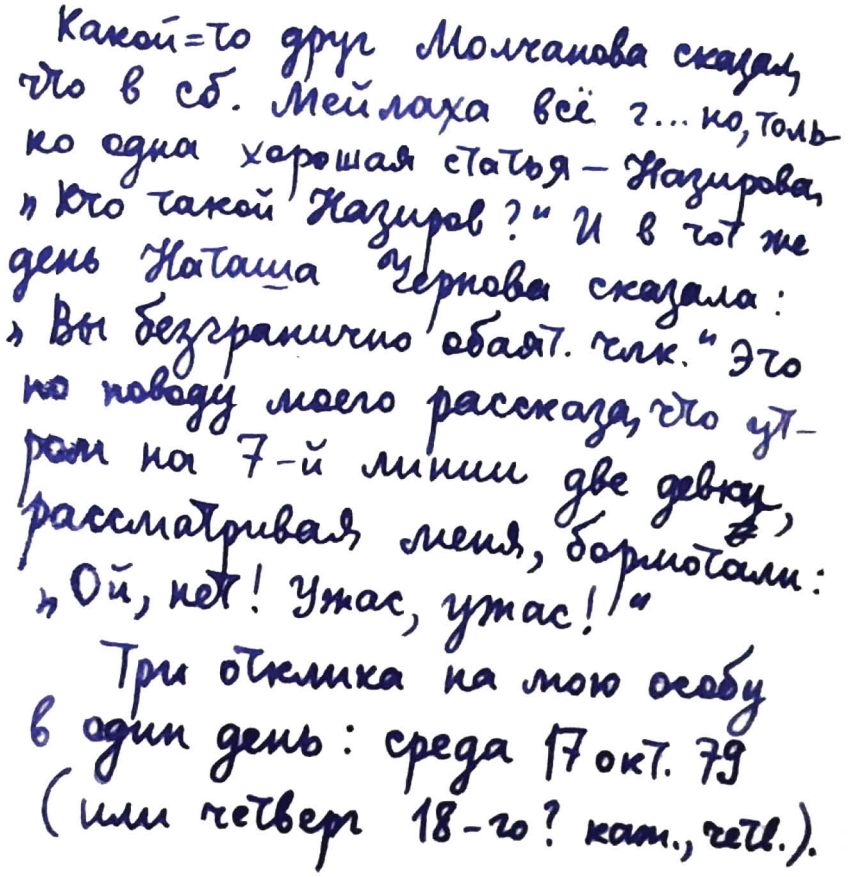
Исследования



Исследования

Кластеризация и тематическое моделирование текстов Р. Г. Назирова

А. А. Лапидус

Национальный исследовательский университет «Высшая школа экономики»

Текстовая близость

Для обработки были взяты тексты Р. Г. Назирова, находящиеся в свободном доступе в репозитории на Github[1].В репозитории имеется таблица с метаданными текстов набора и сведениями о ручной тематической классификации этих произведений. Предобработка производилась с помощью пакета pymorphy2. Тексты были разделены на токены, удалены стоп-слова. Для всех лемм во всех текстах подсчитана мера TF-IDF, которая стала основой векторизации каждого произведения. Между векторами текстов вычислена косинусная близость.

Наиболее близкими текстами разных тематик оказались:

- «Проблема художественности Ф. М. Достоевского» (тема: *Достоевский*) и «Специфика художественного мифотворчества Ф. М. Достоевского» (тема: *миф*). Расстояние: 0.537. Вполне естественная близость, к тому же показывающая условность исходной тематической классификации: статьи о мифе вполне могут быть посвящены творчеству Достоевского и наоборот.

- «Творческие принципы Достоевского» (тема: *Достоевский*) и «Специфика художественного мифотворчества Ф. М. Достоевского» (тема: *миф*). Расстояние: 0.584. Аналогичный случай.

- «Владимир Одоевский и Достоевский» (тема: *Достоевский*) и «О месте Одоевского в русской литературе» (тема: *русская литература*). Расстояние: 0.743. Одна статья является расширенным вариантом другой.

- «Равноправие автора и героя в творчестве Достоевского» (тема: *Достоевский*) и «Автономия литературного героя» (тема: *русская литература*). Расстояние: 0.5.

