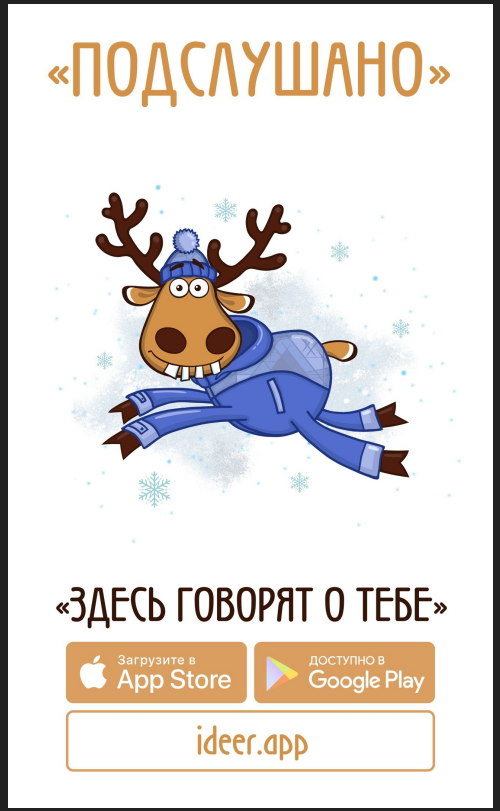
Исследование дискурса сообщества Вконтакте

На примере паблика «Подслушано»

Что такое «Подслушано»

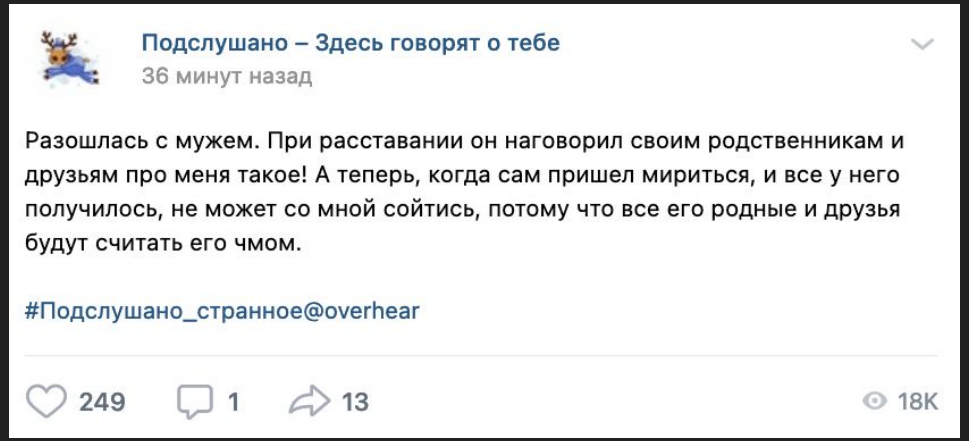
«Подслушано – Здесь говорят о тебе» – это сообщество в Вконтакте с более, чем 4 млн участников, позиционирующее себя как «уникальный ежедневный поток людских откровений». На стене паблика пользователи оставляют свои истории, откровения, наблюдения.

Помимо главного «Подслушано», есть региональные/локальные сообщества подобной тематики. Например, «Подслушано Архангельск» или «Подслушано ВШЭ».



Почему это сообщество интересно

В большинстве сообществ, будь то СМИ или развлекательные паблики, записи на стене состоят из пары слов/предложений, часто рекламного характера. В «Подслушано» формат постов – это маленькие истории, где есть и нарратив, и стиль авторов, пишущих свои мысли простой разговорной речью. Следовательно, в текстах постов можно выделить какие-то паттерны речи, классифицировать посты по темам, найти в них ключевые слова. В целом, ответить на вопрос: о чем пишут люди в одном из самых откровенных сообществ Вконтакте.



