**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**

**«РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ ПЕДАГОГИЧЕСКИЙ УНИВЕРСИТЕТ**

**им. А.И. ГЕРЦЕНА»**



Институт информационных технологий и технологического образования

Отчет по научно-исследовательской работе (семестр 2)

студент группы 1А \_\_\_\_\_\_\_\_\_\_\_\_\_\_/ Момот Д.М.

руководитель практики,

научный руководитель

профессор каф. ИС \_\_\_\_\_\_\_\_\_\_\_\_\_\_/ д.т.н., проф. Фомин В.В.

Санкт-Петербург

2021

Оглавление

[1. План работы 3](#_Toc75771056)

[2. Введение 4](#_Toc75771057)

[3. Классификация текста (неспециализированная) 5](#_Toc75771058)

[3.1. Формирование категорий 5](#_Toc75771059)

[3.2. Индексация документов 5](#_Toc75771060)

[3.3. Обучение классификатора 7](#_Toc75771061)

[3.4. Оценка качества классификаторов 9](#_Toc75771062)

[4. Классификация текста (по сложности) 11](#_Toc75771063)

[5. Заключение 12](#_Toc75771064)

[6. Список использованной литературы 13](#_Toc75771065)

# План работы

Тема научно-исследовательской работы: «Обзор актуальных подходов к задаче классификации текстов по сложности».

Сроки работы: с 08.06.2021 по 21.06.2021.

Форма организации: научно-исследовательский семинар.

В ходе научно-исследовательской работы (далее НИР) необходимо: изучить литературу по тематике НИР, по результатам изучения написать отчет. План выполнения НИР представлен в таблице 1.

|  |  |  |
| --- | --- | --- |
| № | Содержание разделов работы;  основные виды деятельности | Сроки выполнения |
| 1 | Анализ предметной области | 08.06.2021 – 12.06.2021 |
| 2 | Определение целей и задач исследования | 13.06.2021 – 14.06.2021 |
| 3 | Выбор и анализ списка литературы | 14.06.2021 – 17.06.2021 |
| 4 | Анализ методик предметной области для класса выбранных задач | 18.06.2021 – 20.06.2021 |
| 5 | Формирование выводов | 21.06.2021 |

Таблица 1. План работы

студент группы 1А \_\_\_\_\_\_\_\_\_\_\_\_\_\_/ Момот Д.М.

руководитель практики,

научный руководитель

профессор каф. ИС \_\_\_\_\_\_\_\_\_\_\_\_\_\_/ д.т.н., проф. Фомин В.В.

# Введение

В настоящее время человечество генерирует огромное количество текстового контента. Вручную анализировать его уже не представляется возможным, что обуславливает бурный рост подходов, связанных с обработкой естественного языка (natural language processing, NLP).

Одной из задач, решаемых NLP, является задача классификации текстов. Она направлена на автоматическое отнесение текста к одной из заранее заданных категорий. Эта функция является крайне полезной для автоматического тегирования, демонстрации релевантной рекламы, борьбы со спамом.

В настоящее время существует большое количество корпусов текстов – размеченных массивов текстов, где каждому тексту сопоставлены категории, которым он принадлежит. Это существенно облегчает разработку классификаторов.

Сейчас активно развивается классификация текстов по сложности. Так, этому посвящены работы [Futrell], [Mukherjee], [Curto], [Vajjala], [Лапошина]. Основная задача понятна – автоматически определить уровень сложности текста, чтобы облегчить обучение языку. В то же время, точность классификаторов, полученных исследователями, пока не так высока (77-91%), что оставляет пространство для возможных улучшений.

В рамках данной научно-исследовательской работы поставлена задача проанализировать существующие подходы к классификации текстов по сложности и выделить наиболее перспективные из них.

Первый раздел посвящен классификации текстов (более общая задача), второй раздел – уточнение общих техник и методов из первой части. Указывается результат их применения (точность классификации).

# Классификация текста (неспециализированная)

Задача классификации текстов формулируется следующим образом. Существуют заранее заданные категории, и необходимо отнести документ к одной из них.

