Всероссийский конкурс исследовательских и проектных работ школьников «Высший пилотаж»

Алгоритм автоматического деления текстов по литературным направлениям для поэзии Серебряного века[1]

Исследовательская работа

Направление *«Лингвистика»*

Автор: Львовский Яков Сергеевич

Учащийся 11 класса, ОАНО «Школа «Летово», Москва

2024 г.

[1] Автор благодарит В.А. Плунгяна за помощь в постановке задач исследования и ценные методологические указания, Г.А. Мороза за тщательное рецензирование работы и предложение путей её возможного улучшения.

Оглавление

Введение ... 3

Основная часть ... 3

Глава 1. Теоретические подходы и методы ... 3

1.1 Проблема определения литературного направления в филологии ... 3

1.2 Подход к решению проблемы на основе Digital humanities ... 4

1.3 Методы компьютерной лингвистики ... 6

1.3.1 Латентное размещение Дирихле ... 6

1.3.2 Дистрибутивная семантика ... 6

1.3.3 Алгоритм YAKE! для выделения ключевых слов ... 7

1.3.4 Наивный байесовский классификатор ... 8

Глава 2. Данные и алгоритм ... 9

2.1 Создание выборки текстов ... 9

2.2 Выделение тем текстов и ключевых слов ... 10

2.3 Итоговый алгоритм ... 10

Глава 3. Анализ полученных результатов ... 11

3.1 Сводная статистика по направлениям и авторам ... 11

3.2 Рейтинг успешности тем ... 12

3.3 Результаты работы наивного байесовского классификатора ... 14

Заключение ... 16

Список литературы: ... 17

Приложение А ... 18

Приложение Б ... 19

Введение

Темой нашего исследования является связь литературного направления и формальных особенностей текста. Существующие в классическом литературоведении определения особенностей направлений в высокой степени абстрактны, из-за чего литературное направление как термин характеризует скорее философско-идеологическую направленность автора, не говоря ничего о формальных особенностях его текстов. В своём исследовании нам хотелось проверить существование семантической связи внутри текстов представителей одного направления, используя методы обработки естественного языка (в англоязычной литературе принят термин Natural Language Processing, далее в этой работе мы будем использовать аббревиатуру NLP). Эта тема является актуальной, поскольку результаты исследования могут помочь уточнить терминологические границы понятия «литературного направления». Также исследование имеет потенциал для решения задач корпусной лингвистики, в частности, создания автоматической (или полуавтоматической) разметки текстов по литературным направлениям, что существенно упрощает процессы поиска и отбора текстов для филолога-исследователя.

Объектом исследования являются тексты русских поэтов начала XX века. Предметом исследования является связь между принадлежностью текста к литературному направлению и его семантическим содержанием. Целью исследования является разработка алгоритма, который определяет принадлежность текста к определённому литературному направлению.

Гипотеза исследования заключается в том, что выделенные семантические признаки позволят создать алгоритм с сравнительно высокой (больше 50%) точностью, однако не будут являться дискриминирующими в ста процентах случаев.

Основная часть

Глава 1. Теоретические подходы и методы

1.1 Проблема определения литературного направления в филологии

При проведении пилотного исследования, в котором разметка текстов по литературному направлению осуществлялась вручную, мы столкнулись с достаточно серьёзной проблемой – существующая классификация текстов по литературным направлениям опирается в первую очередь на идеологические установки автора текста и его формальную принадлежность к определённым творческим направлениям, практически не связывая литературное направление с особенностями непосредственно текста. Нельзя сказать, что такие связи не проводятся вовсе, однако они по большей части абстрактны. Так, в случае с символизмом говорится в первую очередь о наличии в тексте так называемых «образов-символов» - «многозначных иносказательных выражений скрытого смысла произведения» (зачастую смыслы символу

придаются через культурный контекст, а также контекст творчества автора в целом) [Горшкова et. al. 2018, Барковская 1999]. Связанной с ними оказывается идея многомирия: представленности в стихотворении сразу нескольких «миров» (например, рая и ада), от которых и исходят «образы-символы» (по сути, обогащаясь значениями) [Горшкова et. al. 2018, Барковская 1999, Сухих 2018]. Из многомирия протекает и тематическое отличие поэзии символизма – большой акцент на мистическом [Горшкова et. al., Барковская 1999, Сухих 2018]. Пожалуй, из связанных с языком тенденций можно выделить только любовь символистов к прилагательным[2] [Барковская 1999].

