УДК 004.912

АНАЛИЗ СТИЛЯ РУССКОЯЗЫЧНЫХ ТЕКСТОВ С ИСПОЛЬЗОВАНИЕМ РИТМИЧЕСКИХ ХАРАКТЕРИСТИК

Лагутина Ксения Владимировна, к.т.н., ассистент
Лагутина Надежда Станиславовна, к. ф.-м. н., доцент
Ярославский государственный университет им. П.Г. Демидова
Lagutina Ksenia, PhD, P.G., lagutinakv@mail.ru
Lagutina Nadezhda, PhD, P.G., lagutinans@rambler.ru
Demidov Yaroslavl State University

В работе рассматривается кластеризация русскоязычных текстов разных жанров на основе ритмических параметров, основанных на повторении слов и фраз. Кластеризация характеристических векторов и визуализация результатов осуществляется с помощью алгоритма итар.

Ключевые слова: *автоматическая обработка текста, стилометрия, кластеризация, итар.*

Компьютерная лингвистика, известная также как математическая и вычислительная лингвистика, появилась на стыке лингвистики, информатики, математики и искусственного интеллекта. Первые исследования по машинному переводу, созданию электронных словарей и тезаурусов, разработке методов и алгоритмов морфологического анализа лексики перешли к работам по синтаксическому и семантическому анализу текстов, анализу авторского стиля, определению тональности, эмоций, сарказма. С каждым годом компьютерная лингвистика развивается всё стремительнее, применяя и развивая сложные инструменты и методы.

Одним из разделов компьютерной лингвистики является стилометрия – раздел, занимающийся количественной оценкой лингвистических параметров текстов естественного языка. Эти параметры должны отражать особенности употребления слов, фигур речи, морфологии, синтаксиса, в частности структуры предложений, что определяет узнаваемость стиля написанного текста и обеспечивает уникальность документа [1, с.2]. Возможность выявить такие свойства текста обеспечивает возможность создания методов решения задач автоматической обработки текста, таких как атрибуция и верификация автора, классификация текста по категориям, исследование авторского стиля. Представление текста в виде числового вектора характеристик позволяет использовать большое количество мощных математических методов обработки больших данных, например, методы машинного обучения.

Процесс выбора признаков представляет собой одну из самых больших проблем в стилометрии [2, с.184]. Ученые выделяют порядка тысячи различных параметров на разных по глубине анализа уровнях: лексическом, включая уровни символов и букв, синтаксическом, семантическом, структурном и предметно-специфическом. Однако исследователи уделяют мало внимания пониманию деталей решения задач вычислительной стилометрии. Если можно было бы объяснить решения классификатора, чтобы дать представление о том, почему документы относятся к определенным авторам, жанрам или темам, это могло бы существенно повысить эффективность решения поставленных задач [3, с.458]. Одной из возможных причин описанных проблем является недостаточная интегрированность методов и приемов стилометрических исследований, проводимых представителями разных наук. Специалисты в области компьютерных наук часто не учитывают результаты лингвистических исследований в области теории языковой личности, лингвистики текста, стилистики. Лингвисты в анализе не используют потенциал количественных методов современной теории информации, пользуясь элементарными подсчетами, работая, как правило, с фактами относительного преобладания того или иного свойства/признака текста.

Решением данной проблемы является исследование сложных стилометрических параметров, обладающих потенциалом понимания и интерпретации с точки зрения классической лингвистики и предметной области, к которой относятся тексты. Авторы статьи выделили и успешно использовали для классификации художественных текстов набор характеристик, основанных на повторении слов и фраз, определяющих ритмику текста [4, с.247]. В текущей работе описано продолжение исследования ритмических характеристик как самостоятельных маркеров стиля текста. Для этого была поставлена задача кластеризации текстов разных жанров: научные статьи, отзывы на товары и услуги, политические тексты, реклама, художественные романы.

Кластеризация текстов на основе векторов ритмических характеристик — это задача разбиения множества этих векторов на группы, называемые кластерами. Внутри каждой группы должны оказаться максимально «похожие» вектора, а элементы разных групп должны как можно больше отличаться друг от друга. Визуализация результатов кластеризации осуществляется с помощью метода нелинейного снижения размерности umap [5, с.134]. Получаемая проекция кластеров на двумерную плоскость позволяет увидеть метрические соотношения между текстами.