В процессе классификации текстов можно выделить следующие этапы.

1. Формирование категорий.
2. Индексация документов.
3. Обучение классификатора.
4. Оценка качества классификатора.

Осветим подходы к каждому из этапов в отдельности.

## Формирование категорий

Данный этап является одним из наиболее важных, так как неверный выбор категории приведет к бесполезности всех дальнейших этапов. В работе [Oliveira] авторы приходят к выводу, что для достижения оптимальных результатов требуется человеческое вмешательство.

Этот этап также можно назвать первичной классификацией. Выбирается маленькое подмножество обучающей выборки, и оно анализируется для оптимального набора категорий.

## Индексация документов

Этап индексации документов состоит в приведении текстов к форме, удобной для дальнейшей обработки. Такими формами являются:

1. Векторная модель. Строится векторное пространство, единое для всех документов коллекции. Каждая размерность соответствует уникальному слову данной коллекции. Документу из коллекции сопоставляется вектор, значение каждой координаты – количество вхождений соответствующего слова в текст [Sidorov].
2. Терм-документная матрица. Она является обобщением векторной модели. Ее строки соответствуют документам, столбцы – словам, а в ячейках – некоторая функция «важности» слов (см ниже).
3. Мешок слов (bag of words). Текст представляется как мультимножество уникальных для этого текста слов. Для каждого слова хранится количество его вхождений в текст.
4. Мешок N-грамм (w-shingling). Хранятся уникальные для этого текста последовательности из N подряд идущих слов и количество вхождений каждой такой последовательности. При N=1 подход совпадает с мешком слов. Такой подход позволяет частично компенсировать потерю упорядоченности слов.

Для эффективной индексации коллекции документов необходимо научиться решать ряд более простых задач. А именно:

1. Токенизация – разбиение текста на составные части: абзацы, предложения, слова.
2. Нормализация – приведение слов текста к начальной форме. Это позволяет распознать, что две формы слова соответствуют одной лемме – начальной форме – и отнести их к одной размерности вектора или мешка слов. Данная задача является весьма трудной. В зависимости от желаемого качества, она решается с помощью одного из следующих подходов.
   1. Стемминг – отбрасывание окончания слова. Это наименее интеллектуальный, но наиболее простой способ нормализации.
   2. Лемматизация – приведение слова к начальной форме. Необходимо уметь определять часть речи (PoS-tagging). Этот способ нормализации лучше, чем стемминг, но не устойчив к омонимии.
   3. Устранение омонимии – выбор правильной начальной формы (из нескольких возможных) для приведения к ней.
3. Чистка шума (стоп-слов) – удаление слов, не несущих информации (предлоги, артикли). Этим методом необходимо пользоваться с осторожностью, так как он может ухудшить результаты в случае работы с мешком N-грамм, а также исказить стилевой окрас предложения [Yu].

На этапе индексирования документов обычно выполняется выкидывание лишних признаков (features) – не дающих большого количества информации, но усложняющих вычисления. Это выполняется с помощью следующих методов:

1. Чистка слишком редких слов. Такие слова встречаются в очень небольшом количестве текстов из коллекции, а значит не слишком помогут в классификации остальных текстов.
2. Чистка «неважных» слов. Для каждого слова вычисляется метрика важности, и наименее важные слова отбрасываются. Обычно используется TF-IDF метрика, но также могут использоваться CHI и другие метрики [Deng].
3. Стандартные техники понижения размерности, такие как метод главных компонент, анализ независимых компонент.

## Обучение классификатора

Данный этап является классической задачей классификации. На вход подается текст в удобном для классификации виде (векторная модель, мешок слов, N-граммы), необходимо определить, к какому из заранее заданных классов относится входной текст. При проектировании и обучении классификатора можно выделить следующие группы подходов.