Приписываемые акмеистам характеристики ещё абстрактнее. И исследователи, и сами поэты указывают в первую очередь на уход от сложной метафорики символизма, возвращение образу его «вещной» стороны, конкретности [Горшкова et. al. 2018, Барковская 1999, Сухих 2018]. Достаточно точно этот принцип сформулирован в статье М. Кузмина «О прекрасной ясности», где он поясняется как «соответствие данной формы с известным содержанием и приличествующим ей языком» [Барковская 1999]. Исследований, освещающих лингвистическую сторону поэзии акмеистов, нам найти не удалось.

Из всех исследуемых нами направлений, пожалуй, именно футуризм ассоциируется с конкретной лингвистической особенностью текста – обилием неологизмов [Горшкова et. al. 2018, Сухих 2018]. Некоторые исследователи говорят и о тематической общности текстов футуристов – любви к «урбанизму», обращению к образам современных механизмов и города вообще [Горшкова et. al.]. Оба этих различия проистекают из философии футуризма – протестности, отрицания старых эстетических норм, перехода к «новому искусству» [Горшкова et. al., Барковская 1999, Савченко 2021, Сухих 2018].

Пожалуй, подобное рассуждение можно применить и к системе деления на литературные направления в целом – формулировки общепризнанных различий в поэтике абстрактны в первую очередь потому, что опираются скорее на философию автора, его подход к поэзии. А соответственно задачей нашего исследования становится поиск конкретных языковых проявлений этой философии, возможное уточнение формулировок и исследование того насколько сильна сама связь между философией поэта и его текстом.

1.2 Подход к решению проблемы на основе Digital humanities

Цифровые гуманитарные науки (в англоязычной литературе Digital humanities. Здесь и далее будем использовать сокращение DH) являются достаточно молодым направлением в гуманитарной науке. Характеризуются они в первую очередь *междисциплинарностью*

[2] Далее в исследовании мы не будем возвращаться к этой черте символистов, но отметим, что анализ нашей выборки опровергает эту гипотезу.

проводимых исследований. Большинство DH исследований строятся вокруг комбинации современных автоматических решений для анализа и сбора данных, с традиционными для гуманитарных наук техниками интерпретации.

В рамках подхода DH нам удалось найти немного исследований, ставящих своей целью анализ литературных направлений. Примечательной показалась работа физиков из университета Сан-Карлоса в Бразилии, использоваюших преобразование текстов в так называемые “complex networks” (сложные сети), структуры, хранящие в себе информацию о частотности и совместной встречаемости слов [Amancio et. Al. 2012]. Анализируя различные метрики, применяемые к таким сетям (не будем останавливаться на них подробно), исследователи кластеризуют полученные результаты. Объединившиеся в кластеры тексты действительно в высокой степени коррелировали с принадлежностью текстов к одному литературному направлению. Наш подход, однако отличается тем, что алгоритм строится на выявлении особенностей текстов, принадлежность которых к определённому направлению однозначна.

В рамках нашего исследования отдельного внимания заслуживают работы по цифровому стиховедению (в особенности те из них, которые построены вокруг методов семантического анализа). Многие исследования с подобной методологией построены вокруг такой теоретической проблемы, как *семантический ореол метра*, стиховедческой концепции, согласно которой стихотворный метр со временем накапливает определённый набор ассоциированных с ним тем и образов [Piperski 2017, Šeļa et. al. 2020].

Первым исследованием, о котором нам хотелось бы сказать в этом разделе, будет исследование А.Ч. Пиперски [Piperski 2017]. В данном исследовании к проблеме семантического ореола метра подходят путём выделения ключевых слов, анализируя затем близость выделенных слов с открытыми классическими литературоведами (так, например, для пятистопного хорея, традиционно ассоциируемого с мотивами пути, жизни и ночи, среди ключевых слов оказались «жить», «ночь» «путь», «уходить»).

Ещё больший интерес для нашей работы представляют исследования, использовавшие как методологическую основу более продвинутые методы NLP. Так, чешские стиховеды П. Плехач и Р. Колар использовали для изучения семантического ореола метра чешских поэтов пост-символистов один из наиболее популярных алгоритмов тематического моделирования - Latent Dirichlet Allocation (далее LDA). Создав с помощью LDA список тем (связанных в текстах кластеров слов, более подробно об этом будет сказано позже), Плехач и Колар смотрят на корреляции тем с стихотворным размером и поколением авторов (наблюдая таким образом не

только за самим ореолом метра, но и за динамикой его развития) [Plecháč et. al. 2022]. Подобное исследование проводили и для русского языка Б.В. Орехов, А. Шеля и Р. Лейбов [Šeļa et. al. 2020].

