Exercise 6 - Topic Modeling

Discover and match topics

Deadlines

The deadline for Exercise 6 is **14.12.2020, 00:15 CET**

The deadline for the peer review is **21.12.2020, 00:00 CET.** You will find instructions for the peer review process at the end of this document.

The deadline for feedback to your peer reviewers is **30.12.2020, 00:00 CET**

Learning goals

This exercise is based on the idea of topic modeling and one of its methods called Latent Dirichlet Allocation (LDA). By completing this exercise you should …

- … understand how topic modeling is used as a text-mining tool.
- … be able to apply latent dirichlet allocation for statistical topic modeling.
- … deepen your understanding of the role of identifying abstract topics or matching topics in a document

Please keep in mind that you can always consult and use the exercise forum if you get stuck (note that we have a separate forum for the exercises).

Deliverables

We encourage you to hand in your solutions as a Colab-Notebook. **Download your notebook as a .ipynb file**. That way your reviewers can view and execute your code. Or can view your already executed code.

Please hand in your code and your lab report. Hand in the following files and name them exactly in the following fashion:

-ex06_lda.ipynb

-ex06_labreport.pdf

zip it and name the zip-folder *ex06_ml4nlp1.zip*.

The .ipynb files contain your well documented AND EXECUTABLE code. We recommend you use Google's Colaboratory, where you have access to GPU time.

If you prefer to solve the exercise on your own computer please submit a zip-folder containing the required files:

- ex06_lda.ipynb
- ex06_labreport.pdf

Please submit the lab report in PDF format. The lab report should contain a detailed description of the approaches you have used to solve this exercise. Please also include results. In this exercise description, we highlight places in green where we expect a statement about an issue in your lab report.

Please note:

- Your peers need to be able to run your code. If it does not work, you will not be able to obtain the maximum number of points.
- DO NOT submit the data files!

Data

For this exercise, you will work with a dataset of descriptions of companies (mostly startups). Hence, you have to build a model that can find similarities between companies based on their descriptions.

The dataset can be downloaded via the following link:

- https://drive.google.com/file/d/1r3BN84TzzOz51DLhoUKyVES8OVCELAXZ/view?usp=sharing

Given

For the exercise, you are given this notebook. It already contains code for downloading, preprocessing, inspecting the dataset, and some code for Part 1 and 2 to use as described in the next sections.

Part 1 - Matching companies with TF-IDF (Vector Space Model)

Analyze the code in the TF-IDF section of the notebook (cells 10-14). What does the code in those cells do?

Write 1-2 sentences per cell to explain what is done (e.g. explaining the contents of the variables) and why.

Extend the code in this section to get the top 5 most similar companies to the company "Much Asphalt".
Which are the most similar companies? Do the results make sense?

Part 2 - Matching companies with Latent Dirichlet Allocation (LDA)

Implement a model based on LDA to match companies on the basis of the description and build a model to relate the companies semantically.

You should use an LDA model from sklearn or gensim.

1. Find the top 5 closest matches (companies) to the companies with names:
    - "Much Asphalt"
    - "Vahanalytics"

Additional help: If you want to, you can use the functions and steps provided under the section "Topic Modeling Using LDA" in the colab-notebook as a guide for this part of the exercise.

2. Which method produces more sensible output? Discuss. (Identify manually on the basis of the description of companies)

Peer Review Instructions

First: go to www.eduflow.com/join and join the class with the code **42D2XJ**. Important: Register with the E-mail address you use for OLAT.
As soon as the deadline for handing in the exercise expires you will have time to review the submissions of your peers. You need to do **2 reviews** in order to get the maximum number of points for this exercise.
Here some more rules:

- If you do not submit 2 reviews, the maximum number of points you can achieve is 0.75 (from a total of 1).
- Please use full sentences when giving feedback.
- Be critical, helpful, and fair!
- **All reviews are anonymous: Do not put your name into the python scripts, the lab report or the file names.**
- You must also give your reviewers feedback. The same criteria as above apply.
- Students that consistently provide very helpful feedback can be awarded a bonus in case they earned less than 6 points in total. Ways to obtain points are thus the following:
    - 5 exercises = 5 points
    - 1 presentation or research paper dissection = 1 points
    - consistently good reviews = 1 point

Groups:

- You can create groups of two to solve the exercise together.
- Both students should submit the solutions separately and the same lab report can be submitted. Otherwise, points of team-mates might differ based on feedback on the lab report.
- When submitting the exercise, write a small post in the "Groups"-thread in the exercise forum on OLAT to notify the instructors about the group. **Please notify first hand if you change the group/ decide to work alone.**
- As a group member, you still have to review two submissions with your own eduflow account. However, you may work together in the group to write all 4 reviews.