Задачи

1. Спарсить посты со стены сообщества
2. Применить тематическое моделирование, чтобы найти темы и их ключевые слова
3. Сделать выборку постов по теме
4. Построить граф ключевых слов и посмотреть, в каком контексте они используются
5. С помощью TextRank найти наиболее показательные предложения из постов по выбранной теме и сделать суммаризацию

Данные

- Количество всех постов на странице сообщества «Подслушано»: **136 374**
- Выборка из постов 2020 года: **16 930**

Актуальность

Интернет-сообщества характеризуется своим дискурсом: жанром, речевыми особенностями, закономерностями речи, специфическими темами и особым лексиконом. Для изучения дискурса используется дискурсивный анализ, когда через имеющиеся в тексте языковые единицы выводится структура целого фрейма, то есть понятий и концепций, о которых говорят в этом сообществе. Данная работа направлена на то, чтобы провести дискурс-анализ на большом массиве текстов и использовать методы машинного обучения, которые позволят найти закономерности на тематическом и лексическом уровнях.

Topic modeling

С помощью алгоритма моделирования тем LDA Mallet – Latent Dirichlet Allocation Маллета – и алгоритма для нахождения оптимального количества тем для LDA я выделила 8 тем, которые впоследствии стали:

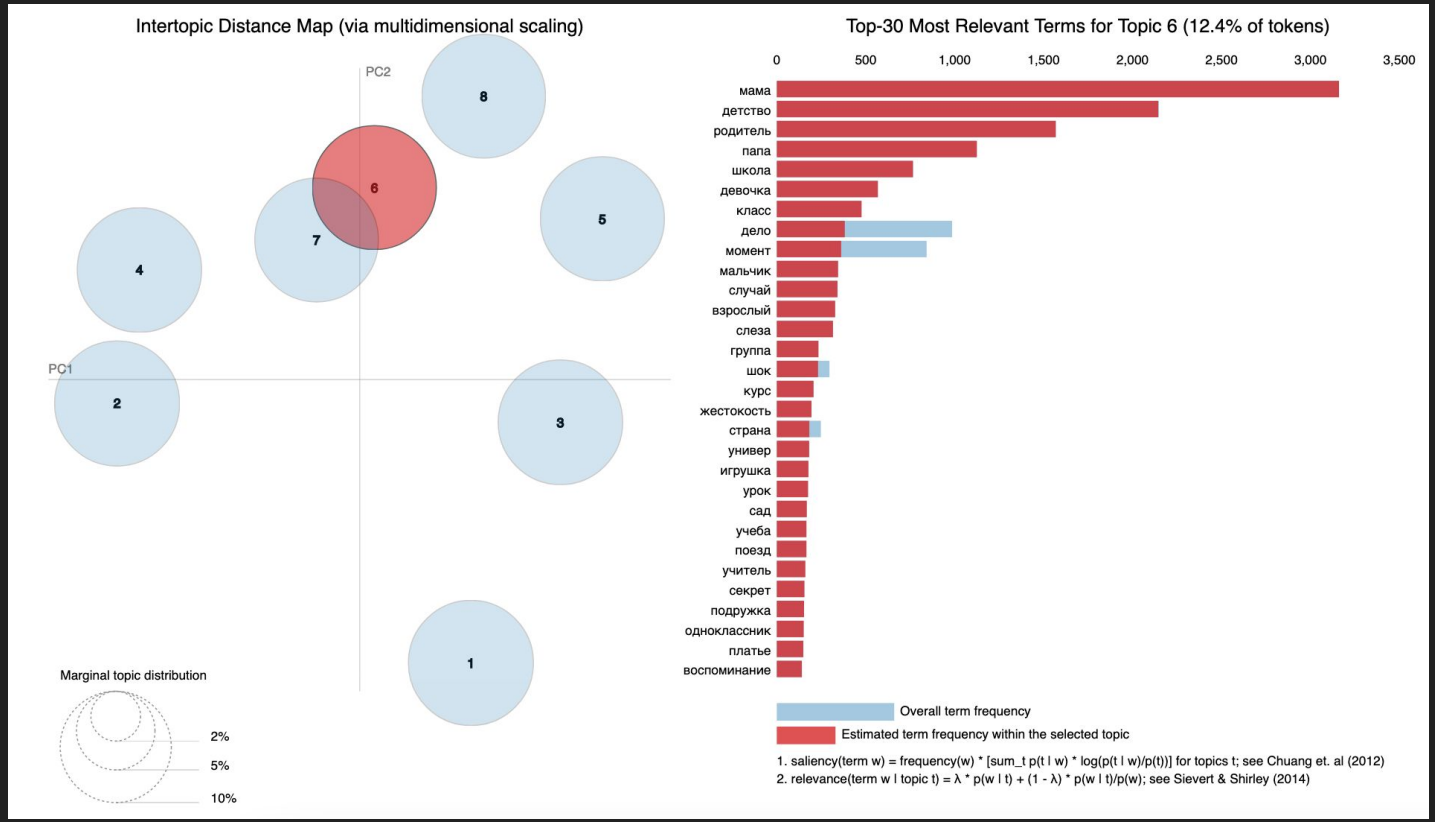
0: конфликты_семья,
1: школа_студенчество,
2: страхи_личное,
3: наблюдения_общество,
4: отношения_интим,
5: подарки_детство_коты,
6: работа,
7: тело_болезни