В качестве исходных данных для кластеризации были взяты русскоязычные тексты пяти жанров, отличающихся по стилю: художественные романы (100 текстов популярных писателей 19, 20 и 21 века), научные статьи (100 текстов из журналов Грамота и Диалог), рекламные интернет-тексты (100 текстов с сайтов auto.ru и detmir.ru), отзывы на отели и рестораны (50 текстов с сайта tripadvisor.com) и политические тексты (50 текстовых расшифровок речей президентов России).

Для каждого текста независимо подсчитывались числовые ритмические характеристики на основе следующих ритмических средств: анафора, эпифора, симплока, анадиплозис, эпаналепсис, многосоюзие, диакопа, эпизевксис, хиазм, апозиопеза, повторяющиеся вопросительные и восклицательные предложения. Алгоритмы их поиска и подсчёта описаны в предыдущей работе авторов [4, с. 3-5]. Характеристики описывают ритм текста как с точки зрения статистики, так и структуры.

- Статистические ритмические характеристики:
  - плотность ритма — количество появлений в тексте ритмического средства, разделённое на количество предложений.
- Структурные ритмические характеристики:
  - доля уникальных слов среди всех, составляющих ритмические средства, т. е. тех, которые появляются только в одном ритмическом аспекте;
  - доли существительных, прилагательных, глаголов и наречий среди слов, составляющих ритмические средства.

Таким образом, текст моделируется как вектор из 16 лексико-грамматических ритмических характеристик.

Полученная матрица векторов подаётся на вход алгоритму кластеризации UMAP. Данный метод уменьшает размерность данных, позволяет их визуализировать в двумерном пространстве и выделить кластеры с близкими по значениям объектами. Результаты работы UMAP представлены на графиках на Рис. 1. Значения гиперпараметров были подобраны вручную: 50 соседей с минимальным расстоянием 0,3. В качестве меры близости рассматривались четыре метрики: Braycurtis, Canberra, корреляция и косинусная мера.

Результаты кластеризации показывают, что романы хорошо отделяются от текстов с другими стилями. То же можно сказать и об отзывах: на каждой картинке появляется отдельный оранжевый кластер. Остальные три жанра сильно смешиваются между собой.

Для того чтобы подтвердить результаты кластеризации, было решено классифицировать тексты на пять жанров и проанализировать ошибки.

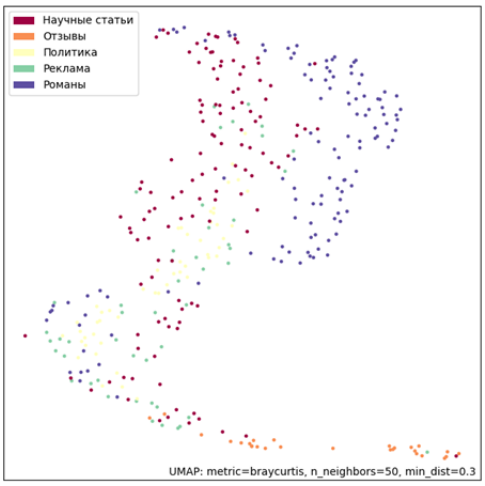
Вектора характеристик были взяты те же, что и для UMAP. Классификатором выступала рекуррентная нейронная сеть со слоем двунаправленной LSTM с 64 блоками и выходным слоем из 5 нейронов, использующим функцию активации Softmax для мультиклассовой классификации. Данная нейронная сеть достаточно популярна в современной литературе о компьютерной лингвистике и регулярно обеспечивает высокие результаты классификации.

Для классификации корпус, состоящий из пяти классов, был разделён случайным образом на обучающую и тестовую выборки в отношении 80%:20%. Это позволило провести пятикратную кросс-валидацию. Оценка качества выполнялась с помощью стандартных метрик: точность, полнота и F-мера. Средние метрики кросс-валидации получились следующими 85.5% точности, 80.6% полноты и 83.0% F-меры. Результаты показывают достаточно высокое качество мультиклассификации.

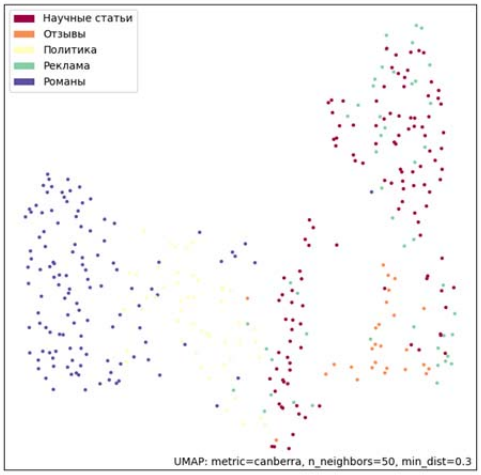
Программные средства для постановки данных экспериментов написаны на языке программирования Python и используют библиотеки Scikit-Learn 0.23.2 и Keras 2.4.3.

