Живые души XX века

по материалам корпуса дневников

Идея

“Большое видится на расстояньи”...
А малое - не видится. Но оно есть.

Учебники истории сохраняют для нас самые значимые события, произошедшие с людьми...

Значимые?
Так это действительно было самым важным, что с ними происходило?

Дневник - личная летопись, где человек фиксирует в истории вещи, важные лично ему.

Цели и задачи

Цель:

Узнать и показать, что считали важным (достойным записи) люди, жившие во времена “великих событий”

Задачи:

- разбить XX век на осмысленные периоды
- провести тематический анализ дневниковых записей в эти периоды
- сопоставить найденные темы с событиями национальной (и мировой) истории
- визуализировать результаты

Meet your dataset: дневниковые записи (источник)

Откуда у нас дневники?

Наши данные - это причёсанный дамп сайта "Прожито" от апреля 2019 года. Таблицы содержит несколько сотен тысяч записей за большой отрезок времени (от XVIII до XXI века, преимущественно — XX век), так что вам будет, где развернуться ;).

Как устроен датасет?

У нас есть две таблицы, `whole_table.csv` и `whole_table_with_lemm.csv`. Вторая отличается наличием колонки с лемматизированными (т.е. с приведенными в начальную форму словами) mystem записями.
В обеих таблицах есть колонка:

- `notes` - содержит сами дневниковые записи
- `dates` - дата записи в формате год/месяц/день
- `id` - айдишник автора (не записи!)
- `author` - имя автора записи

Практически все поля заполнены, у некоторых отсутствует дата (так как таблица отсортирована по датам, они в самом начале).
От авторов исследования: даты отформатированы как YYYY/(M)M/(D)D, отсутствовать могут день, месяц или год, отсутствующие элементы даты заменены нулями.

Данные (источник, сайт prozhito.org)

Ход работы: методы

Ноутбук в Google Colab с процессом работы, репозиторий на GitHub

- Были выбраны только столбцы **date** и **lemm** из таблицы с леммами
- Записи, в которых отсутствуют какие-то элементы даты, удалены
- Таблица была проиндексирована по столбцу **date**
- Лемматизированные тексты были разбиты на списки слов
- Стоп-слова были удалены
- Получившаяся таблица была разбита по периодам на 10 таблиц *(см. следующий слайд)*
- На обработанных текстах из каждой таблицы была обучена модель латентного размещения Дирихле, выделяющая 10 самых важных тем текстов для каждого периода.

Формат тем (топиков):

The probability for each word in each topic, shape (*num_topics*, *vocabulary_size*).

Ход работы: стэк и набор инструментов

Весь код работы написан на **Python** и выполнен в **Google Colab**.

- Для загрузки данных и сохранения моделей использовались модули **PyDrive, pandas, numpy, datetime, ast** и **joblib**.
- Для обработки текстовых данных использовались модули **re, string** и **NLTK**.
- Для topic modelling использовалась **LdaModel** из **gensim**.
- Для красивого отображения содержимого ноутбука использовались модули **pprint** и **tqdm**.
- Для отображения результатов в формате таймлайна использовался открытый бесплатный конструктор **SUTORI**.

Ход работы: параметры LdaModel

- **passes=15** (модель 15 раз проходится по корпусу)
- **iterations=100** (максимальное кол-во итераций во время вычисления распределения тем в корпусе)
- **per_word_topics=True** (для каждого слова в корпусе модель сортирует темы по вероятности принадлежности слова к теме)

Периодизация

Поскольку основная часть текстов корпуса приходится на советскую эпоху, решено было выделить **10 периодов**:

- 3 досоветских (Кровавое воскресенье, период между революциями и Первая мировая война)
- 6 советских
- 1 послесоветский

Периодизация

- *янв. 1905 – янв. 1907* (первая революция)
- *февр. 1907 – май 1914* (период между революциями)
- *июнь 1914 – октябрь 1918* (Первая мировая война и две главные революции в России)
- *ноябрь 1918 – декабрь 1927* (гражданская война, продразверстка, НЭП)
- *январь 1928 – август 1939* (коллективизация, Большой террор)
- *сентябрь 1939 – май 1945* (Вторая мировая война)
- *1946 – 1963* (промежуточный период истории, заканчивается «оттепелью»)
- *1964 – май 1987* (застой)
- *июнь 1987 – июнь 1990* (перестройка и развал СССР)
- *1991 – 2001* (после СССР, «лихие 90-е»)

Распределение кол-ва записей по периодам

first revolution:	*5821 entries*
between revolutions:	*18761 entries*
WWI:	*20362 entries*
civil war:	*35981 entries*
great purge:	*37175 entries*
WWII:	*62686 entries*
before ottepel:	*39742 entries*
stagnation:	*58121 entries*
perestroyka:	*6723 entries*
90s:	*13283 entries*

Результаты

- Для каждого периода выделены самые часто упоминаемые темы и ключевые слова в них.
- Пространство для качественных исследований
- Наглядное представление истории “глазами очевидца” для образовательных целей
- Хотели сделать визуализацию, но не успели. :(Но обязательно доделаем!

Таймлайн

TO DO:

- По результатам заметно, что данные не очень хорошо очищены. Сделать заново предобработку лемматизированных текстов и посмотреть, какие топики выделятся на более чистых данных.
- Интерпретировать полученные на чистых данных результаты.
- Визуализация! И не одним способом, а как минимум десятью!

Потенциал. Возможные идеи

Какие темы постоянны, независимо от истор. событий?

Какие группы авторов активнее всего писали в разные периоды?

Были ли события \ образы, не вошедшие в “официальную историю”, но про которые много писали?

... и т. д.

Исполнители

ОЛЬГИ:

Пичужкина
(обработка и визуализация данных)

4 курс бакалавриата “Фундаментальная и компьютерная лингвистика”, НИУ ВШЭ

Жукова
(гуманитарная экспертиза, таймлайн)

1 курс магистратуры Digital Humanities, НИУ ВШЭ