Authorship identification using deep learning

Füleki Fábián, Jani Balázs Gábor, Torner Márton
Project work for BME Deep Learning course (VITMAV45),
Team: LoremIpsum

Goal:

We are aiming to indentify the most likely author of an article, from group of authors, whose previous works are known and processed.

Szerző azonosítás deep learning használatával

Füleki Fábián, Jani Balázs Gábor, Torner Márton
Projektmunka a BME Deep Learning kurzusára (VITMAV45),
Csapat: LoremIpsum

Cél:
A célunk azonosítani a legvalószínűbb szerzőt, lehetséges szerzők egy olyan csoportjából, kiknek a korábbi munkája ismert és általunk feldolgozott.

Introduction:

Authorship identification is one of the major problems of written media and forensic science. Identifying someone as the author of an article or even just a few sentences can be harmful and helpful as well. Our goal here is to provide a reliable author identifier tool of printed text. Using handwriting would be even more accurate, but we would like to create a general solution for the problem. Modern media is based on internet articles anyway. The author identification could be easily done by using Stylometry techniques, most of the authors have personal habits, commonly used phrases, preferred locutions. We attempt to create a neural network to automate this identification process. Our solution is based on a Bidirectional Long Short Term Memory network,

Subject, previous solutions:

Dataset:

Our primary dataset is the Reuters_50_50 (C50), which is a subset of Reuters Corpus Volume I(RCVI). The RCV1 is archive of categorized newswire stories, made public for research purposes by Reuters, Ltd. The C50 collection consist of 50 texts for each of the 50 top author, for training and separately the same amount for testing purpose (5000 texts in total). This dataset has been previous used by previous studies of authorship recognition and can be found here: *https://archive.ics.uci.edu/ml/machine-learning-databases/00217/C50.zip*

Preparing dataset:

Our goal here is to format the sentences into data points, which can be used as input for the neural network. We are using Spacy for representing each word as a series of numbers. Spacy can determine for each word which part of the sentence it is. This is called part of speech and can be very useful. If a word cannot be processed by Spacy because it isn't in the language pack, we don't include it in the processed sentence. Stop words and punctuations are not really useful in our scenario, so we are removing them as well.

We don't want to lose any information about coherence between sentences, so every input data will consist of three sequential sentences. Therefore every sentence will be included three times in the dataset, this is some kind of data augmentation as well.

Neural network:

Approaches:?

Learning histories…

Final evaluation:

Conclusions, Future plans:

References: