УДК 82-1:004.8

ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ СОВРЕМЕННОЙ СЕТЕВОЙ ПОЭЗИИ

***А.Д. Никифорова*[1]**

[1]Никифорова Анна Дмитриевна – студент,

факультет инфокоммуникационных технологий,

e–mail: annanikiforova564@gmail.com

Национальный исследовательский университет ИТМО,

г. Санкт–Петербург, Российская Федерация

Аннотация: в данной статье с помощью тематического моделирования выделены темы современной сетевой поэзии. В качестве данных используются стихотворения, размещенные пользователями в социальной сети ВКонтакте. Стихотворения обрабатываются алгоритмом латентного размещения Дирихле (LDA). Полученные результаты могут быть использованы для дальнейшего анализа современной поэзии.

Ключевые слова: тематическое моделирование, современная поэзия, сетевой анализ, машинное обучение, LDA.

Введение

Поэзия – уникальная область гуманитарного знания. Она сочетает в себе коммуникативную составляющую – текст, и ритмическую, почти музыкальную подачу этого текста. В поэтических произведениях отражаются эмоции и чувства автора. Язык поэзии метафоричен и позволяет выразить субъективные переживания насчет окружающей действительности. По темам, на которые пишут поэты того или иного времени, можно судить о состоянии общества и событиях, в нем происходящих, а также о том, что волнует людей.

В отличие от поэзии классического периода (начало XIX в. - 1880-е гг.) и поэзии Серебряного века (рубеж XIX – XX вв.), современная русская поэзия рассматривается в работах не так часто. Причем в первую очередь исследователей интересует ее стилистическое своеобразие, а не содержание. Тем не менее, темы, поднимаемые в произведениях современных поэтов, могут дать пищу для размышлений касательно обстановки в обществе и человеческих ценностей. Этим и продиктована актуальность проводимого исследования.

Непредвзятым способом выделить темы можно назвать тематическое моделирование – используемый в машинном обучении подход к определению тем (топиков) для набора документов.

Целью данной работы является выделение тем, поднимаемых в современной поэзии, с помощью методов машинного обучения.

Алгоритм

Для решения поставленной задачи был выбран алгоритм латентного размещения Дирихле (LDA) [1]. Это алгоритм машинного обучения без учителя, в основе которого лежит кластеризация. Он основан на идее, что в схожих текстах будут встречаться схожие слова. Несмотря на то, что поэтические тексты метафоричны, поэтому результаты тематического моделирования для них будут отличаться от результатов тематического моделирования для прозаических текстов [2], работоспособность данного алгоритма продемонстрирована в статье On Poetic Topic Modeling: Extracting Themes and Motifs From a Corpus of Spanish Poetry [3].

Датасет

Существует исследование тем современной поэзии на корпусе текстов с сайта stihi.ru [4]. Нами же в качестве источника данных была выбрана российская социальная сеть ВКонтакте. Тому есть несколько причин. Во-первых, в данной социальной сети существуют открытые

тематические паблики, в том числе поэтические, в которых может публиковаться любой человек. Это гарантирует репрезентативность выборки. Кроме того, публикуемый материал не проходит жесткое рецензирование. Во-вторых, ВКонтакте предоставляет удобный способ извлечения своих данных с помощью API.

Для проведения исследования нами было выбрано десять различных пабликов, в которых публикуются современные поэты [5–14]. Они выбирались по следующим критериям: минимальное количество рекламных постов, минимальное содержание не стихотворных текстов, стандартизированная структура постов (допустим, сначала идет стихотворение, потом опциональна дата, затем имя автора), актуальность – последние публикации не раньше 2020 года.

Предполагалось собрать около 5000 последних постов из каждого паблика, однако не во всех из них столько было опубликовано на момент проведения исследования, поэтому пришлось внести небольшие коррективы (см. таблицу 1). Первоначальный датасет содержал 50440 постов. Из этой выборки ещё на этапе сбора данных были удалены рекламные посты.