Кратко рассмотрев уже проводимые стиховедами исследования, мы можем прийти к выводу, что изучение семантики стихотворного текста методами компьютерной лингвистики позволяет уточнять и переинтерпретировать уже существующие литературоведческие работы.

1.3 Методы компьютерной лингвистики

В этом разделе мы предоставим краткий обзор основных методов и концепций NLP, использованных в нашем исследовании.

1.3.1 Латентное размещение Дирихле

Латентное размещение Дирихле является одним из самых популярных алгоритмов для решения задачи тематического моделирования, т.е. выделения основных тем, содержащихся в тексте. Не будем углубляться в математику, стоящую за алгоритмом, но отметим основные идеи LDA [11]:

- Тема является *вероятностным распределением слов* (т.е. набором слов, каждому из которых приписана вероятность, что оно встретится в тексте с данной темой).
- Одному тексту может соответствовать несколько тем. Самому документу соответствует вероятностное распределение всех тем корпуса.
- Генерация тем и их распределений в документе происходит на основе *терм-документной матрицы* (по сути таблицы, в которой для каждого документа корпуса указана встречаемость в нём каждого уникального слова содержащегося в корпусе).

1.3.2 Дистрибутивная семантика

Важную роль в получившемся у нас в итоге алгоритме играют дистрибутивно-семантические модели (в частности модель Rusvectores).

Основная идея при построении дистрибутивно-семантических моделей (так называемая дистрибутивная гипотеза) заключается в следующем: часто встречающиеся в похожих контекстах лексические единицы обладают схожим значением [Turney et. al. 2010]. Эта идея позволяет строить математические модели, способные:

- Вычислять семантическую близость слов.
- Выполнять простейшие алгебраические операции со словами (так, например, всё та же модель Rusvectores при вычитании из слова «королева» слова «женщина» выдаёт как ближайшее семантически слово «король»).
- Вычислять семантические ассоциаты слова (т.е. набор слов, лексическое значение которых ближе всего к исследуемому слову).

Подобные модели обучаются на больших корпусах текстов, извлекая из них информацию о совместной встречаемости слов и приписывая каждому слову собственный вектор.

В нашем исследовании мы опирались на разработанную Андреем Кузнецовым (университет Осло) и Елизаветой Кузьменко (университет Гроннингена) модель Rusvectores, обученную на Национальном корпусе русского языка и русскоязычных статьях портала Wikipedia [Kutuzov et. al. 2017]. Интересной особенностью этой модели является то, что помимо самого слова в качестве входной информации она принимает и его часть речи. Это сделано для уменьшения случаев некорректной работы модели ввиду морфологической омонимии (ситуации совпадения форм слов, обладающих разными значениями). Простейшим примером морфологической омонимии может послужить слово «печь», которое может означать как существительное, так и глагол.

1.3.3 Алгоритм YAKE! для выделения ключевых слов

Для выделения ключевых слов мы использовали разработанный португальскими исследователями алгоритм YAKE! (Yet Another Keyword Extractor) [Campos et. al. 2019]. Его главное отличие от метода, используемого в исследовании Пиперски (подробно обсуждается в разделе 1.2), заключается в том, что он опирается на статистические характеристики терминов внутри документа, тогда как Пиперски использует метод, построенный на сравнении частотного словаря документа с частотным словарём крупного корпуса текстов на исследуемом языке (так называемого reference corpus). Отметим важные факторы, используемые в алгоритме YAKE! для оценки важности слова:

- Регистр термина (T_{Case}). Основываясь на предположении, что термины, начинающиеся с большой буквы и не находящиеся в начале предложения, являются более важными по сравнению с остальными терминами, авторы алгоритма вводят метрику, сопоставляющую число раз, когда термин начинался с большой буквы или являлся акронимом с его частотностью в целом.
- Позиция термина ($T_{Position}$). Эта метрика показывает, как частотность термина меняется в зависимости от его близости к началу текста (замечено, что в научных и новостных статьях наиболее важные термины сконцентрированы именно в начале текста).
- Нормализованная частотность термина (TF_{Norm}). Для более точной оценки частотности она делится на сумму средней частотности всех терминов документа и стандартного отклонения. Эта процедура по своей сути аналогична удалению стоп-слов (так как также решает проблему слишком высокой частотности терминов, которые характерны для всех текстов на данном языке), однако авторы алгоритма экспериментально показывают большую эффективность выбранного ими подхода.