1. Метрические методы. К этой группе относится метод ближайших соседей и его модификации. Основная идея состоит в том, что на первом этапе создаются кластеры документов, соответствующие категориям, а затем классификатор относит очередной классифицируемый документ к той категории, к которой принадлежат ближайшие к нему документы обучающей выборки.
2. Вероятностные методы. К этой группе относится наивный байесовский классификатор и его модификации. Результатом работы классификатора являются вероятности принадлежности документа к категориям. Он основан на предположении о том, что признаки вносят независимый вклад в вероятности. По обучающей выборке вычисляется вклад значений признаков в итоговые вероятности, и с учетом значений признаков на конкретном документе дается прогноз принадлежности категориям.
3. Линейные методы классификации. К этой группе относится метод опорных векторов (SVM) и его модификации. Строится гиперповерхность, разделяющая объекты выборки (представляемые точками) на категории оптимальным образом.
4. Методы деревьев решений. Дерево решений – это ориентированное дерево, в узлах которого признаки, а в исходящих ветвях – значения этих признаков. В листьях находятся результаты. Классификация происходит путем спуска по дереву вплоть до листьев. Деревья решений могут действовать по одному или в ансамбле. Также к этой группе можно отнести экспертные системы, работающие на основе правил.
5. Нейросетевые методы. Нейросети предоставляют огромное разнообразие вариаций использования, обеспечивая гибкость настройки. Наиболее часто в классификации текстов используются архитектуры сверточная, рекурсивная, рекуррентная и ее разновидность LSTM.

Изначальные допущения и обусловленные ими достоинства и недостатки описанных методов представлены в Таблице 1.

|  |  |  |  |
| --- | --- | --- | --- |
| Метод классификации | Изначальное допущение | Преимущества | Недостатки |
| Метрический | Категориям соответствуют кластеры документов | Простота  Оптимален, если документы хорошо кластеризуются | Неопределенность количества и расположения кластеров |
| Вероятностный | Независимое влияние признаков | Простота  Высокая скорость  Оптимален, если признаки независимы | Часто низкое качество на реальных данных |
| Линейный | Существование разделяющей гиперплоскости |  |  |
| Дерево решений | Существует набор правил, позволяющих высоковероятно классифицировать | Простота  Не требуется много признаков | NP-полнота построения оптимального варианта требует эвристик  Переобучение (чувствительность к шуму, нестабильность) |
| Нейросетевой | Существование нелинейного закона |  |  |

Таблица 1. Изначальные допущения, преимущества и недостатки методов классификации.

## Оценка качества классификаторов

Оценка качества классификаторов во многом схожа таковой в других областях машинного обучения, но имеет несколько нюансов.

Существует 4 возможных исходов применения классификатора по конкретному входу: два верных результата и два ошибочных результата. Они приводятся в таблице 2, называемой матрицей несоответствия (confusion matrix). Исходы обозначаются TP, FP, FN, TN, и так же обозначается количество входов с соответствующим результатом.

|  |  |  |
| --- | --- | --- |
|  | Да (P) | Нет (N) |
| Предсказано Да | TP (true positive) | FP (false positive) – ошибка I рода |
| Предсказано Нет | FN (false negative) – ошибка II рода | TN (true negative) |

Таблица 2. Матрица несоответствия.

Две основные метрики оценки классификаторов – это точность (precision), полнота (recall). Для их объединения обычно берется F-мера (F-measure), иногда взвешенная F-мера. Они вычисляются по следующим формулам.

Для корректного сравнения результатов классификаторов с использованием одной из метрик необходимо обеспечить выполнение условий:

* Сравнение производится на одинаковых коллекциях (одинаковые множества документов и категорий).
* Учебная, тестовая и проверочная выборки выделены одинаково (либо – что рекомендуется – кросс-валидация с одинаковыми параметрами).
* Документы проиндексированы одинаково.

# Классификация текста (по сложности)

В данном разделе описываются конкретные техники, примененные исследователями в данной тематике, и результаты их применения (точность классификации).

