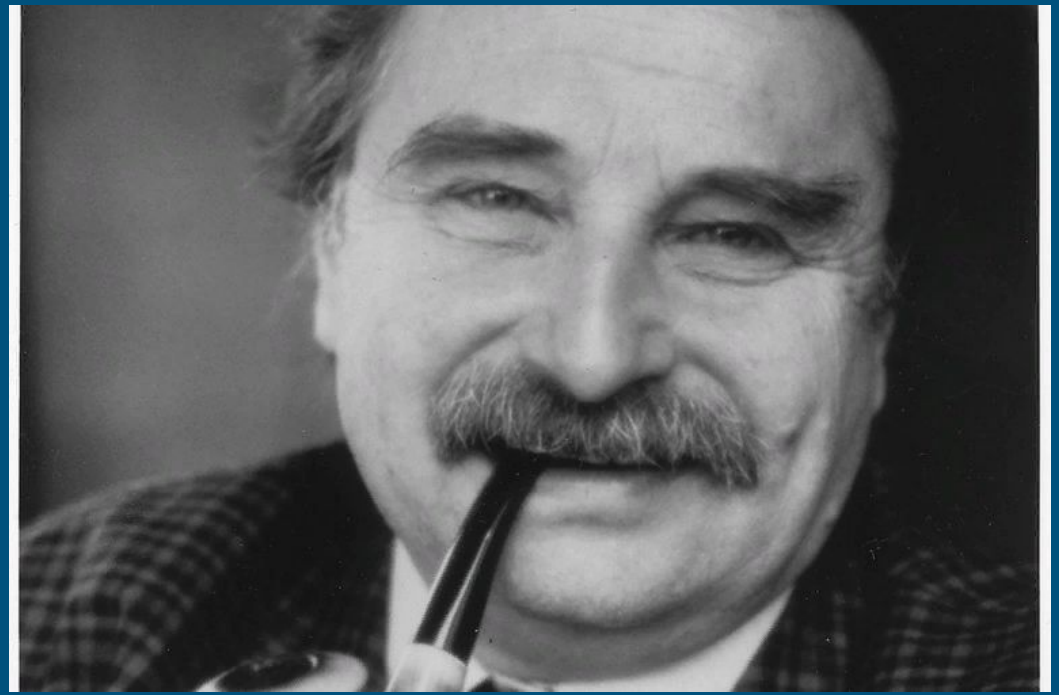
Исследование произведений Милорада Павича “Биография Белграда” и “Пейзаж, нарисованный чаем” как исторических источников с использованием методов цифровой гуманитаристики

Елизавета Дерзаева, Мария Подрядчикова

Москва, 21.05.2019

Милорад Павич (1929 — 2009)



Цифровые гуманитарные науки, или Digital Humanities, — это область исследований, в которой объединены цифровые методы и традиционные методы гуманитарных наук

Цифровые гуманитарные исследования в России

- Сетевой анализ русской драмы (2017-2019)

- Семантическая разметка художественных текстов для количественных исследований в филологии (на примере романа "Война и мир" Л.Н. Толстого) (2019)

- “Нейролирика” (эксперимент: стихи, созданные нейронной сетью) (2019)


МИЛОРАД
ПАВИЧ
Пейзаж, нарисованный чаем
АЗБУКА-КЛАССИКА


МИЛОРАД
ПАВИЧ
БИОГРАФИЯ БЕЛГРАДА

Почему мы выбрали для анализа русский язык

К сожалению, методы естественной обработки текста для сербского языка менее развиты (например, отсутствуют библиотеки для извлечения информации из текста).

Кроме того, поскольку мы не касаемся области определения авторства и стилометрии*, мы считаем, что различие между оригинальным текстом и переводом не будет критически значимым.

*исследование стилистики на основе статистических данных

Цель нашей работы

— проанализировать, каким образом цифровые методы обогащают исследования литературоведов и историков, изучающих наследие Милорада Павича

В чём преимущество современных цифровых исследований?