[1] https://github.com/nevmenandr/nazirov-texts-dataset

Весьма похожи оказались дневники (среднее расстояние 0.55) и рассказы «После выставки» и «Утро в городе Солнца» (косинусное расстояние: 0.678).

На рис. 1 можно посмотреть на граф отношений между текстами Назирова, косинусное расстояние между которыми составило больше 0.5. Видно, что граф крайне разряженный, средняя степень узла 1.11, всего узлов 36, а ребер — 40.

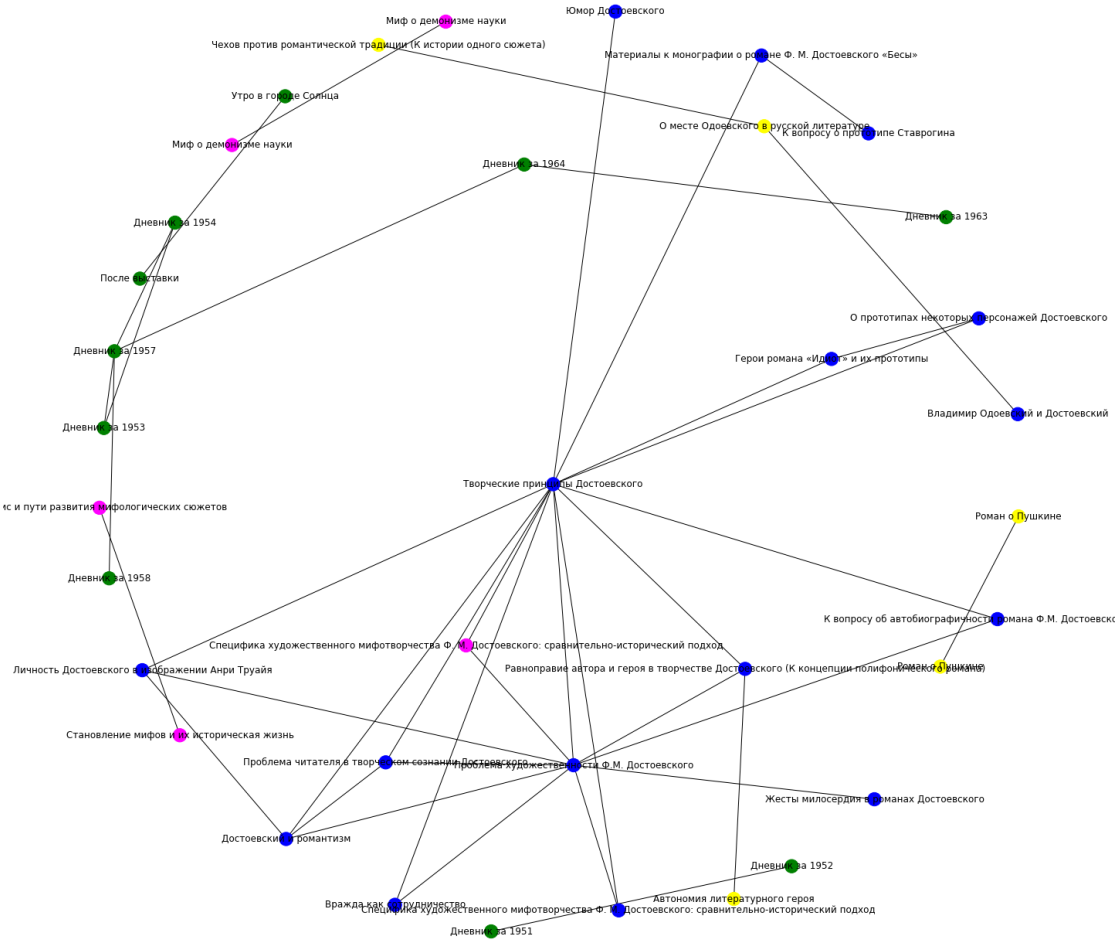


Рис. 1: Граф расстояний в текстах Р. Г. Назирова

Тематическое моделирование

LDA Topic model

У нас есть 4 больших тематических набора внутри текстовой коллекции: дневники, работы о Достоевском, работы о русской литературе, работы о мифологии. Гипотеза состояла в том, что тематическое моделирование отразит эти 4 темы и выделит ключевые слова для каждой из них. То есть если мы зададим 4 темы для алгоритма LDA, то каждая из тем будет соответствовать одному из тематических наборов.

