732A92/TDDE16 Text Mining (2022)

Text clustering and topic modelling

Marco Kuhlmann

Department of Computer and Information Science

LINKÖPING UNIVERSITY

This session

- Questions and answers
- Assessment criteria for the project
- Introduction to the lab

Questions and answers

Overview of text clustering and topic modelling

1. Introduction to text clustering
2. Similarity measures
3. An overview of hard clustering methods
4. Evaluation of hard clustering
5. Soft clustering: topic models

Questions

- What is a corpus, and what is it used for?
- Why is cosine distance not a proper distance metric?
- How to choose hyperparameters for clustering?
- Can we use clustered text as features for classifiers?
- What is a 'generative model'?
- What do the parameters η and α control in LDA?

Project

Project structure

1. Identify your problem	8 hours (w44–w48)
2. Design your approach	32 hours (w49–w50)
3. Evaluate your approach	32 hours (w51–w01)
4. Produce your report	16 hours (w02)

Assessment criteria – Problem at large

Is it clear what was done in the project, and why?
Does the project go beyond what has been covered in the course?
Does the project have enough substance, or would there have been room for more work?

- F – The report does not contain a clear and well-motivated problem statement.
  The project is essentially a repetition of one of the lab assignments.
  For a project with this timeframe1, I (the assessor) would have expected significantly more work.

- E – The report contains a clear problem statement, and the stated problem is well- motivated.
  The project goes significantly beyond the lab assignments.
  The project represents an appropriate amount of work.

- A – The problem is placed in a scientific context, including references.
  The project builds on technical content outside of the course, taken from scientific articles.
  The project represents significantly more work than expected.

Two projects

- **Project 1**

  Problem statement: ‘I want to build a system that predicts the sentiment of Amazon reviews.’

- **Project 2**

  Problem statement: ‘I want to use text mining to discover the main topics in the Holy Bible.’

Assessment criteria – Method

Is the data used in the project suitable for the stated problem?
Are technical concepts, models and algorithms applied correctly?
Are the experimental results validated with appropriate evaluation methods?

- F – The problem should have been approached differently.
  The choice of the data, models, algorithms or evaluation methods is not appropriate, or there is too little information in the report to assess whether the choice was appropriate.

- E – The data used in the project is suitable for the stated problem.
  Technical concepts, models and algorithms are applied correctly.
  The experimental results are validated with appropriate evaluation methods.

- A – The data is created specifically for the project.
  The project involves non-trivial modifications or combinations of models and algorithms.
  The experimental results are validated using several complementary evaluation methods.

Discussion (part 1)

- Give examples of data that you consider suitable/unsuitable to solve the stated problem.
- Give examples of correct/incorrect ways to apply technical concepts, models and algorithms.
- Give examples of evaluation methods that you consider appropriate/inappropriate to analyse the experimental results.

Project 1

The ability to extract meaningful insights from text data has become very important nowadays, with more sophisticated algorithms and computational power we can use huge amounts of data to build powerful models in order to use them for any related problem consisting of text data. Through this project we are going to build a sentiment classifier using some of the most recent state of the art algorithms such as XGBoost and LSTMs that are currently widely used to perform similar tasks using the open source framework Tensorflow/Keras, XGBoost and Scikit-learn.

Project 2

Jesus is the central figure of the Holy Bible and Christian faith. This study used text mining and natural language processing techniques for topic modelling to highlight the main message of Jesus in the Holy Bible of Christians. This study used theory from the Structural Topic Model, Latent Dirichlet Allocation and Zipf's law. The study concluded that the topics that highlight the main message of Jesus are *love*, *father*, *son* and *god* and that *love* was overwhelmingly the most important topic. This study also concluded by showing that the main topics remain consistent for most of the language translations of the Christian Holy Bible.

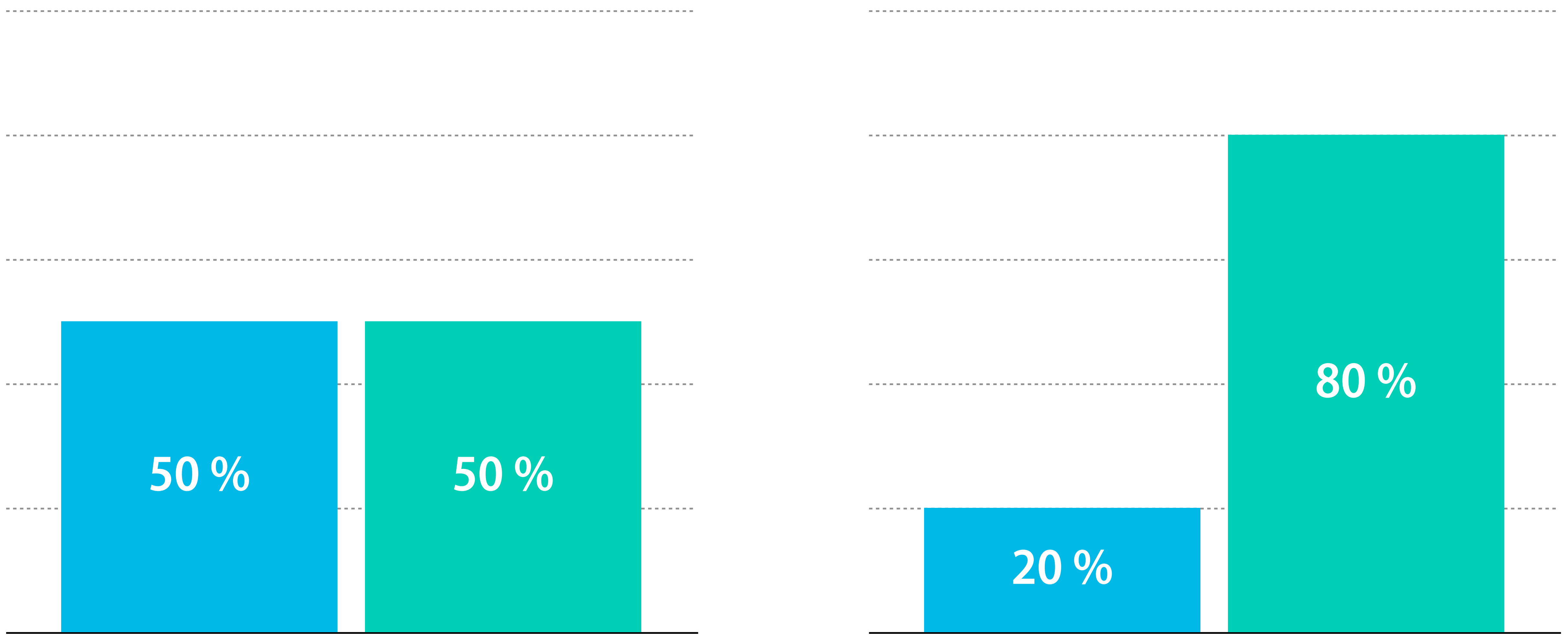
Discussion (part 2)