Наше исследование включает в себя два основных вида цифровых методов: предобработку текста и исследование его содержания.

Предобработка текста

Перечислим основные способы предобработки текста, которые мы использовали.

— токенизация текста (разбиение текста на отдельные слова)
“Кругом одно горе, и все мы в нем точно рыба в воде.” => “кругом, одно, горе, и, все, мы, в, нем, точно, рыба, в, воде”
— лемматизация текста (приведение данных токенов к начальной форме слова)
“кругом, одно, горе, и, все, мы, в, нем, точно, рыба, в, воде” => “кругом, один, горе, и, весь, мы, оно, точно, рыба, вода”
— удаление стоп-слов (оставляем только значимые для понимания текста слова)
“кругом, один, горе, и, весь, мы, оно, точно, рыба, вода” => “кругом, горе, точно, рыба, вода”

Исследование содержания текста

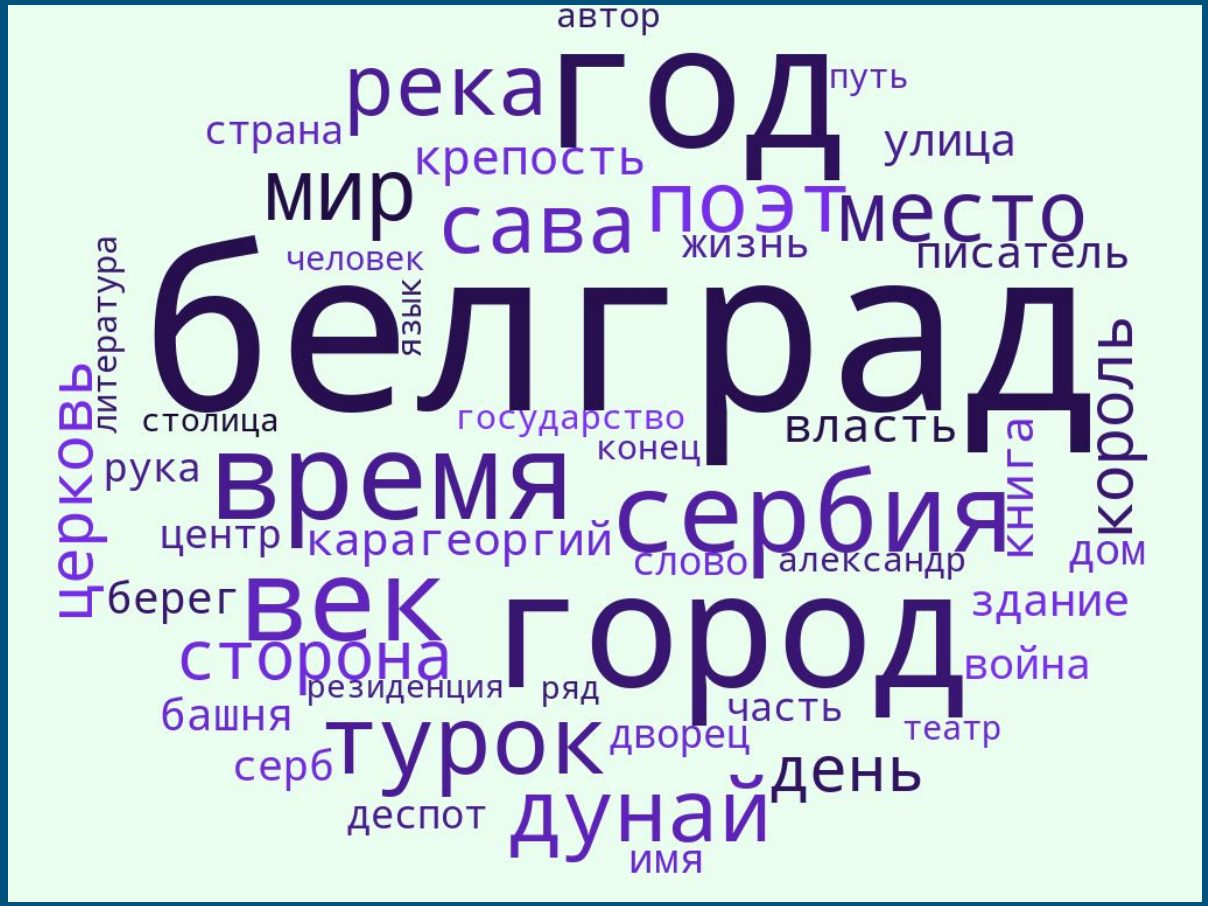
Три задачи:

1. Извлечение ключевых слов

2. Извлечение значимых имён

3. Извлечение значимых локаций

Использованные инструменты: библиотеки для обработки естественного языка (nltk, natasha, sklearn) для языка программирования Python.



топ-50 самых частотных слов “Биографии Белграда”

[('белград', 251), ('год', 132), ('город', 115), ('время', 59), ('век', 52), ('сербия', 49), ('турок', 35), ('дунай', 29), ('мир', 28), ('река', 27), ('поэт', 27), ('место', 26), ('сторона', 25), ('сава', 25), ('день', 23), ('король', 22), ('крепость', 21), ('церковь', 21), ('власть', 20), ('книга', 20), ('карагеоргий', 20), ('улица', 19), ('писатель', 19), ('здание', 19), ('башня', 18), ('берег', 18), ('война', 17), ('дом', 17), ('серб', 17), ('имя', 16), ('жизнь', 16), ('рука', 15), ('центр', 15), ('часть', 14), ('слово', 14), ('государство', 14), ('деспот', 14), ('дворец', 14), ('автор', 14), ('конец', 14), ('александр', 14), ('страна', 14), ('путь', 13), ('человек', 13), ('литература', 13), ('язык', 13), ('театр', 13), ('столица', 12), ('резиденция', 12), ('ряд', 12)]

('карагеоргий', 20) — 21-е место по частотности
('обрен', 10) — 70-е место по частотности

Использование меры tf-idf

Для сравнения того, какие слова наиболее значимы для различных глав “Биографии...”, мы использовали меру tf-idf, которая обращает особое внимание на слова, которые употребляются в данной главе, но не встречаются в других главах.

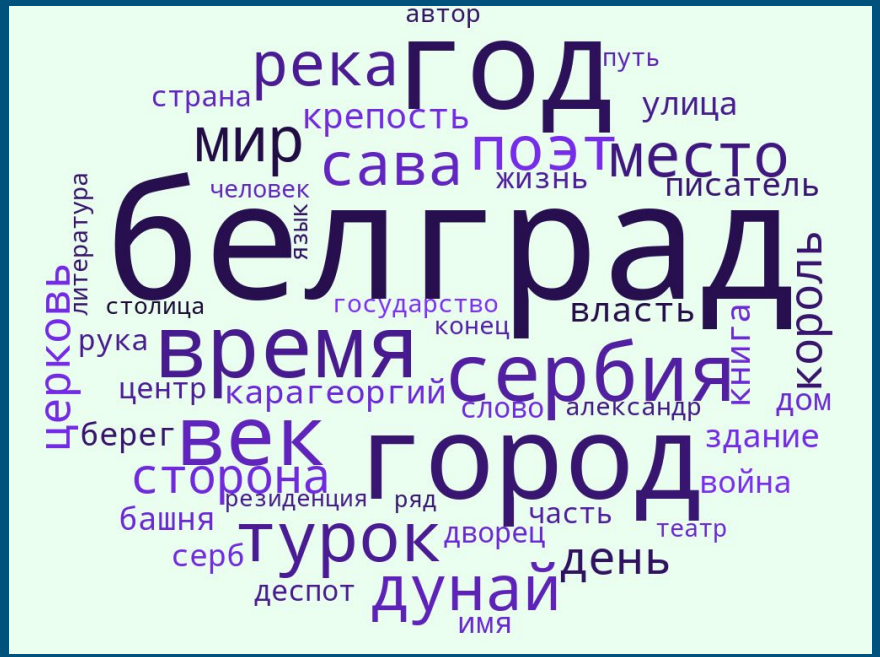
В приведенной далее таблице можно будет увидеть то, как менялся Белград на протяжении времени: древние “купальни”, “племя” и “король” сменяются более современными “государством”, “консулом” и “партией”, а в современность Белград входит с такими политическими терминами, как “демонстрация” и “оппозиция”.