Отчасти гипотеза подтвердилась. См. наборы тем для 4 топиков (перед словом указан коэффициент вероятности этого слова для этой темы):

1. 0.007 * *тургенев*; 0.005 * *одоевский*; 0.005 * *лермонтов*; 0.003 * *раскольников*; 0.003 * *онегин*; 0.003 * *игрок*; 0.003 * *идиот*; 0.003 * *алексей*; 0.003 * *прототип*; 0.003 * *фабула*; 0.002 * *белинский*; 0.002 * *карамазов*; 0.002 * *мышкин*; 0.002 * *петербургский*; 0.002 * *жанр*.

2. 0.004 * *американский*; 0.003 * *вчера*; 0.003 * *сша*; 0.003 * *октябрь*; 0.002 * *правительство*; 0.002 * *партия*; 0.002 * *фильм*; 0.002 * *сталин*; 0.002 * *государь*; 0.002 * *американец*; 0.002 * *президент*; 0.002 * *март*; 0.002 * *министр*; 0.002 * *рабочий*; 0.002 * *окно*.

3. 0.005 * *пророк*; 0.005 * *симон*; 0.004 * *еврей*; 0.004 * *меч*; 0.004 * *соль*; 0.004 * *череп*; 0.003 * *эпос*; 0.003 * *роланд*; 0.003 * *карл*; 0.002 * *аттил*; 0.002 * *рыба*; 0.002 * *рыцарский*; 0.002 * *римляна*; 0.002 * *от*; 0.002 * *иудей*.

4. 0.006 * *мифология*; 0.005 * *фрейд*; 0.004 * *обряд*; 0.003 * *младший*; 0.003 * *архаический*; 0.003 * *бессознательный*; 0.003 * *христианство*; 0.003 * *философия*; 0.003 * *животное*; 0.002 * *маркс*; 0.002 * *моисей*; 0.002 * *иисус*; 0.002 * *богиня*; 0.002 * *тотемный*; 0.002 * *половый*.

Тема 1 соответствует одновременно и набору «русская литература», и набору «Достоевский».

Тема 2 соответствует набору «дневники».

Тема 3 соответствует историческому фону в мифологических работах.

Тема 4 также характеризует работы о мифологии, но на этот раз описывает как раз ядерную лексику поля «миф».

NMF

Алгоритм NMF полностью подтвердил гипотезу и выделил 4 темы, соответствующие тематическим наборам в коллекции.

Выделенные в результате тематического моделирования методом NMF темы можно интерпретировать в соответствии с темами коллекции: Достоевский, дневники, мифология, литература. Результаты можно посмотреть в таблице 1.

Topic # 01	Topic # 02	Topic # 03	Topic # 04
достоевский	советский	миф	пушкин
роман	американский	сказка	гоголь
герой	рука	бог	роман
князь	москва	сюжет	русский
раскольников	улица	мифология	петербург
образ	сша	обряд	одоевский
ставрогин	идти	древний	государь
писатель	вчера	мотив	тургенев
преступление	война	религия	жуковский
автор	город	земля	сюжет
читатель	вечер	народ	фабула
карамазов	фильм	легенда	поэма
рогожина	друг	герой	граф
творчество	сталин	век	герой
далее	хрущёвый	смерть	повесть

Таблица 1: Слова, выделенные для тем алгоритмом NMF

Кластеризация

Кластеризация производилась на основе векторов текстов со значениями TF-IDF для лексем.

В результате кластеризации выделяются 4 кластера по основным темам коллекции: литература, дневники, Достоевский, мифология.

Кластер	Тема	Число текстов
1	литература	16
	миф	2
2	разное	19
	литература	1
	миф	1
3	Достоевский	25
	литература	3
	миф	1
4	миф	36
	литература	6

Таблица 2: Результаты кластеризации

В кластер, связанный с литературой, попадают две статьи про мифологию («Подлинный смысл Поликратова перстня», «Сюжет об оживающей статуе»).

В кластер, к которому относится большинство статей о Достоевском, попадают несколько текстов о литературе и мифологии («Автономия литературного героя», «Фигура умолчания в русской литературе», «О месте Одоевского в русской литературе», «Специфика художественного мифотворчества Ф. М. Достоевского»). В статье «Автономия литературного героя» черты литературных героев рассматриваются в том числе на примере героев Достоевского. В статье «Фигура умолчания в русской литературе» приводится рассуждение о князе Мышкине из романа «Идиот». В работе «О месте Одоевского в русской литературе» Назиров пишет о связи творчества Одоевского и Достоевского.

