GlotSparse: Creating Corpora in Under-Resourced Languages

- **Title:** GlotSparse: Creating Corpora in Under-Resourced Languages
- **Supervisor:** Amir Hossein Kargaran
- **Examiner:** Prof. Hinrich Schütze
- **Open for: MSc/BSc**

Introduction

The goal of this project is to find and create monolingual datasets for very under-resourced languages. These languages are probably not among the 300 languages in Wikipedia. There are more than 7k languages around the world, and if we want to have at least 30K sentences for each of these languages, at best, we can achieve this for 500 languages that meet the criteria. Getting 30k sentences for a language should not be a hard task, as even one local news website can provide 30k sentences.

In this project, we first start by running language identification and script identification around the web or Common Crawl to find possible cases where a website is gathering data for a low-resource language. Some examples are included in the GitHub repository and how they are achieved. This GitHub repository is the start of this project: https://github.com/cisnlp/GlotSparse and this is the data gathered till now: https://huggingface.co/datasets/cis-lmu/GlotSparse

Objectives

1. Collect a list of possible websites for each language, except for religious data.
2. Find a good pipeline and modify or write one to crawl the content of a website.
3. Respect the website's terms of service and robots.txt file to avoid any legal issues during the crawling process.
4. Crawl and save the data in the same format as https://huggingface.co/datasets/cis-lmu/GlotSparse.
5. Perform some preliminary analysis on the text, such as Zipf's law (showcases are available in Ahmadi et al.), or calculate similarity, such as perplexity divergence (showcases are available in Imani et al.), of gathered texts with closely related languages in Bible text.

Prerequisites

- Enthusiasm (for publishing results at a conference/workshop)
- Proficiency in speaking and writing English
- Good Python programming background (e.g., knowledge of NumPy and Pandas, Matplotlib libraries, and how to write classes in Python)
- Basic knowledge of NLP and data cleaning

Supervisor

Hello, I am Amir. If you choose this project, I will be your supervisor. You will receive 1 hour per three weeks of guidance and help. Our work begins with a comprehensive literature review of the task and available resources. We want to focus on languages missed by the community. I will ensure you receive the allocated time from my side. For any questions, contact me directly: amir@cis.lmu.de

References

- [Ahmadi et al.] Approaches to Corpus Creation for Low-Resource Language Technology: the Case of Southern Kurdish and Laki
- [Imani et al.] Glot500: Scaling Multilingual Corpora and Language Models to 500 Languages