4️⃣ Natural Language (NLP)

Learning steps

- ☑️ ~~yt video~~
- ☑️ ~~notebook/own implementation~~
- ☑️ ~~book chapter~~

Resources

- website 💻
    - lesson 4
- notebooks 🗒️
    - Getting started with NLP for absolute beginners
- book 📘
    - book chapter 10
    - solutions to exercises

Quick notes

YouTube video (→ link)

- using Huggingface
    - we will do it in a completely different way than in the book (RNNs vs Transformers)
        - not even gonna use FastAI at all
            - good to use the same concepts in different ways
        - fast-tuning a pre-trained Huggingface Transformers
            - state-of-the-art model in NLP
            - lower level library ⇒ we have to go deeper to make it work
- fine-tuning pre-trained models

    - similar to having a lot of parameters already set to optimal values ⇒ fine-tuning is the process of finding values for the other parameters and barely touching the ones with optimal values
- UMLFit
    - turned into an academic paper
    - goals of the model
        - language model trained on all of Wikipedia
            - goal: predict the next word of a Wikipedia article
            - the model needed to be good at a lot of things
                - language, syntax, etc
        - language model IMDb
            - goal: predict the next word of an IMDb review
        - classifier IMDb
            - goal: fine-tune these weights to do a sentiment analysis of IMDb reviews
            - used RNNs
- transformers were invented during the release of the UMLFit project
    - take good advantage of Google's TPUs
    - didn't allow to predict next word's in sentences
        - instead, use chunks of Wikipedia, delete words, predict the deleted words (quite similar)
- Zeiler & Fergus
    - see features detected at layer of a computer vision model
- US patent Kaggle competition
    - real competitions vs playground
        - real ones are actual projects that organizations put money on solving
    - steps taken
        - see notebook for more details
            - transform problem into a classification problem by creating an input column
            - use tokenization to create numerical representations of tokens (words or pieces of words)
- four key libs for data science on Python
    - numpy, matplotlib, pandas, pytorch
    - good book about these: Python for Data Analysis (Wes McKinney, free edition on his website)
- things to pay attention to when creating an AI model
    - be wary of incorrect metrics (+overfitting, use test_train_split)

        - make sure to use pictures and understand what is happening when the metric is getting better
    - alphas in scatter plots allow to have darker areas where there are more individuals
    - be careful of outliers

Exercises