- Релевантность термина к контексту (T_{Rel}). Авторы исходят из предположения, что чем более разнообразны слова, встречающиеся в левом и правом контексте термина, тем менее он значим (поскольку не обладает специфичным контекстом). Для решения этой проблемы и вводится метрика T_{Rel}.
- Доля уникальных предложений ($T_{Sentence}$). Деля число предложений, в которых встречается термин, на общее число предложений документа, авторы исходят из предположения, что важность термина коррелирует с процентом предложений, в которых он встречается.

Ещё одной важной чертой YAKE! является то, что в качестве ключевых слов выделяются в том числе n-граммы (последовательности из нескольких слов). Общая важность n-граммы складывается из важности всех встречающихся в ней слов. Последним шагом алгоритма является сопоставление всех ключевых слов и удаление похожих пар (степень их похожести определяется разными способами, основанными главным образом на числе единиц, которые надо заменить для превращения одной n-граммы в другую).

Стоит сказать, что в случае с исследованием стихотворных текстов выбор YAKE! может быть расценен как не самое удачное решение, так как сам алгоритм создан скорее для обработки достаточно крупных текстов. Отметим, однако, что его важным преимуществом по сравнению с более классическими методами выделения ключевых слов является то, что он не требует наличия reference corpus (а в случае с поэтическими текстами конца XIX – начала XX века потребовалось бы создание частотного словаря текстов именно этого периода, что потребовало бы достаточно серьёзных затрат по времени).

1.3.4 Наивный байесовский классификатор

Последним важным методом NLP, который нам необходимо осветить, является наивный байесовский классификатор. Этот классический алгоритм машинного обучения отлично подходит для решения задачи классификации (т.е. распределения документов по уже известным классам, исходя из обучающей выборки, в которой каждому тексту уже приписан класс). Не вдаваясь в математические тонкости, отметим только, что при вычислении вероятности принадлежности слова к классу наивный байесовский классификатор рассматривает два ключевых параметра [10]:

- Доля документов, относящихся к определяемому классу из всех документов корпуса.
- Сумма (в некоторых вариациях формулы произведение) долей каждого слова документа среди всех слов документов, относящихся к определяемому классу.

Соответственно документ считается относящимся к тому классу, вероятность принадлежности к которому выше.

В нашем исследовании мы использовали этот алгоритм для определения направления стихотворения (соответственно с классами «акмеизм», «символизм» и «футуризм»), разделив всю выборку на обучающую и тестовую в отношении 66% к 34% соответственно.

Глава 2. Данные и алгоритм

2.1 Создание выборки текстов

Для проведения исследования была создана выборка общим объёмом 332 текста и 32810 словоформ. В выборку вошли следующие авторы:

- Символисты: В.Я. Брюсов, К.Д. Бальмонт, Д.С. Мережковский, А.А. Блок, Н.М. Минский, З.Н. Гиппиус, Ф.К. Сологуб.
- Акмеисты: Н.С. Гумилёв, А.А. Ахматова, О.Э. Мандельштам, Г.В. Иванов.
- Футуристы: В.В. Хлебников, В.В. Маяковский, А.Е. Кручёных, В.В. Каменский, Е.Г. Гуро.

Выбор авторов был основан на изучении специализированной литературы и консультации со специалистами.

От каждого автора взято приблизительно 10-15 текстов. Тексты выбирались случайно.

Общую таблицу по объёму корпуса каждого направления можно представить следующим образом:

Таблица 1. Распределение числа текстов и числа словоформ по направлениям

	Число текстов	Число словоформ
Символисты	130	14605
Акмеисты	100	9206
Футуристы	99	8998

Выбраны только тексты с годом создания от 1880 до 1919.

Сводную таблицу по авторам можно посмотреть в приложении Б.

Средняя длина текста – 100 словоформ.

Источником текстов послужил поэтический подкорпус Национального корпуса русского языка (https://ruscorpora.ru/new/search-poetic.html).

Малый размер выборки мог исказить результаты, которые мы представим в данном исследовании. В дальнейшем планируется проведение более подробных исследований с расширенной выборкой.