Для дальнейшей работы я сократила выборку до постов 2020 года по теме школы_студенчества. Выборка = **2073** поста.

Topic modeling

[ссылка на html]

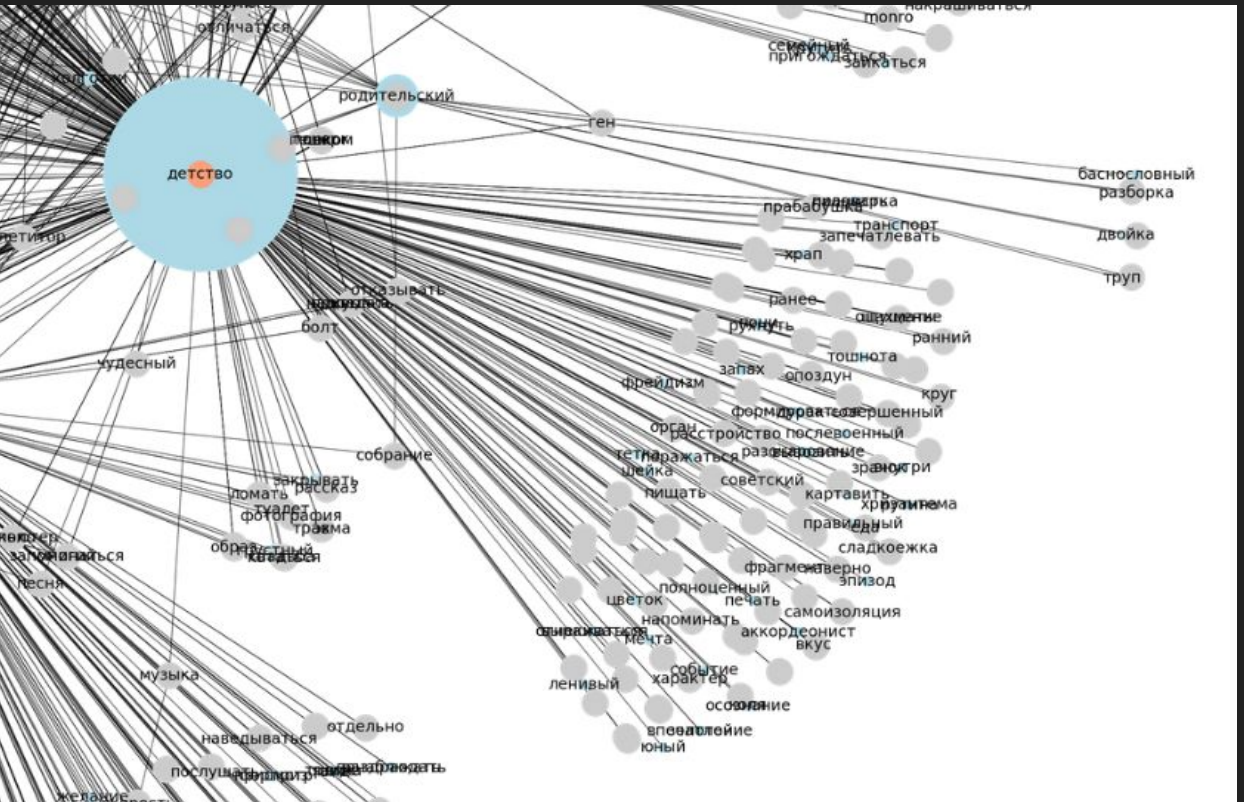
По каждой теме я получила ключевые слова. В теме про школу_студенчество выделились 10 ключевых слов: мама, детство, родитель, папа, школа, девочка, класс, дело, момент, мальчик.



