2025-11-04 Call with Yali AI

Attendents: Jan, Philbert, Schadrack

Jan: I am still crawling, currently 35 mio words. Next tasks: 1) Add Kinyarwanda Wikipedia and the corpus of Mike which is on huggingface. We need compute and programmers.
Jan: We cannot offer a funded project for now, we need to work on a win win basis
Philbert: Yes that is ok.
Philbert: We can offer compute and programmers. We can also look into the Kinyarwanda Wikipedia. We should also look into data quality / data cleaning.

Tasks

1) Data quality assessment of the three sources (Mikes, mine, Wikipedia)
2) Improving data quality
3) Merging the data

Schadrack: We want to build our own tool for data scraping. We would like to use your tool. We shared in the team 10 sources for data that we can get.
Jan: If I share the crawler it would be important to make these works as open source and if your work contributed back to the crawler.
Schadrack: Yes I agree
Philbert: I agree also. Also, the Rwandan government (MINICT) is working on a data generation project for Kinyarwanda also (text, image, …).
Philbert: Also, how can we add data that is not online? E.g., from the Rwandan Institute for Cultural Heritage.
Jan: You should speak to Arnaud and learn about his OAI - PMH crawler. He wants to crawl digital libraries for Kinyarwanda data.
Jan: What is your strategy towards sharing the training data / publishing the data as open source.
Schadrack: I think we can share data publicly. But, the end goal would be sharing the dataset with the public if it was useful.
Jan: Open source is important for me because my job is to promote open source. The data I generate should be open source. It would help me a lot if the data that we generate together would be open source.
Jan: Are other languages apart from Kinyarwanda interesting for you as well?
Schadrack: This is important but we want to get started with Kinyarwanda.

Jan: Who would contribute what?

Jan:

- Contribute the data that I already generated (and which will be created in future)
- The crawler itself
- My general expertise

Yali:

- First step: Data quality assessment
- Compute resources

Schadrack: Which compute resource do you need?
Jan: Right now I use 4 CPU Cores, 32 GB RAM and a large HDD (~200 GB). It also works with only 16 GB RAM but a bit slower. This runs per language.
Jan: It might be interesting to add a 2nd language.
Philbert: We will discuss what we can contribute.
Jan: We could aim at a publication on LREC2026. Deadline is in october. Would this be an interesting aim for YALI?
Philbert: We see ourselves also as researchers. A publication would be nice. We could also provide additional guides on the data cleaning process.

Next steps:

1) Jan sends the data that to Yali
    a) Crawled data
    b) Wikipedia plain text dump
2) Yali adds the Mbaza data and works on
    a) Data quality assessment
    b) Deduplication
3) Yali to provide a shared folder / a data server.
4) Yali to check computational resources to move the crawler to Yali

Jan: Can we transfer the data via SSH or do you mean Google Drive?
Schadrack: I think we can offer SSH. We can offer a data server for data sharing.
Philbert: Can you add me to the code base of the Crawler?

Schadracks Emails

niyibizischadrack05@gmail.com

niyibizi@yalilabs.com