2.2 Выделение тем текстов и ключевых слов

Для выделения тем, содержащихся в поэтических текстах, мы использовали разработанный Дэвидом Мимно (исследователь из департамента наук об информации Корнелльского университета) веб-инструмент для тематического моделирования (https://mimno.infosci.cornell.edu/jsLDA/). Используемый в нём алгоритм идентичен используемому в более популярном инструменте MALLET и принципиально не отличается от стандартных алгоритмов LDA.

Важным шагом перед выделением тем являлась предобработка поэтических текстов. Она состояла из двух этапов: удаление стоп-слов и лемматизация. Под удалением стоп-слов мы понимаем достаточно стандартную для NLP процедуру – очистку текста от слов, высокая частотность которых характерна для большинства текстов на языке, но не зависит от конкретного текста (в случае с русским языков это большинство служебных слов). В нашем исследовании мы использовали список стоп-слов из библиотеки nltk (Natural Language Toolkit) языка Python, с ним можно ознакомиться в приложении А. Лемматизацией же мы называем процесс приведения всех слов текста в начальную форму. Её мы осуществили с помощью модуля pymorphy2 всё того же языка Python. Проведение предобработки важно, поскольку увеличивает точность работы алгоритма, предоставляя ему более полную информацию о частотности и совместной встречаемости слов.

Корпус предобработанных текстов от одного направления загружался в веб-инструмент, после чего алгоритм выделял в нём 50 тем, повторяя этот процесс 50 раз (за счёт большого числа повторений достигается большая точность в выделении тем). Затем темы выгружались и переводились в формат, необходимый для работы с моделью Rusvectores (каждому слову приписывалась его часть речи).

Выделение ключевых слов мы проводили по алгоритму YAKE! используя разработанный авторами алгоритма пакет yake языка Python.

2.3 Итоговый алгоритм

После проведения всех вышеозначенных шагов исследования был разработан итоговый алгоритм:

1. Для каждого стихотворения выделялись ключевые слова по алгоритму YAKE!.
2. Ключевые слова стихотворения сопоставлялись с темой: находилась средняя близость каждого ключевого слова с каждым словом темы. Это значение и считалось значением близости темы и стихотворения.

3. Процесс вычисления близости темы и стихотворения проводился для всех тем каждого направления.
4. Находилась среднее от тем по каждому направлению. Это значение считалось значением близости текста к направлению.
5. Тексту приписывалось направление, с которым у него получалось наибольшее значение близости.

Подробная оценка результатов работы алгоритма приведена в следующей главе.

Глава 3. Анализ полученных результатов

3.1 Сводная статистика по направлениям и авторам

Итоговый алгоритм правильно распознал направление у 188 текстов из 332 (56.6% случаев). Число распознанных текстов по направлению представлено в этой таблице:

Таблица 2. Процент распознанных текстов по направлениям

Направление	**Процент распознанных текстов**
Футуризм	47% (47 текстов из 100)
Символизм	58% (75 текстов из 121)
Акмеизм	57% (58 текстов из 101)

Также определённый интерес представляет статистика по числу распознанных текстов у каждого автора.

Таблица 3. Процент распознанных текстов по авторам

Автор	**Процент распознанных текстов**	**Направление**
Ф.К. Сологуб	60%	Символизм
Н.М. Минский	40%	Символизм
В.В. Маяковский	58%	Футуризм
Г.В. Иванов	70%	Акмеизм
В.В. Хлебников	32%	Футуризм
А.А. Блок	60%	Символизм
А.А. Ахматова	46%	Акмеизм
В.Я. Брюсов	61%	Символизм
А.Е. Кручёных	70%	Футуризм
Н.С. Гумилёв	58%	Акмеизм

Е.Г. Гуро	33%	Футуризм
В.В. Каменский	33%	Футуризм
О.Э. Мандельштам	58%	Акмеизм
Д.С. Мережковский	78%	Символизм
К.Д. Бальмонт	34%	Символизм
З.Н. Гиппиус	70%	Символизм

3.2 Рейтинг успешности тем

Чтобы проверить гипотезы филологов о тематических различиях между текстами различных направлений, нами был составлен так называемый *рейтинг успешности тем*. Суть заключается в том, что для тем каждого направления мы посчитали среднюю близость между каждой темой и каждым текстом выбранного направления. Отсортировав темы по этому параметру, мы и получили рейтинг. Далее мы попробуем сопоставить наиболее «успешные» темы каждого направления с тематическими характеристиками, которые выделяются в филологии.

