Prof. Björn Gambäck
TA. Tollef Jørgensen

Lab 5+6

Information access, deep learning and transfer learning

Deadline: 23:59, April 13th, 2022

Guidelines

For this final lab, there will be no boilerplate code provided. By now you should have well structured and easily reusable code from previous labs, which should be used where appropriate. This includes file loading, text cleaning, tagging, etc. Deliver in your desired format, whether this is a set of '.py' files or a single '.ipynb' notebook. Useful info can be found in the GitHub[1]) repository regarding general usage of the libraries and hints/tips. Note that this code is mostly considered a guideline, and is not necessarily the correct approach. Keep in mind that code quality is important. We expect you to use object oriented functionality and create wrappers where necessary.

Provide your outputs in a simple report, along with textual answers (remember to justify your choices and note down any observations or issues). Zip if there's more than 1 file.

Format: first_last_lab6.zip/ipynb.

Exercise 1 - Named entity recognition with Harry Potter

You will find the text for 'Chamber of Secrets' on GitHub[2], which you are to use for all tasks in this exercise. Load the file and separate chapters into a proper data structure. Each chapter should be a single string of text. For visualization purposes, use a few sentences from chapter 1, or whatever subset which makes sense.

1. With your knowledge of POS tags and chunking, attempt to fetch entities from text. Print out the NLTK trees.[3]
2. Chunk pronouns in the same text, and reflect on how you would attack the problem of attributing a pronoun to an entity.
3. Use NLTKs built-in functionality for NER. Print the top 10 most frequent entities in chapter 2.
4. Now implement spaCy. See code on GitHub. Perform the previous task and discuss the results.
5. Using the results from spaCy, plot the frequency of characters (e.g. PERSONs) in the entire book.
6. Visualize the dependency tree using spaCy, discuss how you could utilize the results to improve upon what you figured out in task 2.

Exercise 2 - Link that entity!

Continue using the data from Exercise 1. With HTTP-calls ("requests" library) in Python, build functionality to fetch information from entities using the WikiData knowledge base. This can be simplified by using the "qwikidata" library found here: `https://github.com/kensho-technologies/qwikidata`. This task is fairly open, but requires you to fetch either relationships (e.g. father, mother) or other info such as aliases from entities in the text.

[1] `https://github.com/ph10m/TDT4310-Exercises`

[2] `https://github.com/ph10m/TDT4310-Exercises/tree/main/lab5-6/data`

[3] Hint: in Jupyter Notebooks, you can use the 'IPython.display' module to print more than one tree in loops

Prof. Björn Gambäck
TA. Tollef Jørgensen

Exercise 3 - Sentiment Analysis with Keras

Implement a neural network with LSTM with dropout. Download a commonly used dataset for sentiment analysis: IMDB Reviews `https://www.tensorflow.org/datasets/catalog/imdb_reviews`. This is found in the "tensorflow-datasets" package, installable with pip, or via keras `https://keras.io/api/datasets/imdb/`.

1. Instantiate the dataset and clean it as you see fit. Encode the labels as you want (e.g. 0 for negative).
2. Setup a shallow feed-forward NN to give your setup an initial benchmark (no LSTM)
3. Setup a NN with LSTM. Feel free to follow the code from the ATAP book. If you wish, you can implement GloVe embeddings in the initial layer.
4. Test out a few different model setups, different dropouts, etc. This is heavily based on your available hardware.
5. Verify the model on a sample of texts from Chamber of Secrets. Explain your initial thoughts and reflect on how you could create a dataset more suited to the domain of fantasy books.

Exercise 4 - GPT-2 fine-tuning and generation

Using the Hugging Face library *transformers*[4], fine-tune a generalized GPT2-model to generate sentences based on all chapters of Chamber of Secrets except chapter 1. Then generate sentences based on the first few words in the original sentences of chapter 1. Explain your results.

There are *a lot* of resources on the topic here. Use this an opportunity to learn something cool and avoid copy-pasting directly.

Exercise 5 - Preparing for your project with Transformers

By now you have probably decided on your topic for the main project. Use this task as an introductory part, whether that is to explore entity recognition, dependency parsing, sentiment analysis, summarization, or anything else. Explore the library and experiment with already fine-tuned models related to your problem. See e.g. `https://huggingface.co/models?pipeline_tag=text-generation&sort=downloads`.

Note down your results here and what you may have found :-)

Good luck with your projects!

[4] `https://huggingface.co/docs/transformers/index`