В работе [Mukherjee] основной упор делается на **лексических цепочках** (несколько слов, связанных между собой семантически; близко к понятию «словосочетание», но понятие «лексическая цепочка» шире). По словам авторов, такой подход показывает значительно лучшие результаты, чем классический «мешок слов» (90% против 78%). Использованные авторами классификаторы достаточно стандартны: логистическая регрессия, наивный байесовский алгоритм, деревья решений, линейная SVM и SVM с ядром RBF и случайный лес.

Работа [Curto] использует более классические признаки, но в большом количестве: **52 признака**. Результаты составили 75% для варианта с 5 уровнями сложности и 81% для трехуровневого варианта. Использовались стандартные алгоритмы классификации: наивный байесовский, SVM, логистическая регрессия, k ближайших соседей, решающие деревья, случайный лес и другие.

В работе [Vajjala] авторы провели эксперименты с самостоятельно созданным корпусом текстов. У них есть три варианта каждого текста (простой, средний и продвинутый), с предложениями, которые различаются лишь в нескольких словах. В связи со схожестью предложений, наилучшим оказался подход **с n-граммами букв**. Авторы предполагают, что это связано с тем, что характерные признаки оказываются заключены в приставках и суффиксах (использовался английский язык, в котором, например, большое количество сложных слов имеют суффикс «tion»). Результат составил 77%.

В отечественной науке (даже в настоящее время), насколько можно судить по доступным публикация, например [Кисельников], [Солнышкина], [Мизернов], [Лапошина], [Кутузов], [Глушань] преимущественно освещается «качественный» подход. Используются понятия «сложности» (объективная характеристика), «трудности» (трудность чтения читателем) и «читабельности» (степень понимания текста читателем). Преимущественно используются аппроксимирующие формулы (наподобие американского «автоматического индекса удобочитаемости»).

# Заключение

В процессе выполнения учебной практики выполнен обзор существующих подходов и методов классификации текстов на естественном языке. Проведен анализ их преимуществ и недостатков. Рассмотрены метрики оценки качества классификаторов. Рассмотрены существующие работы в области классификации текстов по сложности, выделены их ключевые идеи.

Таким образом, план выполнен в полном объеме.

# Список использованной литературы

[Oliveira] Oliveira E. Automatic classification of journalistic documents on the Internet. / E. Oliveira, D. B. Filho // Transinformacao – 2017. – V. 29, № 3 – P. 245–255.

[Sidorov] G. Sidorov. Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model / G. Sidorov, A. Gelbukh, H. Gomez-Adorno, and D. Pinto // Computacion y Sistemas – 2014 – Vol. 18, No. 3 – pp. 491-504.

[Yu] Yu B. An evaluation of text classification methods for literary study / B. Yu // Literary and Linguistic Computing – 2008. – V. 23, № 3 – P. 327–343.

[Deng] Deng Z.H. A comparative study on feature weight in text categorization / Z. H. Deng, S.W. Tang, D. Q. Yang, M. Zhang, L. Y. Li, K. Q. Xie // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) – 2004. – V. 3007. – P. 588–597.

[Futrell] Futrell R. The Natural Stories corpus: a reading-time corpus of English texts containing rare syntactic constructions / Richard Futrell. Edward Gibson. Harry J. Tily. Idan Blank. Anastasia Vishnevetsky. Steven T. Piantadosi. Evelina Fedorenko // Lang Resources & Evaluation – 2021. – №55. – P. 63 – 77.

[Mukherjee] Mukherjee P. Using Lexical Chains to Identify Text Difficulty: A Corpus Statistics and Classification Study / Partha Mukherjee, Gondy Leroy, David Kauchak // IEEE J Biomed Health Inform – 2019 September. – №23(5). – P. 2164 – 2173.

[Curto] Curto P. Automatic Text Difficulty Classifier. Assisting the Selection Of Adequate Reading Materials For European Portuguese Teaching / Pedro Curto, Nuno Mamede, Jorge Baptista // CSEDU2015 – 7th International Conference on Computer Supported Education. – P. 36 – 44.

[Vajjala] Vajjala S. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification / Sowmya Vajjala, Ivana Luciˇc // Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications. – P. 297 – 304.