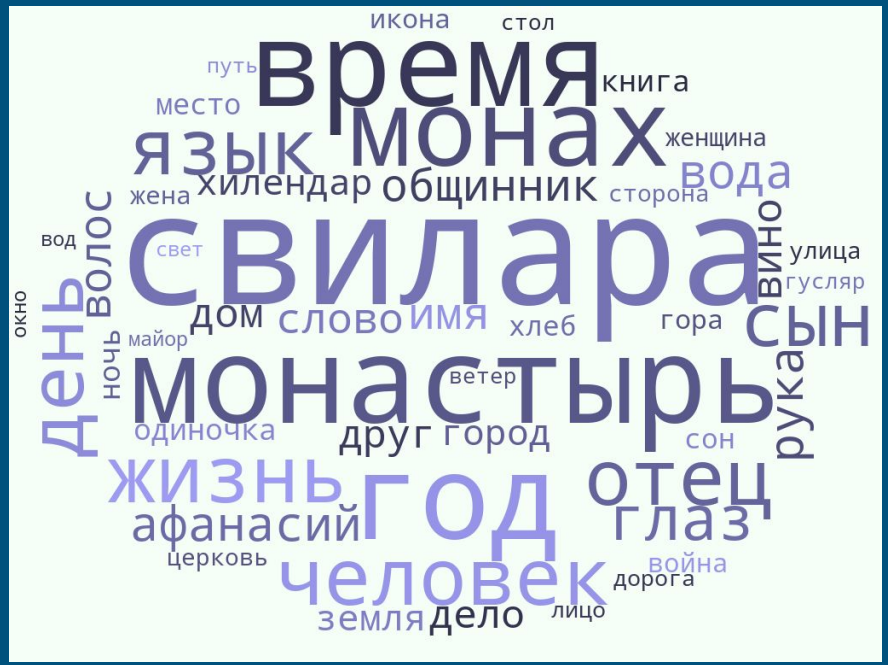
основание Белграда	['купальня', 'сингидунум', 'имя', 'поселение', 'белград', 'река', 'вод', 'город', 'дунай', 'век', 'эра', 'аргонавт', 'неолит', 'племя', 'место', 'душа', 'территория', 'лагерь', 'год']
Средневековье	['король', 'белград', 'год', 'сингидунум', 'город', 'император', 'век', 'правило', 'королевство', 'время', 'королева', 'имя', 'прибытие', 'симонида', 'почесть', 'радость', 'кателина', 'архиепископ', 'византиец']
XV век	['белград', 'деспот', 'город', 'турок', 'возвышение', 'год', 'христианство', 'запад', 'стефан', 'поэма', 'ариосто', 'иерусалим', 'богородица', 'царьград', 'лазарь', 'башня', 'константин', 'осада', 'падение']
шестнадцатый-восемнадцатый века	['белград', 'леандр', 'город', 'год', 'турок', 'учитель', 'птица', 'серб', 'ученик', 'крепость', 'власть', 'сторона', 'торговля', 'взгляд', 'время', 'век', 'язык', 'река', 'вампир']

