University *of Ljubljana*
Faculty *of Computer and Information Science*

Submission 1 - Project selection and simple corpus analysis

Martin Jurkovič, Blaž Pridgar, and Jure Savnik

Abstract

Keywords
Public domain book corpus, GPT, Fine tuning, Slovenian literature

Advisors: Slavko Žitnik

Introduction

With the public demo of ChatGPT the field of natural language processing (NLP) has seen the most attention from the public eye any machine learning (ML) field has ever seen. With insurmountable amount of written language, be it books, articles, blogs, tweets, may of the sources can't be publicly used in the ML domain because of copyright issues. That's why we decided to focus on books which are in public domain, meaning they are not being subject to copyright or other legal restrictions. One of the most popular public domain book repositories is Project Gutenberg [1] (*www.gutenberg.org*), which holds only one Slovenian book. That is why we decided to create a public domain book corpus of publicly available books in Slovenian language. We gathered the books from the website *www.dlib.si* of the *Digital Library of Slovenia* along with their metadata.

The second step of our project is to use this corpus to fine tune a Slovenian generative text model [2]. This model is trained on corpuses Gigafida, KAS, slWaC, and MaCoCu and we would like to fine tune it with additional literature found on *dlib.si* and *Wikivir - sl.wikisource.org*.

Related work

Public domain book corpus

Gerlach, M. and Font-Clos, F. built and processed a *Standardized Project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics* [1]. They processed the books by taking raw text, extracting only text, tokenizing this text with a tokenizer from the *nltk* library and produced word counts. Besides the title and author of the book they have also provided a broad characterization of the books in terms of their length, language, date of publication and subject. They have also quantifies variability in the corpus based on subject and time of publication.

Another notable paper is *Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books* by Zhu, Y., Kiros, R. [3]. They created a corpus of 11.038 books from the web and provided six summary statistics. The researches doenloaded every free book longer than 20.000 words from the website *www.smashwords.com*, whch contains ebooks by independent authors. They have also aligned the books with corresponding movie releases and mostly focus on that topic, with almost no mention of corpus preprocessing or summarization.

Generative text modelling

A group of for researches from Huawei Noah's Ark Lab wrote a paper on fine tuning a generative text model for generating Chinese poetry [4]. This is a similar task to what we aspire to achieve with Slovene literature. In their example they fine tuned a model trained on a corpus of Chinese news with 235M sentences with an additional dataset of around a milion Chinese poems.

Another relevant paper written by Jieh-Sheng and LeeJieh Hsiang describes how they fine tuned OpenAIs GPT-2 with a corpus of more than 500.000 patent claims to develop a pre-trained model for generating patent claims [5].

Methods

Results

Discussion

Acknowledgments

References

[1] Martin Gerlach and Francesc Font-Clos. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *CoRR*, abs/1812.08092, 2018.

[2] cjvt/gpt-sl-base · hugging face.

[3] Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *CoRR*, abs/1506.06724, 2015.

[4] Yi Liao, Yasheng Wang, Qun Liu, and Xin Jiang. Gpt-based generation for classical chinese poetry, 2019.

[5] Jieh-Sheng Lee and Jieh Hsiang. Patent claim generation by fine-tuning openai gpt-2. *World Patent Information*, 62:101983, 2020.