**Отчет по работе над статьей по анализу ценностей в корпусе русского рэпа**

*Бойченко Антон*

*11.04.19*

В прошлый раз был представлен план статьи и была описана база данных исследования. С тех пор была проделана работа по предобработке текстов корпуса и построению тематических моделей. Сперва рассмотрим произведенный процесс предобработки данных.

Все тексты были предобработаны: удалена пунктуация, тексты приведены к нижнему регистру, токенизированы и лемматизированы (приведены к начальной форме) с использованием морфологического анализатора Yandex mystem3, затем были удалены стоп-слова. Отдельно стоит отметить, что на этапе предобработки были удалены латинские буквы, тем самым тексты не содержат английских и иных слов, использующих латиницу. Такая особенность предобработки обусловлена первичным анализом и построением тематических моделей, на основании которых удалось понять, что английские слова в текстах встречаются (1) достаточно редко и лишь искажают анализ, (2) не несут содержательного смысла в большинстве случаев, (3) объединяются в одну тему, интерпретация которой сводится к пониманию того, что в нее включены английские слова.

По завершении предобработки были удалены дубликаты и искаженные тексты. Удаление дубликатов производится именно после предобработки, так как непредобработанные тексты могут различаться технически (регистр, пунктуация, количество пробелов и проч.), но содержательно представлять идентичные наблюдения. Для программного же обеспечения технически разные тексты не могут считаться дублирующими друг друга и будут входить в анализ несколько раз, что искажало бы его результаты. Предобработка позволяет нивелировать подобные различия и выбрать из всей базы набор действительно уникальных текстов. После проведения предобработки 10 889 наблюдений было отобрано для анализа.

Важно отметить, что существует несколько ограничений предобработки в рамках настоящего исследования. Во-первых, поскольку оба сайта являются частично созданными пользователями, могут встречаться орфографические ошибки в написании некоторых слов, которые не учтены в морфологическом анализаторе. Таким образом, два слова, несущие один и тот же смысл, будут восприниматься программой как разные, искажая результаты. Несмотря на это, ошибки встречаются редко, так как предполагается, что тексты изначально собираются из официальных источников артистов. Во-вторых, некоторые слова обрабатываются таким образом, что искажается их изначальное контекстуальное значение. Например, одним из наиболее частых случаев является слово «бабки», имеющее два значения: множественное число слова «бабка» и разговорное слово «бабки» в значении «деньги». Поскольку анализатор возвращает слова в единственном числе, он возвращает слово «бабка», однако значение «деньги» встречается гораздо чаще в контексте рэпа. Учитывая проблему, подобные случаи были вручную заменены с учетом первичного анализа контекста использования.

Отдельно стоит отметить, что после сбора и предобработки текстов, припевы, которые повторяются несколько раз, сохранились в текстах. Припевы были оставлены в текстах по причине того, что припев видится именно той составляющей текста песни, которая специально выбирается исполнителем для повторения и с большей вероятностью запоминается аудитории. Также именно припев может содержать «хук» – повторяемое несколько раз слово или выражение, которое наибольшим образом запоминается и повторяется.

Теперь, разобравшись с предобработкой текстов, перейдем к рассмотрению результатов. Как уже было замечено в прошлый раз, для построения моделей используется метод BigARTM, доказавший возможность преодоления проблем методов LDA и PLSA. При этом же, благодаря усилиям коллег был создан новый метод тематического моделирования, основанный на категориальном методе главных компонент. Рассмотрим модель, полученную с использованием этого метода.

*Таблица 1*

**Итог построения тематической модели с использованием метода CatPCA с использованием корпуса с припевами текстов**

|  |  |
| --- | --- |
| **Токены** | **Темы** |
| любитель, фигура, член, стыдно, четко, понятный, горячий, быдло | Телесные контакты |
| любовь | - |
| измена, весы, правильный, респект, концерт, пример | Моральная оценка поведения |
| х\*й | - |
| рэп, хип | Создание хип-хопа |
| высотка, случай, жарко, сценарий, тусовка, сотка | - |
| волос, дрожь, смело, шея, нежно, дыхание, разлука | Отношения и секс |
| градус, бокс, бар, сталь, честь | - |
| классика, телок, телка, продюсер, модный | Объективизация женщин |
| награда, клетка, труд, похожий, рожа | - |
| подвал, долгий, срок, голый, искра | Заключение |
| гад, парад, тупо, баран | - |
| смс, вено, признание | - |
| земной, погон, печень, трудный | - |
| кадр, картинка, секунда, экран | Кинематограф (предположительно) |
| тот, сюжет, мука, горизонт | - |
| цифра, кнопка, трубка, пробка | - |
| обед, ужин, завтрак, дорого | Прием пищи |

Как видно из данных, модель оказывается очень сложно интерпретируема. Предполагаемые интерпретации выписаны в правый столбец таблицы. Далее была протестирована модель с использованием этого же метода, однако исключавшая припевы из анализа.