Начнём с символизма. Самые «успешные» темы, выделенные из текстов символистов, выглядят следующим образом:

1. Свет, мгла, туман, черный, мерцание, лицо, молитва, мечта, холодный, туча. Степень близости: 0.109.

2. Мертвый, снег, солнце, тяжелый, холод, покой, тусклый, белизна, луч, сухой. Степень близости: 0.104.

3. День, душа, печалить, свой, тоска, весь, звезда, бесконечный, мир, святой. Степень близости: 0.103.

4. Огонь, радость, цвет, пред, жизнь, таинственный, легкий, смех, нагой, птица. Степень близости: 0.101.

5. Трава, сердце, небо, голос, природа, воспоминание, еще, толпа, бездна, друг. Степень близости: 0.099.

Следующие черты тем могут быть соотнесены с особенностями текстов символистов:

1. Наличие антонимических (хотя бы частично) пар внутри одной темы (например: свет-мгла, трава-небо, небо-бездна). Такая особенность может быть соотнесена с характерным для символистов *двоемирием*.

2. Наличие в темах слов, связанных с миром, отдалённым от земного (например: бесконечный, небо, солнце, звезда, молитва). Это можно интерпретировать и как проявление всё того же двоемирия, и как связь с любовью символистов к мистическим темам.
3. Наличие слов, связанных с *уходом от реальности* и сугубо абстрактными понятиями (мечта, воспоминание). Также говорит о меньшем акценте на объектах окружающего мира (в отличие от того же акмеизма).
4. Наличие слов, связанных с мистическим содержанием (таинственный, душа, святой).

Общая абстрактность слов, составляющих темы символистов (конкретные аспекты раскрыты в списке выше) позволяет нам говорить о том, что тематическое моделирование косвенно подтверждает гипотезы литературоведов о признаках общности символистских текстов.

Проведём аналогичную процедуру и для акмеистов. Их самые «успешные» темы выглядят так:

1. Царица, уста, птица, казаться, плакать, глаз, невидимый, хитон, лес, рыдать. Степень близости: 0.097.

2. Тень, закат, сон, цветок, прежде, это, гордый, сумрачный, свет, планета. Степень близости: 0.093.

3. Умирать, что-то, скоро, тишина, смеяться, дрожать, стена, жгучий, печаль, страстный. Степень близости: 0.090.

4. Слеза, жена, чертог, шептать, усталый, любить, гора, покинуть, безмолвный, алый. Степень близости: 0.088.

5. Черный, рот, видеть, тонкий, темный, крик, ворота, рука, серый, смерть. Степень близости: 0.087.

Сравнивая эти темы с темами символистов, можно прийти к нескольким важным выводам.

Во-первых, идея о большем акценте акмеистов на «вещности» находит в этих темах своё подтверждение, на что указывает меньшее (по сравнению с символистами) присутствие в темах абстрактной лексики.

Во-вторых, некоторые черты, которые можно было бы интерпретировать как характеризующие символистов (например, присутствие лексики, ассоциированной со смертью, что могло бы рассматриваться как связь с тем же двоемирием) на самом деле таковыми скорее не являются, поскольку приблизительно в таком же объёме присутствуют в темах акмеистов.

Такое рассмотрение тем акмеистов и символистов достаточно хорошо стыкуется с результатами работы модели. Несмотря на то, что и для символизма, и для акмеизма, число правильно распознанных текстов стремится к 60%, значительную часть из них алгоритм не различает (37% текстов символистов были распознаны как акмеизм, 35% текстов акмеистов – как символизм). Так и различия, которые видны при сравнении тем, пусть и заметны, но не позволяют установить явную границу между направлениями.

Рассмотрим, наконец, темы футуристов:

1. Небо, бог, падать, собака, труп, красный, ваш, крик, река, пасть. Степень близости: 0.078.

2. Небо, петься, дом, роза, видеть, кричать, фонарь, восток, луна, спать. Степень близости: 0.072.

3. Взор, кровь, хохотать, пинь, родный, цепь, трава, выходить, падать, вечный. Степень близости: 0.072.

4. Играть, сиять, вечер, что-то, бубенец, белый, смеяться, золото, коза, локоть. Степень близости: 0.071.

5. Дым, песня, солнце, улица, бросать, рог, иной, холод, туча, скакать. Степень близости: 0.0713.