Таблица 1 – Количество постов из пабликов ВКонтакте

Название паблика	Количество постов
Чай со вкусом коммунальной квартиры	9931
Написал я пару строк	4989
ВСЕ ПОЭТЫ ВКОНТАКТЕ	5000
СтиХозА • поэзия • проза	4835
хруСТАЛЬНЫЕ СТИХИ	4981
Лечебница	4981

Dum spiro...| Стихи, фото, арт	4995
СТИХИ	4709
Тысяча снежинок. Стихи.	5000
Бабочки на стене	1000

Прежде чем проводить дальнейшие манипуляции с датасетом, нужно было обработать его. Мы удалили хэштеги, а также последнюю строчку каждого текста. Это продиктовано тем, что в большинстве стихотворений последняя строчка содержит инициалы автора, ссылку на паблик автора или дату написания стихотворения. Так или иначе, она не несет особого смысла для тематического моделирования. Затем была проведена первичная предобработка текста [15]. Были удалены все символы не русского алфавита, т.е. пунктуация, латинские символы и цифры. Слова были приведены к нижнему регистру, а тексты токенизированы для удобства дальнейшей работы.

На первом графике (рис. 1) представлено распределение количества записей в зависимости от количества слов в них в изначальном датасете. Как видно по графику, имеется подозрительный пик около нуля. Он объясняется тем, что в выборку попали посты, содержащие только медиафайлы без текстовой информации, а также тексты, не являющиеся полноценными стихотворениями – например, цитаты. Посты, содержащие большое количество слов, тоже не внушают доверия. Скорее всего, они даже не являются стихотворениями в силу объема.

Было решено обрезать изначальный датасет. Нижняя и верхняя граница были выбраны как 30 и 200 слов соответственно. Размер получившегося датасета составил 34292 текста. Распределение текстов в нем можно увидеть на втором графике (рис. 2).

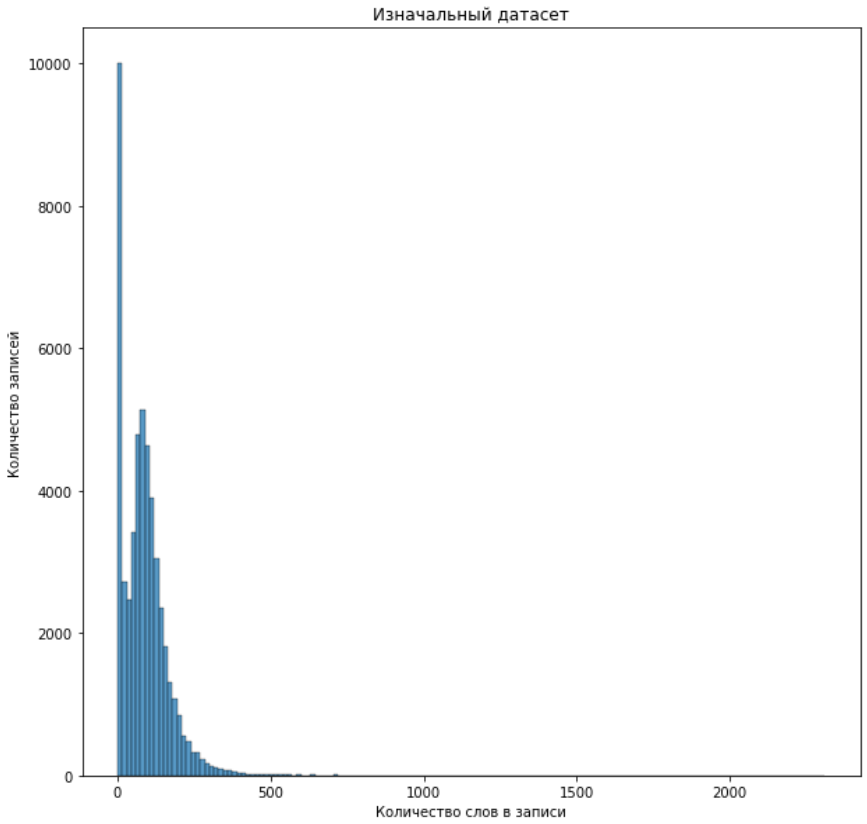

Рисунок 1 – График распределения количества записей в зависимости от количества слов в них в изначальном датасете