*Таблица 2*

**Итог построения тематической модели с использованием метода CatPCA с использованием корпуса без припевов текстов**

|  |  |
| --- | --- |
| **Токены** | **Темы** |
| лужа, хор, повтор, достойный, корень, горе |  |
| рэп, хип | Музыка/реп |
| кол, сумма, царь, замес, танк, шум, воин |  |
| компания, пламень, внутри, вино, нога |  |
| телка, классика, продюсер, телок, парад, модный |  |
| тормоз, резко, газ, второй, поворот, мотор, машина, руль |  |
| расклад, бро, рок, печень |  |
| золото, голод, норма, доллар, холод | Материальные цености |
| крик, стон, тупик, равный |  |
| возможность, оскал, точка, реальность |  |
| нежно, нежный, свидание, тишина | Отношения полов |
| клан, око, верный, словом |  |
| правильный, смысл, поступок, вариант, нация |  |
| милиция, бабушка, секунда, тип, странный, мораль |  |
| особый, целый, стадо, баба |  |
| задница, прикол, зад |  |
| смешной, дешевый, радость, смех | Синонимы |
| концерт, пример, космос, широкий, жестко |  |
| вечный, вечность, вечно | Синонимы |
| принц, ясный, принцесса, король | Семантические ассоциаты |

Сравнивая первую и вторую модель видно, что интерпретация второй затруднена еще больше. Предположительно причиной тому становится небольшая средняя длина текстов, которые стали меньше после удаления припевов. Таким образом, метод CatPCA не будет использован в будущей публикации.

Рассмотрим модели с использованием метода BigARTM с включением и невключением припевов. Модель с припевами выглядит следующим образом:

*Таблица 3*

**Итог построения тематической модели с использованием метода BigARTM с использованием корпуса без припевов текстов**

|  |  |
| --- | --- |
| **Токены** | **Темы** |
| рука голова грязь нога пьяный хер вес антон грустно монета | Не поддается интерпретации |
| клуб водка пиво черный девчонка бутылка пати пацан парень дом | Тусовки и алкоголь |
| любовь твой сердце друг мой слово глаз один душа слеза | Любовь |
| стиль дон закон слово пацан новый рука клан бок номер | Братство и клановые отношения |
| хип рифма рэп песня слово мой готовый хоп текст микрофон | Создание музыки/репа 1 |
| солнце небо ветер сон дым снег весна лето море вода | Погода |
| деньги игра новый дело правда человек должный свой один правило | Бизнес |
| сука белый п\*здец че рука брат голова друг косяк пара | Специфичная реп лексика и ругательства 1 |
| мама брат ребенок сын мать отец детство папа семья дом | Семья и дом |
| высоко страна свобода россия власть человек родина народ душа слово | Мораль |
| окно дождь тихо последний лист осень капля холодный голос стена | «Грустная» погода |
| война тело страх глаз лицо кровь смерть мир лес раб | Страх и страдания |
| рэп твой русский сука нах\*й х\*й парень трек рот рэпер | Специфичная реп лексика и ругательства 2 |
| утро завтра новый ночь день вечер окно час друг домой | Время |
| кровь бой ад боль душа рай мой воля бог судьба | Борьба за жизнь |
| рэп бит трек музыка рука дым куплет голова микрофон звук | Создание музыки/репа 2 |
| мир небо свет путь сила земля дорога бог душа жизнь | Апелляция к духовным понятиям и фатализм |
| город улица москва район дом столица двор машина дорога метро | Город и его части |
| че пацан ак витя х\*й братан за\*бись дело ноггано бла | Обращения |
| твой ночь девочка тело подруга клуб глаз рука секс детка | Тусовки и секс |

Модель же без припевов представлена в Таблице 4.

