Автоматическая обработка фольклорных текстов для чайников[1]

Евгения Коровина, Никита Петров

При подготовке фольклорных текстов к публикации исследователи зачастую тратят большое количество времени на их предварительную обработку, разделение на смысловые группы, подготовку именного указателя, определение количества текстов, записанных в том или ином году, в той или иной местности, от информантов мужского или женского пола. Частично решения таких задач можно автоматизировать. На семинарах участники узнают, что такое регулярные выражения и как можно использовать Python и R для анализа корпуса текстов, а также каким образом можно визуализировать результаты обработки. На семинарах мы будем работать с двумя корпусами текстов: 500 текстов мифологических рассказов, записанных в Архангельской области, и 200 текстов детских страшных историй, записанных в селе и в городе в 1990–2015 гг.

Для работы участникам необходимо предварительно ознакомиться (учить необязательно!) с языком Python (https://www.python.org/) и R — свободным программным обеспечением для статистических вычислений (https://www.r-project.org/).

Тематические блоки занятий:

0. Подготовка корпуса текстов к анализу. Установка Python и R.

1. Регулярные выражения и их использование.

2. Разделение корпуса на отдельные тексты, поиск личных имен и создание указателя годов записи, сортировка текстов по длине.

3. Как устроены сочетания слов? Различные методы измерения тесноты связей между словами.

[1] Работа выполнена при поддержке гранта Российского научного фонда (проект №14–18–03384) «Истории, пересказываемые тысячелетиями: реконструкция динамики глобального распространения фабульных и образных элементов устных нарративов».

4. Словарный состав и частотный словник, стоп-слова и семантическое ядро текста.

5. Подготовка текстов к автоматической рубрикации. Классификация текстов.

6. Визуализация данных — возможности R.

Рекомендуемая литература

Кабаков Р.И. R в действии. Анализ и визуализация данных в программе R. URL: http://www.ievbras.ru/ecostat/Kiril/R/Biblio/R_rus/Kabacoff2014ru.pdf

Python 3 для начинающих. URL: http://pythonworld.ru/samouchitel-python