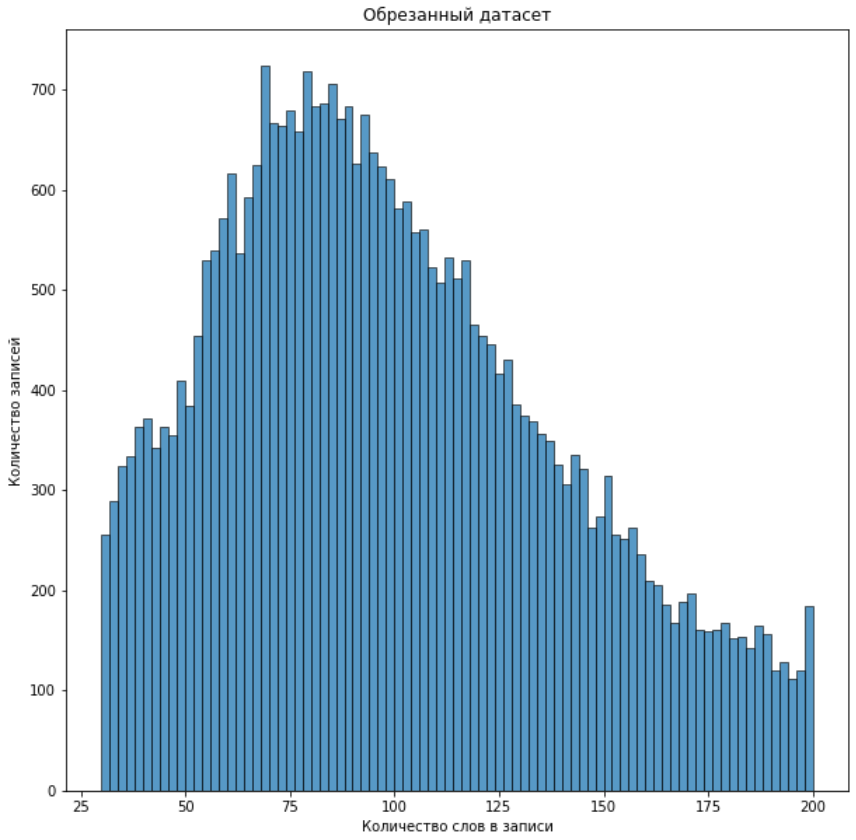

Рисунок 2 – График распределения количества записей в зависимости от количества слов в них в обрезанном датасете

Дальнейшая обработка текстов состояла в лемматизации (приведении слов к начальной форме) и удалении стоп–слов [15]. Для лемматизации была использована библиотека pymorphy2. Список стоп–слов был взят из библиотеки nltk и дополнен вручную.

Затем был создан словарь из 9639 уникальных токенов, а также корпус, представляющий собой мешок слов.

Выбор количества тем

Особенностью алгоритма LDA является то, что количество тем – это параметр, дающийся алгоритму изначально, а не настраиваемый в ходе обучения модели. Поэтому мы построили несколько моделей LDA, чтобы оценить, какое количество тем будет оптимальным. В качестве критерия использовалась когерентность (степень согласованности документов между собой) [16]. Результаты представлены на графике (рис. 3). Как

видно по графику, наиболее заметные пики приходятся на семь и двенадцать тем.

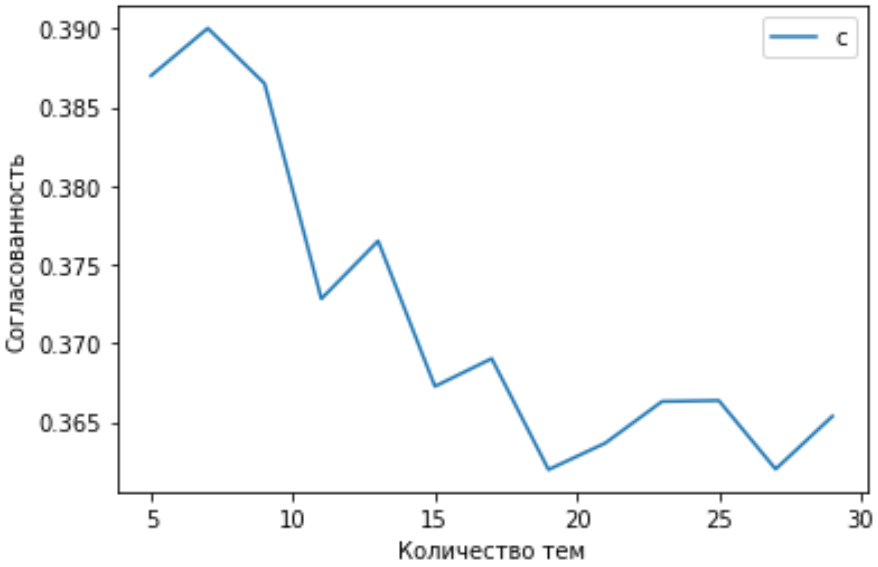

Рисунок 3 – График когерентности

Результаты

Топики, которые выделила модель LDA на семь тем, представлены в таблице 2. Топики, выделенные моделью на 12 тем, представлены в таблице 3.

