (unsupervised learning)



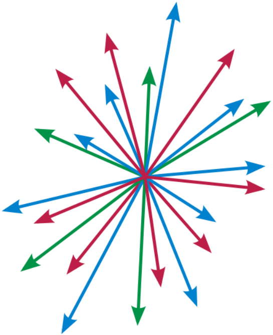
Overview of lectures

- `https://www.uio.no/studier/emner/matnat/ifi/IN2110/v22/`
- Vector space models (non-probabilistic ML)
- Representing documents
- Representing word meaning
- Classification (supervised learning)
- Clustering (unsupervised learning)
- Logistic regression
- Sequence classification



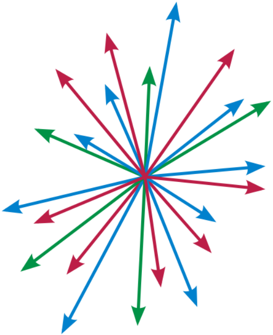
Overview of lectures

- https://www.uio.no/studier/emner/matnat/ifi/IN2110/v22/
- Vector space models (non-probabilistic ML)
- Representing documents
- Representing word meaning
- Classification (supervised learning)
- Clustering (unsupervised learning)
- Logistic regression
- Sequence classification
- Syntax and statistical parsing



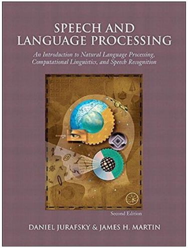
Overview of lectures

- `https://www.uio.no/studier/emner/matnat/ifi/IN2110/v22/`
- Vector space models (non-probabilistic ML)
- Representing documents
- Representing word meaning
- Classification (supervised learning)
- Clustering (unsupervised learning)
- Logistic regression
- Sequence classification
- Syntax and statistical parsing
- Applications
  - MT
  - Interactive systems



Reading list

- Selected chapters of the following books.
- Both are freely available online.
- Jurafsky & Martin:
  Speech and Language Processing
  3rd ed. draft of Jan. 12, 2022:
  `https://web.stanford.edu/~jurafsky/slp3/`
- Manning, Raghavan, & Schütze (2008):
  Introduction to Information Retrieval:
  `https://nlp.stanford.edu/IR-book/information-retrieval-book.html`

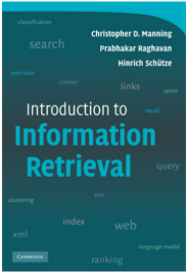


SPEECH AND LANGUAGE PROCESSING

An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition

Second Edition

DANIEL JURAFSKY & JAMES H. MARTIN



Christopher D. Manning
Prabhakar Raghavan
Hinrich Schütze

Introduction to Information Retrieval

classification
search
precision
crawler
links
spam
recall
query
clustering
index
web
xml
language model
ranking
CAMBRIDGE

Obligatory assignments

- Two obligatory exercises, each in two parts; four submissions:
- 1a+b and 2a+b.
- Possible to earn maximum of 10 points for each submission.
- In order to pass and qualify for the exam you need to collect at least 60% of the points across all exercises, i.e. 12 points across a+b.
- Extensions can only be given in case of illness, and re-submissions will not be possible.
- Group work encouraged! (Max. 3 pers.)
- See course page for the schedule:

  `https://www.uio.no/studier/emner/matnat/ifi/IN2110/v22/obliger/`
- Final exam: 4 hours digital exam at Silurveien (June 1st)

Course staff

- Forelesere:
  - Pierre Lison
  - Jan Tore Lønning
  - Lilja Øvrelid
  - Erik Velldal
- Gruppelærere:
  - Annika Willoch Olstad
  - Lilja Charlotte Storset
- Rettere:
  - Fredrik Aas Andreassen
  - Alexandra Wittemann
- Obliger / programmeringsomgivelse:
  - Egil Rønningstad

Course communication

- Questions?
  - Will use GitHub issues as dicussion board; `https://github.uio.no/IN2110/v22`
  - `in2110-hjelp@ifi.uio.no` reaches all course staff.
- Messages:
  - Check the course pages regularly.
  - Make sure to click ‘watch’ in the GitHub repo.
  - ‘Obligs’ etc will be distributed through the repo.
  - Participate in the discussion board (‘issues’).

Language technology in Python

- Will be using Python for labs and assignments.
- First lab sessions:
  - Group 1 (physical): Wed. 26th Jan. 14:15–16:00
  - Group 2 (digital): Thu. 27th Jan. 12:15–14:00

Some of the tools we'll be using in the labs

NumPy

matplotlib

scikit learn

jupyter

spaCy

- The Python add-ons ecosystem is vast (and can be confusing to navigate);
- NumPy for efficient multi-dimensional arrays and linear algebra;

Some of the tools we'll be using in the labs

NumPy

matplotlib

scikit learn

jupyter

spaCy

- The Python add-ons ecosystem is vast (and can be confusing to navigate);
- NumPy for efficient multi-dimensional arrays and linear algebra;
- SciKit-Learn for machine learning (and data preparation);

Some of the tools we'll be using in the labs

NumPy

matplotlib

scikit learn

jupyter

spaCy

- The Python add-ons ecosystem is vast (and can be confusing to navigate);
- NumPy for efficient multi-dimensional arrays and linear algebra;
- SciKit-Learn for machine learning (and data preparation);
- MatPlotLib for visualization and data analysis;

Some of the tools we'll be using in the labs

NumPy

matplotlib

scikit learn

jupyter

spaCy

- The Python add-ons ecosystem is vast (and can be confusing to navigate);
- NumPy for efficient multi-dimensional arrays and linear algebra;
- SciKit-Learn for machine learning (and data preparation);
- MatPlotLib for visualization and data analysis;
- JuPyter as an integrated development environment and authoring tool;

Some of the tools we'll be using in the labs

- The Python add-ons ecosystem is vast (and can be confusing to navigate);
- NumPy for efficient multi-dimensional arrays and linear algebra;
- SciKit-Learn for machine learning (and data preparation);
- MatPlotLib for visualization and data analysis;
- JuPyter as an integrated development environment and authoring tool;
- NLTK and spaCy for text pre-processing (from tokenization to parsing).

Next week

- Vector space models.
- Geometric framework for representing data and measuring similarity.