В кластере, включающей в основном тексты о мифологии, встречаются статьи с темой «литература»: «Продолжение как форма обновления традиции», «Гюго-Флобер, или невозможная любовь дикаря, «О влиянии фрейдизма на современную литературу», «Сюжет как компромисс».

tSNE

На рис. 2 видно, что в результате снижения размерности векторов текстов с помощью метода t-SNE по полученному отображению на плоскость тексты коллекции кластеризуются в соответствии с темами. Некоторые статьи о литературе расположены близко к статьям о мифологии, и некоторые тексты, относящиеся к темам "Литература"и "Мифология"оказываются близкими к кластеру с текстами о Достоевском.

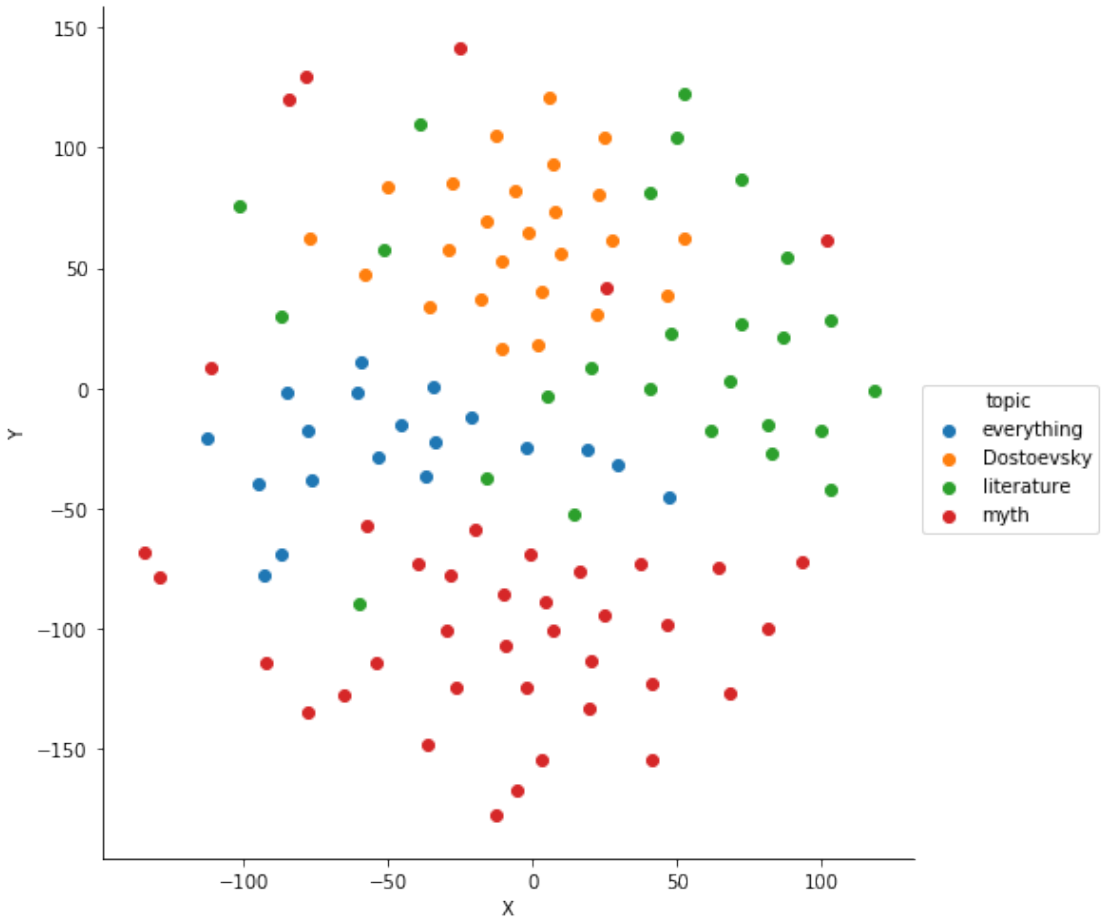


Рис. 2: Визуализация снижения размерности векторов текстов Р. Г. Назирова