девятнадцатый век: политика	['год', 'белград', 'турок', 'каменский', 'андерсен', 'обрен', 'движение', 'партия', 'бантыш', 'князь', 'карагеоргий', 'город', 'крепость', 'сербия', 'время', 'восстание', 'друг', 'мечеть', 'газета']
девятнадцатый век: культура	['белград', 'год', 'карагеоргий', 'король', 'кафан', 'сербия', 'консул', 'доситея', 'вывеска', 'князь', 'обрен', 'знак', 'династия', 'милоша', 'обрено', 'симич', 'жизнь', 'государство', 'век']

первая мировая война и время между войнами	['белград', 'война', 'барилль', 'командование', 'военный', 'бруно', 'австро', 'югославия', 'город', 'год', 'австриец', 'время', 'сторона', 'памятник', 'войско', 'событие', 'земуна', 'рождение', 'деятельность']
середина XX века и послевоенное время	['белград', 'пётр', 'илич', 'произведение', 'карагеоргий', 'рассказ', 'писатель', 'попович', 'война', 'мир', 'страна', 'поэт', 'представитель', 'принц', 'путь', 'литература', 'европа', 'александр', 'томислав']
рубеж XX и XXI веков	['белград', 'выбор', 'страна', 'сербия', 'дос', 'партия', 'город', 'милошевич', 'демонстрация', 'слободан', 'год', 'полиция', 'конец', 'югославия', 'оппозиция', 'зоран', 'демонстрант', 'коштуница', 'поддержка']
современность	['белград', 'театр', 'автор', 'книга', 'город', 'отель', 'год', 'роман', 'время', 'мир', 'конь', 'центр', 'кафе', 'сава', 'серия', 'издание', 'горан', 'выставка', 'здание']

топ-50 частотных слов для “Биографии…” и “Пейзажа…”





XX век в “Биографии...” и “Пейзаже...”

первая мировая война и время между войнами	['белград', 'война', 'барилль', 'командование', 'военный', 'бруно', 'австро', 'югославия', 'город', 'год', 'австриец', 'время', 'сторона', 'памятник', 'войско', 'событие', 'земуна', 'рождение', 'деятельность']
середина XX века и послевоенное время	['белград', 'пётр', 'илич', 'произведение', 'карагеоргий', 'рассказ', 'писатель', 'попович', 'война', 'мир', 'страна', 'поэт', 'представитель', 'принц', 'путь', 'литература', 'европа', 'александр', 'томислав']
первая часть “Пейзажа” (война и послевоенное время)	['свилара', 'год', 'афанасий', 'время', 'человек', 'отец', 'рука', 'язык', 'день', 'жизнь', 'глаз', 'монастырь', 'книга', 'война', 'сын', 'вода', 'земля', 'вино', 'монах']

Проблемы при извлечении имён собственных

— практически не выделяет биграммы и триграммы (сочетания из двух или трёх слов): и Александр Карагеоргиевич, и Александр Пушкин могут превратиться в Александра

— сербские фамилии распознаются как отчества (согласно анализу, в сочетании “Горан Брегович” Горан — имя, Брегович — отчество)

— принимает имена людей за названия локаций и наоборот (Сингидунум и Калемегдан попадают в список личных имён)