Таблица 2 – Результаты LDA для семи тем

Авторская интерпретация темы	**Наиболее часто встречающиеся в теме слова**
Внутренний мир человека	боль губа слово тело кожа кровь палец взгляд внутри чувство мысль дышать пустой остаться лицо стена море стих грудь видеть голос имя помнить плечо голова сон стать нить память слеза рана последний ладонь горло волос воздух живой вода сила больной

Авторская интерпретация темы	**Наиболее часто встречающиеся в теме слова**
Счастье	счастье хотеть стать свет друг сон знать слово чувство мечта ночь боль слеза судьба рядом взгляд мысль улыбка любимый надежда нежный верить забыть ждать тепло небо милый встреча обнять сила радость огонь родный сказать просить смочь нежность звезда остаться миг
Природа	дождь снег осень окно город ветер весна лето небо зима солнце лист дом время белый вечер сон тёплый осенний утро ночь рассвет свет тепло улица идти снежный ждать цвет холод крыша мороз серый листва запах рисовать звезда дерево метель человек
Человек в мире	небо свет земля звезда ветер ночь море луна солнце путь вода птица река волна огонь тень дорога крыло чёрный край тьма сон облако берег человек лес песня петь видеть лететь трава белый поле идти камень живой закат стать гора тёмный
Домашняя обстановка. Быт	друг хотеть говорить ночь знать давать сказать девочка спать дом утро сидеть ждать чай кошка кофе окно помнить пить мальчик скучать человек стать вечер плакать дверь хотеться рассказать думать дело приходить любимый звонить нужный кот писать работа время новый забыть

Авторская интерпретация темы	**Наиболее часто встречающиеся в теме слова**
Жизненный путь. Поиск себя	человек знать друг стать слово время сказать верить хотеть говорить никто идти понять путь искать судьба писать нужный ждать новый бояться бывать остаться бог забыть сильный простой чувство видеть смочь сила дело простить пройти ответ дорога думать счастье просить стих
Война и родина	дом мама ребёнок бог идти война мать детство страна стать ждать человек отец старый смерть земля сын новый папа народ взрослый город нога вернуться сказка пёс родный русский последний детский стена стоить дорога господь назад бой россия сестра девчонка зверь

Таблица 3 – Результаты LDA для 12 тем

Тема	**Наиболее часто встречающиеся в теме слова**
(Несчастная) любовь	друг губа чувство слово сон знать боль тело видеть целовать море хотеть стих взгляд кожа дышать имя далее ночь остаться плечо нить стать забыть забывать мысль чувствовать внутри давно чужой память строчка пытаться сказать вино палец помнить клетка живой страсть

Внутренний мир человека	хотеть боль слово стать слеза чувство сила знать мысль обнять стих слышать просить радость внутри тело грудь голос улыбка пустота голова кричать ангел господь остаться рядом рана зло кровь взять песня нежность лёгкий сильный грусть эго дышать хотеться видеть больно
Городской пейзаж	город дождь время окно улица человек помнить стать писать лицо серый танцевать кофе стекло крыша стоить дом пустой говорить слово идти голова взгляд утро дверь прохожий чёрный дым осень капля мысль небо волос кот палец стучать танец лить нравиться грудь
Поэт и поэзия	поэт давать свет ночь чёрный ветер путь живой цвет небо человек стих деревянный остаться вино смерть последний луна дорога королева лёд земля маска вечность строчка знать лицо холодный посмотреть голова рисовать зелёный боль трава видеть далёкий красный кровь стать писать
Домашняя обстановка. Быт	друг говорить ждать человек ночь дом кошка давать чай скучать рассказать утро кофе новый вечер знать время хотеть спать хотеться приходить сказать нужный вместе пить стать стихотворение собака прийти вопрос стих остаться окно звонок вспоминать работа сидеть написать встреча одиночество

