DH-RAG Assistant: Entry Point in Digital Humanities for Russian

Yana Khlusova

October 23, 2024

Abstract

This technical report is devoted to the description of the idea of virtual assistant for beginners in the Digital Humanities field in Russia. It describes the prerequisites of such a work, collected data, preprocessing steps and further plans.

1 Introduction

Digital Humanities — the area of scientific research that uses Digital approaches to gain new knowledge in Humanities. In Russia this field is still not so popular and well-known [**Antopolskiy et al.**, **2023**] so there are several projects aimed to solve the problem.

In July 2024 Russian project "Systemniy Blok" (System Blok, SB) [2] launched a portal dedicated to Digital Humanities "Digital Humanities: Entry Point" [3]. The main goal of the portal is to give an overview for those who are interested in and to popularize DH in Russia. The portal consists of articles from SB connected to the main topic. These articles are written on the basis of real scientific works but in the popular form, so that every person may understand and appreciate them.

In 2023 Siberian Federal University published a monograph [**Antopolskiy et al.**, **2023**] written by famous Russian DH-scientists — Boris Orekhov, Anastasiia Bonch-Osmolovskaya, Dinara Gagarina and many other. This is the very first book in Russian which shortly but consistently covers all aspects of the sphere and, hence, is a great resource for the beginners. Thus, given the above, it was decided to develop a virtual assistant with usage of Retrieval Augmented Generation [**Lewis et al.**, **2020**] The assistant is thought to receive questions and give answers in Russian language.

2 Methods

The implementation of virtual assistant involves RAG QA-system development. This implies collecting textual data for information retrieval, converting it into embeddings and storing into the format of vector-database. This is supposed to be realized with the help of LangChain framework [5].

3 Data

At this moment the knowledge base for assistant is made only of the monograph mentioned above. It has open access and exists in the form of PDF-file. The last point imposes the specifics of preprocessing steps. It included transformation of the PDF-file into a plain text format (optical character recognition, OCR) and postprocessing steps.

Several methods of OCR were tested. The first one consisted of the usage of pdf2image and pytesseract python packages. The data were converted into images using the first package and then the text was recognised using the latter option. It took about 7–8 min to recognise all 273 monograph's pages. The work with pytesseract requires specification of language of the document. For the very first probe Russian and English were chosen (because the monograph contains web-links). However, the later research highlighted small parts of French and even Chinese texts. The first one is also used in the references list, the second is an example of multilingual tokenization problems. So this step highlighted the problem of differentiation between the main text of PDF-files and text from the illustrations.

Another option consisted of using instruments provided by LangChain. This allowed the usage of PyPDF and PyMuPDF document loaders [5]. It took just several seconds for both of them (about 12 for PyPDF and 0.7 for

PyMuPDF) to cope with the monograph without specifying the language. Furthermore, the quality of recognised text was higher than with the first option. Another positive point is that these loaders split text into parts (by the number of document pages), which is probably designed to avoid text splitting step in building a knowledge base. However, there are two negative points of such a decision: firstly, splitted text still has too much characters (about 2800) which may be more than context window of LLM; secondly, the text contains special symbols as and 0, which disappear while printing, but still are the parts of the strings. Thus, it implies postprocessing steps for cleaning these symbols and the necessity of reconnecting parts of the text to split them into smaller parts.

The main difference between PyPDF and PyMuPDF document loaders is the amount of metadata provided. PyPDF makes only source and page numbers. PyMuPDF adds much more data, including keywords, file path, total pages, etc. Because of the fact that I used only one file without any special features, PyPDF was preferred.

Talking about other sources for knowledge base, there was also an option to use Wikidata for our purpose but there were several obstacles to it. First of all, there are not many Russian Wiki Pages in the Digital Humanities category [6]. It means that if one would like to use it, the system would not be able to give full and reliable answers. There is much more DH-content in English Wiki Pages [7], however, it forces me to think about translation. This is the field of future research.

4 Further steps

The next stages are partially described in the previous section: it is crucial to postprocess recognized text. This will be made with regular expressions (python re package). Meanwhile, the question of content of PDF-pages is open. Probably, the examples and illustrations could be removed using RegEx, too. However, this is the field of future research.

Then, it is necessary to choose an appropriate way of Russian text vectorization and LLM capable of generating answers in Russian. After that the QA-chain should be developed.

5 References

[1] Antopolskiy, A.B., Bonch-Osmolovskaia, A.A., Borodkin, L.I., Volodin, A.Yu., Gagarina, D.A., Grishin, E.S., Kizhner, I.A., Orekhov, B.V., Rumyantsev, M.V., Smetanin, A.V. Cifrovie Gumanitarnie Issledovaniia: monographiia [Digital Humanities: monograph]. https://bik.sfu-kras.ru/shop/publication?id=BOOK1-

[2] Systemniy Blok [System Blok]. https://sysblok.ru/

[3] Digital Humanities, ili Cifrovie Metody v Gumanitarnikh Naukakh: Tochka Vkhoda [Digital Humanities: Entry Point]. https://sysblok.ru/dh/

[4] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. 10.48550/arXiv.2005.11401.

[5] LangChain. https://www.langchain.com/

[6] Category: Cifrovie Gumanitarnie Issledovaniia [Digital Humanities]. https://ru.wikipedia.org/wiki/

[7] Category: Digital Humanities. https://en.wikipedia.org/wiki/Category:Digital$_{h}umanities$