Их интерпретация, пожалуй, представляется нам наиболее сложной по следующей причине: критерии для отделения футуризма от остальных направлений являются в наименьшей степени тематическими. Пожалуй, выделенные темы могли бы указать разве что на важность неологизмов в поэзии футуристов. Однако в 5 наиболее «успешных» темах не наблюдается ни одного неологизма (разве что, возможно, «пинь»). Сопоставляя с другими направлениями, можно, пожалуй, сказать, что ближе всего эти темы к представлению о «вещности» акмеистов (что подтверждается и статистикой, 32% текстов футуристов были распознаны как акмеизм). Практически не удаётся проследить в этих темах и любовь футуристов к теме города и образам «механического» (можно попробовать так интерпретировать наличие слов «дом», «улица», «фонарь» и возможно «дым», но этого явно недостаточно чтобы говорить о тематической общности). Таким образом, гипотеза об отсутствии тематической общности в текстах футуристов также подтверждается результатами работа алгоритма и анализом тем.

3.3 Результаты работы наивного байесовского классификатора

Чтобы ещё точнее оценить результаты нашего исследования и придумать возможные линии его развития, мы провели сравнение результатов работы нашего алгоритма с наивным байесовским классификатором. В таблице ниже представлены результаты его работы:

Таблица 4. Результаты работы наивного байесовского классификатора на всей выборке

	Precision	Recall	F1-score
Акмеизм	0.38	0.45	0.41
Символизм	0.58	0.73	0.65
Футуризм	0.85	0.47	0.61

Здесь и в таблицах 5-6: precision – доля верно распознанных текстов среди всех, recall – доля текстов, которые были распознаны неверно, а F1-score – их среднее гармоническое.

Заметно, что наиболее хорошо распознанным направлением оказался футуризм (что ровно противоположно результатам работы нашего алгоритма). Поскольку выбранный нами метод векторизации (count vectorizer) опирается в первую очередь на частотность терминов в корпусе, можно сделать предположение, что ключевое отличие футуристов лежит не в области семантики стихотворений, но в том, какие пласты лексики являются основными для представителей футуризма (заметим, что к этому же выводу можно прийти, посмотрев на процент пересекающейся лексики в темах футуристов, акмеистов и символистов). Эту гипотезу можно связать с *протестностью* текстов футуристов, которая могла выражаться в выборе лексики из тех пластов, которыми как правило не пользовались более классические поэты.

В поисках более конкретного объяснения, мы решили проверить гипотезу о том, что настолько высокая точность распознавания футуристов связана с большим числом неологизмов в текстах Велимира Хлебникова. Чтобы протестировать это предположение мы применили наивный байесовский классификатор к выборке без текстов Хлебникова. Результат, однако, нас удивил.

Таблица 5. Результаты работы наивного байесовского классификатора на выборке без текстов Хлебникова

	Precision	Recall	F1-score
Акмеизм	0.38	0.45	0.41
Символизм	0.67	0.87	0.75

Футуризм	1.00	0.14	0.24

В поисках объяснения мы решили повторить эксперимент, но на этот раз убрали тексты Кручёных (так как наш алгоритм распознавал его тексты точнее, чем тексты других футуристов).

Таблица 6. Результаты работы наивного байесовского классификатора на выборке без текстов Кручёных

	Precision	Recall	F1-score
Акмеизм	0.48	0.43	0.46
Символизм	0.64	0.77	0.70
Футуризм	0.59	0.43	0.50

Как видно из таблицы, отсутствие текстов Кручёных значительно снижает распознаваемость футуристов (но зато позволяет классификатору точнее определять акмеистов). Можно сделать предположение, что конкретно его текстам присуща наиболее специфичная лексика, тогда как у остальных футуристов различия с другими направлениями не так сильны.

Заключение

Наше исследование показало, что деление на литературные направления не связано напрямую с семантическим наполнением текста, а значит, гипотеза о том, что оно носит в первую очередь исторический характер (т.е. зависит от философских взглядов автора и его принадлежности к определённым организациям), подтвердилась. Однако работа наивного байесовского классификатора показала, что тексты футуристов могут быть с достаточно высокой точностью отделены от текстов символистов и акмеистов. Наша гипотеза заключается в том, что это связано с выбором футуристами лексики слабо характерной для поэтического языка начала XX века. В дальнейшем представляет интерес повторение нашего исследования на выборке большего масштаба, а также исследование лексики в поэзии футуристов путём сопоставления с частотным словарём поэзии начала XX века и статистического анализа результатов.