Размышления о судьбе	знать друг человек хотеть стать сказать слово верить счастье время забыть говорить нужный ждать понять судьба смочь чувство путь остаться никто искать простить уйти бояться бывать идти счастливый новый помнить пройти просить мысль рядом ответ мечта уходить думать хотеться сила
Война и родина	мама дом ребёнок детство ждать стать мать сказка бог страна знать отец родный старый война взрослый сын новый чудо народ сестра писать давно идти семья папа верить человек россия свет детский смерть окно время вернуться место русский дорога никто девчонка
Мечты и реальность	стена чёрный нога тело кожа песок говорить видеть кровь мальчик мысль камень волос лицо бог идти искать рот бросить море потолок корабль пол играть капитан живой стоять рухнуть палец крест вода остаться муза поэт место ладонь зеркало ад пустой круг
Семейная жизнь	женщина стать мужчина знать девочка красивый простить сказать любимый сильный хотеть ребёнок жена научиться счастье бояться стоить милый умный простой верить дело прекрасный просить немного сделать идти женский купить прийти страшный право улыбка дурак хотеться бог муж брать сила устать
Природа	снег осень ветер свет небо дождь весна сон солнце окно ночь счастье зима звезда тепло нежный лист белый лето тёплый рассвет мечта огонь время осенний губа стать вечер луч луна снежный дом

	холодный утро прийти яркий нежно нежность мороз взгляд
Жизненный путь. Поиск себя	человек бог свет небо путь земля судьба идти стать огонь боль сила крыло дорога слово страх время надежда искать тьма видеть шаг смерть знать слеза верить век вечный пустой мечта взгляд сон ночь мысль внутри живой птица война покой солнце
Природа	небо море ночь звезда ветер река солнце лес земля облако волна вода сон птица трава лето луна спать поле рассвет идти дерево дом окно синий берег лететь петь крыло нога край плакать песня крыша падать чёрный бежать туман дорога тень

Сопоставив и усреднив результаты двух LDA–моделей, мы получили следующие темы современной сетевой поэзии: (Несчастная) любовь, Внутренний мир человека, Жизненный путь и поиск себя, Размышления о судьбе, Поэт и поэзия, Природа, Война и родина, Семейная жизнь и быт, Городской пейзаж.

Наибольший интерес среди данных тем представляет городской пейзаж. Выделившись из более крупной темы природы (см. модель LDA на семь тем в сопоставлении с моделью на 12), он начинает фигурировать как отдельная тема. Получается, город становится новой составляющей лирического пейзажа современной поэзии.

Заключение

Мы провели исследование современной сетевой поэзии с точки зрения поднимаемых в ней тем с использованием алгоритма машинного обучения LDA. По итогу, было получено девять тем.

В дальнейшем планируется изучить, что представляет собой тема городского пейзажа, выделившаяся из темы природы, в чем ее

уникальность и какую роль играет город в произведениях современных поэтов.

Ссылки на источники/References

1. Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022. – URL: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf
2. Liza M. Rhody. Topic Modeling and Figurative Language. – URL: http://web.archive.org/web/20220331215147/http://journalofdigitalhumanities.org/2–1/topic–modeling–and–figurative–language–by–lisa–m–rhody/
3. Navarro–Colorado B (2018) On Poetic Topic Modeling: Extracting Themes and Motifs From a Corpus of Spanish Poetry. Front. Digit. Humanit. 5:15. doi: 10.3389/fdigh.2018.00015. – URL: https://www.frontiersin.org/articles/10.3389/fdigh.2018.00015/full
4. Тематическое моделирование корпуса текстов с сайта stihi.ru. – URL: http://poetrytopicmodeling.tilda.ws/
5. Чай со вкусом коммунальной квартиры [Электронный ресурс]. – URL: https://vk.com/public9073074
6. Написал я пару строк [Электронный ресурс]. – URL: https://vk.com/napisalya
7. Все поэты вконтакте [Электронный ресурс]. – https://vk.com/public49870925
8. СтиХозА • поэзия • проза [Электронный ресурс]. – URL: https://vk.com/public_stihoza
9. Хрустальные стихи [Электронный ресурс]. – URL: https://vk.com/public38087663
10. Лечебница [Электронный ресурс]. – URL: https://vk.com/h0spital
11. Dum spiro...| Стихи, фото, арт [Электронный ресурс]. – URL: https://vk.com/night.song

12. СТИХИ [Электронный ресурс]. – URL: https://vk.com/stihi_o_liubvi
13. Тысяча снежинок. Стихи [Электронный ресурс]. – URL: https://vk.com/public54602160
14. Бабочки на стене [Электронный ресурс]. – URL: https://vk.com/butterflies_wall
15. Тематическое моделирование жанров паралитературы. – URL: https://vk.com/@sysblok-kak-sdelat-tematicheskoe-modelirovanie
16. Derek O'Callaghan, Derek Greene, Joe Carthy, Pádraig Cunningham. An analysis of the coherence of descriptors in topic modeling. – URL: http://derekgreene.com/papers/ocallaghan15eswa.pdf