UDC 004.912

DIGITAL HUMANITIES FIELD ANALYSIS BASED ON GOOGLE SCHOLAR DATA

Anastasiia Chernysheva
(ITMO University, Saint Petersburg)
Supervisor – Maksim Khlopotov, Ph.D.
(ITMO University, Saint Petersburg)

Introduction. Digital Humanities is a new, rapidly developing field, which is gradually becoming a subject of interest for Russian scientists and researchers. So far, this area of knowledge is believed to be represented mostly by natural language processing and data visualization. However, the full range of areas covered by or closely related to Digital Humanities, is not specified [1].

The main **goal of our research** is to study Digital Humanities as a subject area by extracting and analyzing keywords from Google Scholar scientific data.

The first step is data collection. Google Scholar does not provide any kind of API, which would allow researchers to use its resources. Therefore, data collection was carried out through Scholarly Python package for Google Scholar data retrieval [2], which has a relatively poor, but still sufficient set of features necessary for our research.

As a result of the first step, the following data was collected:

- scientific interests of the researchers who mentioned Digital Humanities as one of their subjects of interest in the Google Scholar profile;
- information on mentioned above authors' publications for the past 5 years including titles, abstracts, journals, etc.

The second step is pre-processing and analysis of the collected data. First, a list of keywords based on authors' scientific interests was compiled. Processing of the list included translating all terms into one language (English), with a subsequent manual correction of these translations, search and replacement of synonymous terms. The result list consisted of 1555 keywords. Terms with the highest occurrence rate among them are Natural Language Processing, Data Visualization, Computational Linguistics, Library Science, Information Technologies, History, Media Studies, Literature and Machine Learning.

The next step is processing and analyzing information on publications. The following work was done there:

- specifying top of journals preferred by the scientists;
- detecting what languages abstracts and, therefore, articles are written in and automatically translating non-English texts into English;
- keywords extraction by methods for automatic keyword extraction, which showed the need for further research in this area.

Thus, we obtained some **results** during the steps of our research:

- Google Scholar data analyzing methods are explored;
- list of keywords based on scientific interests of researches in the field of Digital Humanities is compiled and analyzed;
- list of keywords extracted from abstracts is compiled and analyzed.

References:
1. Digital Humanities, Web: https://ling.hse.ru/Projects_DigHum.
2. Scholarly 0.2.3, Web: https://pypi.org/project/scholarly/.

Author: Anastasiia Chernysheva

Supervisor: Maksim Khlopotov