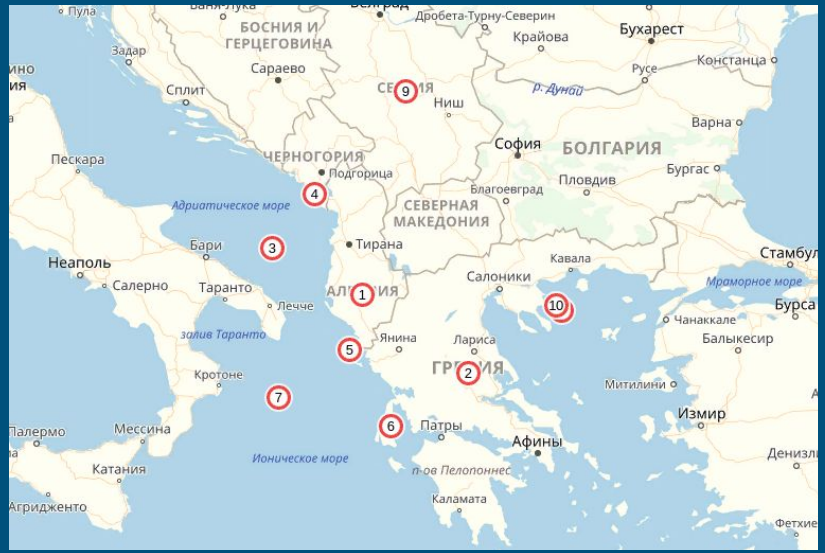
Извлечение имён из текстов

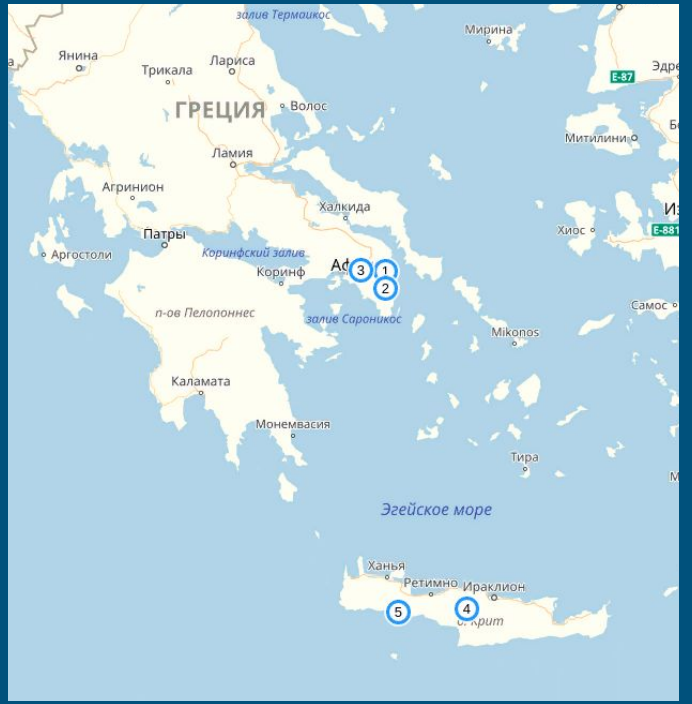
“Биография Белграда”	“Пейзаж, нарисованный чаем”
[('леандр', 12), ('сингидунум', 10), ('александр', 6), ('стефан', 5), ('йован', 5), ('каменский', 5), ('карагеоргий', 5), ('андерсен', 5), ('калемегдан', 4), ('небойша', 4)]	[('свилара', 141), ('афанасий', 45), ('карамустафа', 17), ('коста', 13), ('василий', 12), ('савва', 12), ('йоан', 7), ('йован', 6), ('степанида', 6), ('неманя', 5)]

Извлечение названий локаций

“Биография Белграда”	“Пейзаж, нарисованный чаем”
[('белград', 234), ('сербия', 40), ('тот', 29), ('дунай', 28), ('сава', 26), ('царьград', 11), ('земуна', 10), ('восток', 9), ('великое', 9), ('европа', 9)]	[('тот', 44), ('белград', 18), ('карамустафа', 18), ('куда', 16), ('афон', 16), ('нея', 12), ('греция', 12), ('никола', 11), ('красный', 9), ('который', 8)]

Маршрут Косты Свилара (“Пейзаж…”)





— слайд из презентации к выступлению к. фил. н. Бориса Орехова на конференции Яндекс Data Science, 02.03.2019 г.

Заключение

Таким образом, цифровые методы позволяют оптимизировать работу над многими задачами, важными для литературоведов и историков.

Кроме того, в ходе цифрового анализа данных можно получить информацию, неочевидную при традиционных исследованиях.

Спасибо!