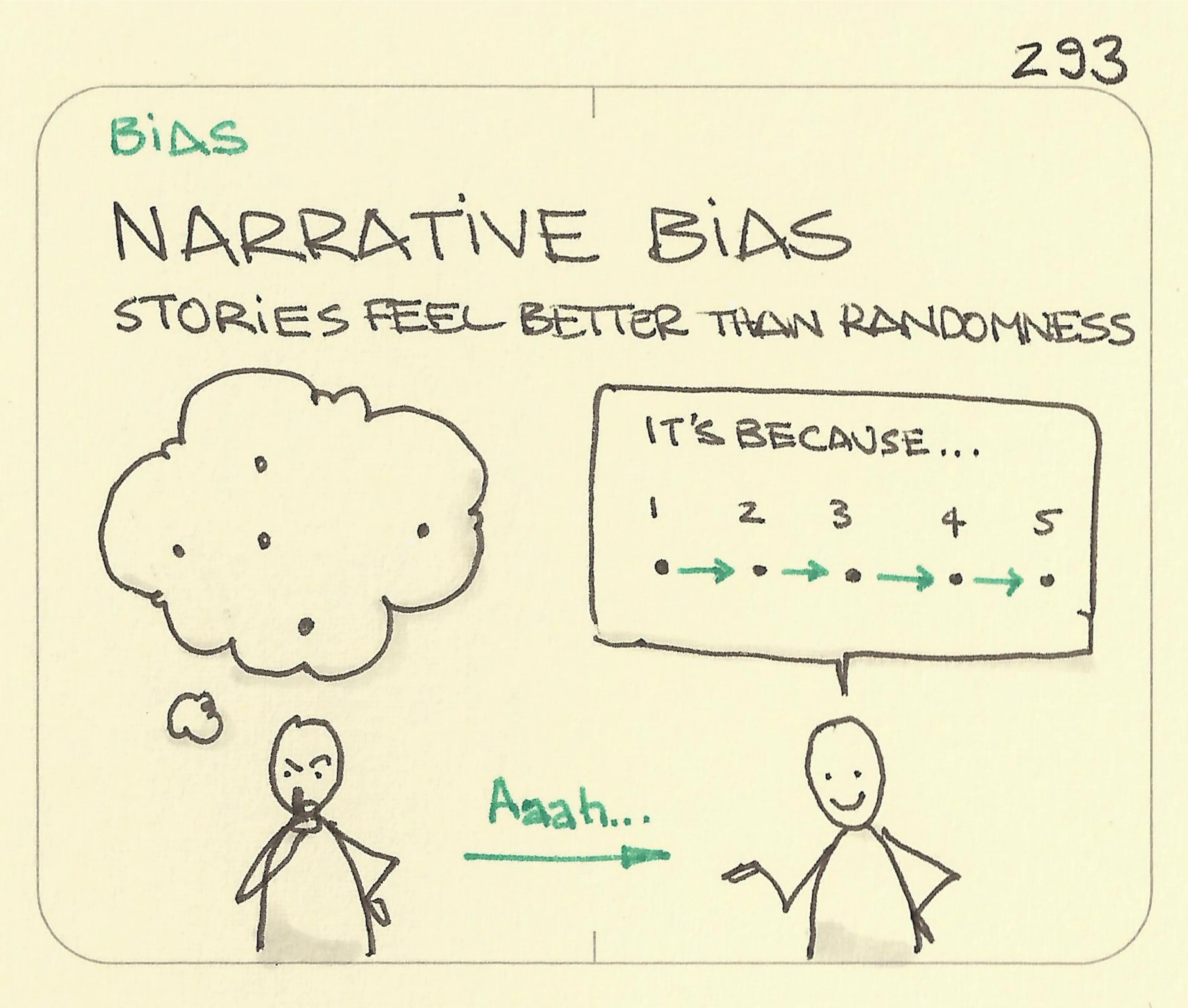
Based on the abstract:

- Do you consider the data that was used suitable/unsuitable to solve the stated problem?
- Do you consider the application of technical concepts, models, and algorithms correct/incorrect?
- Do you consider the evaluation methods that were used appropriate/inappropriate to analyse the experimental results?

Accuracy and imbalanced data sets

Is 80% accuracy good or bad?





Source: sketchplanations.com (licence)

Introduction to the lab