Авторы выбрали один из раундов кросс-валидации с характеристиками, наиболее близкими к средним, и проанализировали ошибки классификации. Получились следующие результаты:

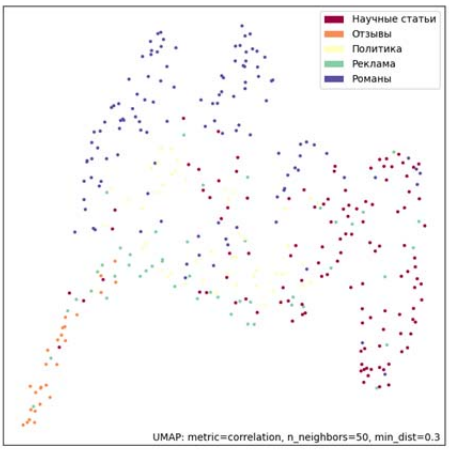
- реклама 4 раза была принята за политические тексты и 3 раза за научные статьи. То есть 7% рекламных текстов не были отнесены к своему жанру;
- одна политическая статья была принята за научный текст. То есть 2% политических текстов не были отнесены к своему жанру;
- пять научных текстов были приняты за политические. То есть 5% научных текстов не были отнесены к своему жанру.



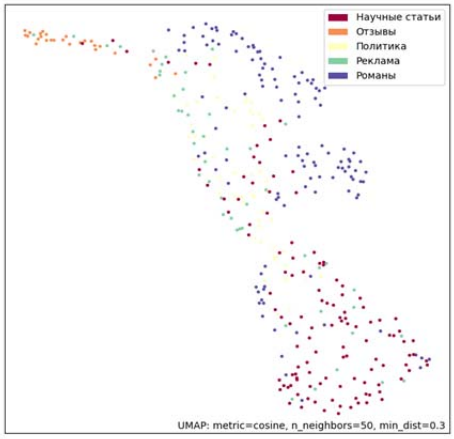
а)



б)



в)



г)

Рисунок – Кластеризация ритма текста с помощью UMAP, метрика близости векторов: а) Braycurtis, б) Canberra, в) корреляция, г) косинусная мера

Ошибки классификации подтверждают результаты кластеризации. Отзывы и художественные романы отделяются от остальных классов без ошибок, тогда как остальные три жанра ошибочно принимаются друг за друга. Наиболее вероятной причиной таких результатов является то, что рекламные, научные и политические тексты содержат достаточно мало ритмических средств, поэтому слабо отличаются по ритмическим характеристикам. В романах встречается достаточно большое число ритмических средств, так что они хорошо выделяются на фоне других жанров. Отзывы, наоборот, практически не содержат лексических и грамматических повторений, так что тоже кластеризуются и классифицируются хорошо.

Таким образом, кластеризация и визуализация русскоязычных текстов в различных жанрах позволяют выявить тексты с отличающимся стилем: художественные романы и отзывы. Класси-

фикация текстов по жанрам подтверждает качество результатов кластеризации.

Работа поддержана стипендией Президента Российской Федерации для молодых ученых и аспирантов, осуществляющих перспективные научные исследования и разработки по приоритетным направлениям модернизации российской экономики: № СП-2109.2021.5.

Список использованных источников

1. Neal T. Surveying stylometry techniques and applications /T. Neal [et al.]. –ACM Computing Surveys (CSuR). – 2017. – V. 50. – №. 6. – P. 1-36.

2. Lagutina K. A survey on stylometric text features /K. Lagutina [et al.]. –Proceedings of the 25th Conference of Open Innovations Association FRUCT. – IEEE, 2019. – P. 184-195.

3. Daelemans W. Explanation in computational stylometry /W. Daelemans. –International conference on intelligent text processing and computational linguistics. – Springer, Berlin, Heidelberg, 2013. – P. 451-462.

4. Lagutina K. Authorship verification of literary texts with rhythm features /K. Lagutina [et al.]. – Proceedings of the 28th Conference of Open Innovations Association FRUCT. – IEEE, 2021. – P. 240-251.

5. Ko H. K. Progressive Uniform Manifold Approximation and Projection /H. K. Ko, J. Jo, J. Seo. – EuroVis (Short Papers). – 2020. – P. 133-137.