**ДАТИРОВКА ПИСЬМЕННОГО ИСТОЧНИКА НА ОСНОВАНИИ ЧАСТОТНОГО АНАЛИЗА СЛОВ**

¹Крыленков Л.В., ²Козлов Д.А., ²Беспалов А.М.

¹Санкт-Петербургский политехнический университет [Петра Великого](https://ru.wikipedia.org/wiki/%D0%9F%D1%91%D1%82%D1%80_I)

²Лицей Физико-техническая школа, Санкт-Петербург, Россия

sonkot25@mail.ru

Проблема датировки текстов - частая задача, которая стоит перед литературоведами, историками, генеалогами. Привязка текстов ко времени позволяет выстраивать последовательность происходящего, определять взаимосвязи между событиями. Обычно датировка текстов - это долгая и кропотливая исследовательская работа. В настоящее время у множества людей появился запрос на исследование истории своей семьи; появляется множество опубликованных семейных архивов - документов, писем, записок, зачастую не имеющих даты. В то же время, благодаря интересу к семейной истории, в сети появился и корпус четко датированных текстов - дневниковых записей, которые опубликованы в рамках проекта "Прожито". Возникает естественная идея, использовать этот корпус четко датированных текстов, чтобы попробовать на его основе датировать тексты с неизвестным временем создания.

Метод и промежуточные действия:

1. Была собрана в sql формате первичная база данных, состоящая из дневников и воспоминаний людей, относящихся к временному промежутку 1770-2019 гг.
2. База данных была переработана, удаление записей, не соответствующих требованиям по критериям: язык, тип записи (воспоминания людей о прошлом могут содержать, как лексику современности, так и прошлых лет, поэтому оставлены в основном дневники), наличие датирования. Преобразование базы данных в более удобный для чтения формат json.
3. Обработка текстов записей из дневников. Удаление знаков препинания (сохранение орфографических знаков, таких как дефис), приведение слов к общему виду.
4. Был проведён частотный анализ записей, принадлежащих одному десятилетию.
5. Была создана первая модель, позволяющая датировать письменный источник на основании проведённого частотного анализа.

Результат, полученный к данному моменту:

Модель была протестирована на исходной выборке для сбора данных и на других текстах – дневниках этого столетия. На выборке, которая использовалась для определения частотности слов по десятилетиям она дает хороший результат, что однако неудивительно.

На реальных же данных модель пока работает неидеально. Результат датировки может несколько отличаться от истинной даты, но top-5 accuracy (то есть точность среди пяти вариантов ответа программы) высока – верное десятилетие обычно слабо отклоняется от предлагаемого ответа и входит в top-5 предлагаемых. Точные результаты модель выдает, если источник является, как и база, письмом или дневником.

В дальнейшем планируется а) реализация иных возможных моделей и сравнение их точности и б) усовершенствование текущей модели с помощью некоторых приемов и эвристик.