Key words contexts

Для каждого ключевого слова по теме я решила найти своим кастомным методом соседние слова и составить пары встречаемости слов.

Key words contexts



Например, слово «детство» встречается в окружении таких слов, как сладкоежка, картавить, прабабушка, советский. Слово «класс» – чаепитие, птушник. Сейчас сложно на примере одной темы и одного паблика сказать, как характеризуют эти слова посты пользователей.

TextRank

Я решила попробовать преобразовывать текст в граф, где вершинами стали предложения, а ребрами связь между ними, которая определяется количеством одинаковых слов в предложениях. Таким образом, у каждого предложения был бы свой рейтинг и предложения как бы ссылались друг на друга, как в PageRank.

TextRank

Выделив предложения с большим рейтингом, можно сделать суммаризацию текста по предложениям. Я делала суммаризацию не на полном тексте, а на маленькой выборке с целью экономии ресурсов. Получилось, что самыми «показательными» предложениями оказались:

1. А я стояла и делала вид, что всё в порядке, и делала это три года, пока меня не увезли на скорой с нервным приступом.
2. Так что я знаю только, что это мальчик, его имя и что он родился в 2000 году.
3. Когда в детстве не было еды, то я рисовал ее в тетрадке, а потом вырывал листок и ел.
4. У меня в детстве не было бабушек и дедушек, то есть номинально они были, только очень далеко, и мы почти не общались.
5. В детстве у меня любимыми игрушками были не куклы, не машинки и не мягкие игрушки, а бумажные.)]

TextRank

Если ограничиваться темой, которую я назвала школа_студенчество, оказывается, здесь преобладают воспоминания (тяжелые и неприятные) из детства.

1. А я стояла и делала вид, что всё в порядке, и делала это три года, пока меня не увезли на скорой с нервным приступом.
2. Так что я знаю только, что это мальчик, его имя и что он родился в 2000 году.
3. Когда в детстве не было еды, то я рисовал ее в тетрадке, а потом вырывал листок и ел.
4. У меня в детстве не было бабушек и дедушек, то есть номинально они были, только очень далеко, и мы почти не общались.
5. В детстве у меня любимыми игрушками были не куклы, не машинки и не мягкие игрушки, а бумажные.)]

Future work

С помощью данного проекта я примерно поняла, что хочу видеть в диссертации. Я бы хотела изучать дискурс не только одного глобального «Подслушано», но и региональных – по городам. Возможно, в других сообществах поднимаются свои специфические темы. Помимо тематического моделирования, можно классифицировать посты по тегам вроде #стыдно #бесит, которые есть в конце почти каждого поста. Также я бы использовала TextRank, чтобы выделить ключевые слова.

Еще можно посмотреть, как развивались темы в разные годы (например, сообщество «Подслушано» существует с 2013 года), какие смысловые сочетания слов были самые употребляемые в разный период и как это может характеризовать действительность, например, в разных городах России.