Код самого алгоритма и данные, использовавшиеся в исследовании, можно посмотреть в репозитории на платформе Github.

Список литературы:

1. Горшкова Н.Д., Малинина М.Г., Чурляева Т.Н. Поэзия Серебряного века: учебное пособие . - Новосибирск: Издательство НГТУ, 2018. - 39 с.
2. Барковская Н.В. Поэзия “серебряного века”. Учебное пособие. - 2-е изд. - Екатеринбург: Урал. гос. пед. ин-т. Екатеринбург, 1999. - 170 с.
3. Савченко Т.К. Поэзия Серебряного века как системный феномен // Словесное искусство Серебряного века и русского зарубежья в контексте эпохи («IV смирновские чтения»), материалы IV Международной научной конференции. - 2021. - С. 181-193.
4. Сухих И.Н. Русская литература для всех: От Блока до Бродского . - Москва: ИГ "Лениздат", 2018. - 736 с.
5. Piperski A. Ch. Semantic halo of a meter: a keyword-based approach // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2017” . – 2017
6. Šeļa A., Orekhov B., Leibov R. Weak Genres: Modeling Association Between Poetic Meter and Meaning in Russian Poetry // CHR 2020: Workshop on Computational Humanities Research. - Amsterdam: 2020
7. Plecháč P., Kolár R. Metre and Semantics in the Poetry of Czech PostSymbolists Accessed via LDA Topic Modelling // Atudia Metrica et Poetica. - 2022. - №9.1. - С. 7–19.
8. Kutuzov A., Kuzmenko E. WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models // Social Networks and Texts: 5th International Conference, AIST 2016. - Yekaterinburg: Springer International Publishing AG, 2016. - С. 155-161.
9. R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, A. Jatowt YAKE! Keyword extraction from single documents using multiple local features. // Information Sciences. - №509. - С. 257-289.
10. “Наивный Байес, или о том, как математика позволяет фильтровать спам.” // Хабр URL: https://habr.com/ru/articles/415963/ (дата обращения: 24.01.2024).
11. “Вероятностные модели: LDA, часть 2.” // Хабр URL: https://habr.com/ru/companies/surfingbird/articles/230103/. (дата обращения: 24.01.2024).
12. Turney P.D., Pantel P. From Frequency to Meaning: Vector Space Models of Semantics // Journal of Artificial Intelligence Research . - 2010. - №37. - С. 141-188.
13. D. R. Amancio, O. N Oliveira Jr, L. F. Costa Identification of literary movements using complex networks to represent texts // New Journal of Physics. - 2012. - №14. - С. 1-15.

Приложение А

Список стоп-слов библиотеки nltk языка Python: вас, на, когда, тот, хорошо, потому, ни, наконец, все, есть, надо, во, при, разве, того, чтоб, себе, раз, о, опять, три, эти, у, от, тогда, том, ты, над, больше, как, будет, ж, можно, хоть, теперь, тут, один, вам, может, она, из, про, этой, даже, иногда, ему, если, об, них, вот, нибудь, всю, он, какая, лучше, между, здесь, меня, до, этом, тоже, ведь, или, там, они, два, быть, куда, под, себя, чем, уж, тебя, всех, был, тем, с, чуть, нет, такой, никогда, ну, более, будто, его, а, совсем, почти, где, сейчас, эту, него, всегда, им, но, свою, через, моя, чего, много, чтобы, ним, мы, к, сам, я, какой, со, да, по, было, конечно, не, была, же, всего, ей, за, зачем, кто, что, еще, для, без, впрочем, только, потом, так, ее, и, их, то, в, ничего, бы, нее, уже, вдруг, после, другой, были, ней, мой, нельзя, перед, ли, этот, мне, нас, этого, вы.

Приложение Б

Распределение числа текстов и числа словоформ по авторам (среди текстов выборки)

Автор	Число текстов	Число словоформ
А.А. Ахматова	26	1792
А.А. Блок	20	1789
А.Е. Кручёных	27	1415
В.В. Каменский	9	842
В.В. Маяковский	17	2288
В.В. Хлебников	37	3798
В.Я. Брюсов	26	2225
Г.В. Иванов	20	1555
Д.С. Мережковский	28	3325
Е.Г. Гуро	9	655
З.Н. Гиппиус	10	1355
К.Д. Бальмонт	26	3595
Н.М. Минский	10	1266
Н.С. Гумилёв	27	4110
О.Э. Мандельштам	27	1749

Ф.К. Сологуб	10	1051