University of Ljubljana
Faculty of Computer and
Information Science

FRI Natural language processing course 2024
Project report

Slovenian Instruction-based Corpus Generation

Jernej Ulčakar, Jon Kuhar, Bella Muradian

Abstract

This study explores the utilization of web scraping techniques to extract and analyze data from online forums. Through scraping scripts, we gathered a substantial corpus of forum posts and comments. This corpus was then parsed with scripts.

Keywords
LLM,Slovene,Training

Advisors: Slavko Žitnik

Introduction

Large Language Models (LLMs) have shown great promise as highly capable AI assistants that excel in complex reasoning tasks requiring expert knowledge across a wide range of fields, including in specialized domains such as programming and creative writing. They enable interaction with humans through intuitive chat interfaces, which has led to rapid and widespread adoption among the general public. In this project we use knowledge from research on InstructGPT to create our multilingual LLM fine-tune. This method permits using very low parameter counts compared to the full model. In the case of InstructGPT the original model had 175B parameters and the fine-tune had 1.3B parameters. We use the same methods to construct and train a fine-tune for a multi-lingual that supports slovene.

We trained the model on conversations from slovene forums such as Med.Over.Net, slo-tech and Reddit Slovenia. We collected this data using WinHTTracker scraper and python for parsing.

Related Works

Llama 2 (2023) [1] is a collection of pretrained and fine-tuned large language models (LLMs) ranging in scale from 7 billion to 70 billion parameters. They introduced LLMs that are fine-tuned, called Llama 2-Chat, and are optimized for dialogue use cases.

BLOOM (2023) [1] is a collection of pretrained and fine-tuned A 176B-parameter open-access language model designed and built thanks to a collaboration of hundreds of researchers. BLOOM is a decoder-only Transformer language model that was trained on the ROOTS corpus [2], a dataset comprising hundreds of sources in 46 natural and 13 programming languages (59 in total). BLOOM achieves competitive performance on a wide variety of benchmarks, with stronger results after undergoing multitask prompted fine-tuning. To facilitate future research and applications using LLMs.

Training models (2022) [3] this paper contains instructions in how to fine-tune language models with user intent on a wide range of tasks by using human feedback. Starting with a set of labeler-written prompts and prompts submitted through the OpenAI API, they collected a dataset of labeler demonstrations of the desired model behavior, which they later used to fine-tune GPT-3 using supervised learning and the collect the dataset of rankings of model outputs. These are used to further fine-tune the model using reinforcement learning from human feedback. The model names are InstructGPT. In human evaluations on our prompt distribution, outputs from the 1.3B parameter InstructGPT model are preferred to outputs from the 175B GPT-3, despite having 100x fewer parameters. Moreover, InstructGPT models show improvements in truthfulness and reductions in toxic output generation while having minimal performance regressions on public NLP datasets. Even though InstructGPT still makes simple mistakes, their results show that fine-tuning with human feedback is a promising direction for aligning language models with human intent.

Methods

We employed WinHTTrack, a powerful web scraping tool, to scrape data from three distinct online forums: Med.Over.Net, slo-tech.com and forum.over.net.

- **WinHTTrack Configuration**: We configured WinHTTrack to navigate through the forum websites systematically, simulating user interactions to access all accessible content, including posts and comments.

- **Customized Scraping Scripts**: To ensure efficient parsing of the forums' diverse structures and functionalities, we developed customized scraping scripts tailored to each website's specific layout and navigation protocols. These scripts enabled targeted extraction of forum data while maintaining compliance with the websites' terms of service.
- **Data Extraction**: Using WinHTTrack, we initiated the scraping process, allowing the tool to systematically crawl through the target forums and retrieve HTML content representing forum pages, threads, and posts.
- **Data Preprocessing**: Upon completion of the scraping process, we conducted data preprocessing steps to clean and structure the extracted HTML content. This involved parsing the HTML files to extract relevant text data, removing noise, and organizing the data into a structured format suitable for further analysis.
- **Quality Assurance**: To ensure data integrity and accuracy, we performed regular checks during the scraping process, monitoring for any anomalies or errors in the extracted data. Additionally, we implemented safeguards to prevent overloading the forum servers and to maintain ethical data scraping practices.

Through these methods, we successfully parsed data from Med.Over.Net, slo-tech.com/forum, and forum.over.net, enabling subsequent analysis of forum content to uncover valuable insights into user behavior, discussion trends, and community dynamics.

Results

Parsing the data was very well done. The only junk remaining was inside the posted content of each user as well as the users type of speech (this includes typos and accent-like typing).

More random text

This text is inserted only to make this template look more like a proper report. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam blandit dictum facilisis. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Interdum et malesuada fames ac ante ipsum primis in faucibus. Etiam convallis tellus velit, quis ornare ipsum aliquam id. Maecenas tempus mauris sit amet libero elementum eleifend. Nulla nunc orci, consectetur non consequat ac, consequat non nisl. Aenean vitae dui nec ex fringilla malesuada. Proin elit libero, faucibus eget neque quis, condimentum laoreet urna. Etiam at nunc quis felis pulvinar dignissim. Phasellus turpis turpis, vestibulum eget imperdiet in, molestie eget neque. Curabitur quis ante sed nunc varius dictum non quis nisl. Donec nec lobortis velit. Ut cursus, libero efficitur dictum imperdiet, odio mi fermentum dui, id vulputate metus velit sit amet risus. Nulla vel volutpat elit. Mauris ex erat, pulvinar ac accumsan sit amet, ultrices sit amet turpis.

Phasellus in ligula nunc. Vivamus sem lorem, malesuada sed pretium quis, varius convallis lectus. Quisque in risus nec lectus lobortis gravida non a sem. Quisque et vestibulum sem, vel mollis dolor. Nullam ante ex, scelerisque ac efficitur vel, rhoncus quis lectus. Pellentesque scelerisque efficitur purus in faucibus. Maecenas vestibulum vulputate nisl sed vestibulum. Nullam varius turpis in hendrerit posuere.

Nulla rhoncus tortor eget ipsum commodo lacinia sit amet eu urna. Cras maximus leo mauris, ac congue eros sollicitudin ac. Integer vel erat varius, scelerisque orci eu, tristique purus. Proin id leo quis ante pharetra suscipit et non magna. Morbi in volutpat erat. Vivamus sit amet libero eu lacus pulvinar pharetra sed at felis. Vivamus non nibh a orci viverra rhoncus sit amet ullamcorper sem. Ut nec tempor dui. Aliquam convallis vitae nisi ac volutpat. Nam accumsan, erat eget faucibus commodo, ligula dui cursus nisi, at laoreet odio augue id eros. Curabitur quis tellus eget nunc ornare auctor.

Discussion

Use the Discussion section to objectively evaluate your work, do not just put praise on everything you did, be critical and exposes flaws and weaknesses of your solution. You can also explain what you would do differently if you would be able to start again and what upgrades could be done on the project in the future.

References

[1] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[2] Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González

Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gerard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Alexandra Luccioni, and Yacine Jernite. The bigscience roots corpus: A 1.6tb composite multilingual dataset, 2023.

[3] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.