*Таблица 4*

**Итог построения тематической модели с использованием метода BigARTM с использованием корпуса с припевами текстов**

|  |  |
| --- | --- |
| **Токены** | **Темы** |
| рука, голова, дело, антон, вес, сука, стена, ствол, пара, пацан | Не поддается интерпретации |
| черный, пиво, дом, план, пацан, мама, че, спорт, водка, модный | Тусовки и алкоголь |
| любовь, твой, друг, сердце, мой, глаз, слово, один, чувство, душа | Любовь |
| пацан, слово, че, закон, новый, бок, стиль, дело, рука, такой | Братство и клановые отношения |
| хип, рифма, песня, слово, мой, текст, рэп, музыка, наш, готовый | Создание музыки/репа 1 |
| море, солнце, дым, весна, вода, ветер, лето, снег, сон, берег | Погода |
| деньги, новый, человек, жизнь, путь, игра, правда, мир, один, дело | Бизнес |
| че, брат, рука, белый, голова, вася, друг, пацан, мент, сука | Специфичная реп лексика и ругательства 1 |
| мама, брат, сын, бог, отец, мать, ребенок, смерть, жизнь, кровь | Семья и дом |
| страна, свобода, власть, слово, рука, россия, родина, голова, закон, мой | Мораль |
| окно, дождь, тихо, осень, лист, дверь, холодный, ночь, стена, глаз | «Грустная» погода |
| война, мир, тело, кровь, страх, смерть, земля, воин, раб, новый | Страх и страдания |
| рэп, твой, сука, русский, нах\*й, х\*й, парень, жопа, рот, трек | Специфичная реп лексика и ругательства 2 |
| утро, новый, вечер, ночь, день, дом, завтра, друг, пара, работа | Время |
| кровь, бой, голова, мозг, твой, мир, душа, белый, лицо, больной | Борьба за жизнь |
| рэп, бит, трек, рука, хип, музыка, стиль, микрофон, слово, куплет | Создание музыки/репа 2 |
| небо, свет, мир, огонь, звезда, земля, душа, глаз, путь, рука | Апелляция к духовным понятиям и фатализм |
| город, улица, москва, район, дом, столица, двор, метро, питер, дорога | Город и его части |
| че, х\*й, пацан, ак, витя, п\*здец, пох\*й, один, жесть, братан | Обращения |
| клуб, твой, ночь, девочка, подруга, секс, рука, тело, детка, дама | Тусовки и секс |

Видится, что модель с припевами оказывается лучше интерпретируемой (предполагаемая причина тому была обозначена выше), поэтому выбрана будет именно модель с использованием метода BigARTM на корпусе с припевами. Фактором, определяющим выбор модели, служила именно возможность содержательной интерпретации тем, однако стоит все-таки обозначить полученный коэффициент разреженности матриц. Разреженность матрицы Фи (тем- термов) составила 0,86, а матрицы Тета (тем-документов) 0,0003. Таким образом, можно делать выводы о том, что отдельные темы описываются достаточно определённым уникальным набором слов/термов, документы же (в данном случае песни) не могут быть определены конкретной отдельно выбранной темой. Первый вывод дает нам основания для весьма устойчивой содержательной интерпретации получаемых тем и извлечения из них необходимой для исследования информации — тех идей и ценностей, которые транслируются в текстах русского рэпа.

Хочется обратить внимание, что помимо тех тем, что определяют некую идею или мысль, излагаемую в тексте, например, любовь или повествование про семью, существуют такие темы, которые не описывают содержания того текста, который характеризуют, и являются по сути набором неких текстовых атрибутов и стилистических приемов. Такими темами становятся «Мат» и «Природа». Описание текста через одну из этих тем не означает, что текст написан про мат или полностью из него состоит, равно как и это не означает, что текст повествует о любви к природе или ее описании. Тематическое моделирование порождает подобные темы, так как тексты до определённой степени описываются подобными темами. Например, текст будет описываться темой «Мат», если таковой в ней вообще встречается и будет описываться темой «Природа», если, например, заключает в себе строчки «Тихо, как падал снег, падал весь квартал мы падали на полпути во сне в поисках нала». Можно обратить внимание, что слово снег в данном случае используется лишь как вспомогательное и не определяет тематику всего текста.