**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 1)

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

руководитель практики,

научный руководитель

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

Санкт-Петербург

2020

Оглавление

[1. План работы 3](#_Toc61563060)

[2. Введение 4](#_Toc61563061)

[3. NLP: уровни и задачи 5](#_Toc61563062)

[4. NLP: подходы к моделированию структур 6](#_Toc61563063)

[5. Задача классификации текстов: описание 7](#_Toc61563064)

[6. Задача классификации текстов: индексация документов 8](#_Toc61563065)

[7. Задача классификации текстов: построение и обучение классификатора 8](#_Toc61563066)

[Пороговые классификаторы 8](#_Toc61563067)

[Линейный потоковый классификатор 9](#_Toc61563068)

[Линейная регрессионная модель 9](#_Toc61563069)

[ДНФ-классификатор 10](#_Toc61563070)

[Нейросетевой классификатор 10](#_Toc61563071)

[8. Задача классификации текстов: оценка классификатора 11](#_Toc61563072)

[9. Заключение 11](#_Toc61563073)

[Список использованной литературы 12](#_Toc61563074)

[Приложение 1. Список использованных терминов из лингвистики 14](#_Toc61563075)

## План работы

Тема научно-исследовательской работы: «Исследование применения алгоритмов обработки естественного языка для задач классификации текстов».

Сроки работы: с 15.12.2019 по 28.12.2019.

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 | 16.12.2019 – 18.12.2019 |
| 2 | Определение целей и задач исследования | 18.12.2019 – 20.12.2019 |
| 3 | Выбор и анализ списка литературы | 20.12.2019 – 26.12.2019 |
| 4 | Анализ методик предметной области для класса выбранных задач | 26.12.2019 – 27.12.2019 |
| 5 | Формирование выводов | 28.12.2019 |

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 информации, и это количество экспоненциально возрастает. Данное явление получило название «информационный взрыв». При этом значительная часть информации предстает в неструктурированном виде: изображения, видео, текст.

В то же время, человек может извлечь из текста достаточно большое количество информации, в том числе полезной. Однако обработать колоссальные объемы текстовых данных «вручную» – немыслимо. Итак, имеется противоречие: генерируется много текста на естественном языке, он содержит полезную информацию, но человеческих ресурсов на ее извлечение не хватает. Разрешением этого противоречия (но не только его) занимается дисциплина «обработка естественного языка», также известная как NLP.

Одна из задач, встающих перед NLP, является задача классификации текстов, то есть отнесения данного текста к одному из заранее определенных классов. Это позволяет разметить большие массивы текстов, что является первым шагом к превращению их в структурированную информацию.

Данная работа ставит целью описание существующих методов классификации текстов не естественном языке, их преимуществ, направлений модификации и оптимизации.

## NLP: уровни и задачи

Обработка естественного языка (NLP) – это дисциплина на стыке искусственного интеллекта и математической лингвистики. Она изучает проблемы компьютерного анализа и синтеза текстов на естественном языке. При этом структуры естественного языка рассматриваются как объекты, которые могут быть смоделированы [1].

Структуры языка принято разделять на несколько взаимосвязанных иерархических уровней. Каждый уровень имеет свои правила сочетания составных единиц. На каждом уровне возникают свои задачи. Они описаны в таблице 2 (содержит данные из [3, 4, 5, 10]).

В филологии принято также выделять фонологический (фонемы либо графемы, самый атомарный) и лексический (слова из лексикона, между морфологическим и синтаксическим) уровни, однако они не имеют своих специфических задач и потому в таблице не приводятся.

Таблица 2. Уровни естественного языка, их единицы и задачи.

|  |  |  |
| --- | --- | --- |
| **Уровень** | **Составные части** | **Задачи** |
| Морфологический  (словоформа -> лемма, теги) | Морфемы  Словоформы | Лемматизация и стемминг  Полный морфоанализ  Разрешение морфологической омонимии |
| Синтаксический (предложение + результат морфоанализа -> АСД) | Словосочетания  Предложения  Сверхфразовые единства | Токенизация/сегментация  Векторизация  Парсинг (построение АСД)  Извлечение названий, терминов |
| Семантический (АСД -> семантическая структура) | Семы | Определение семантики слов и словосочетаний, разрешение неоднозначности  Установление семантических отношений между словами и словосочетаниями  Построение модели представления смысла  Локальный семантический анализ  Тематическое моделирование  Анализ тональности (эмоциональной окраски) |
| Прагматический / дискурсивный | Структуры связного текста | Суммаризация, реферирование, аннотирование  Разрешение анафор, кореференций, извлечение отношений  Дедубликация |

## NLP: подходы к моделированию структур

Подходы к моделированию языковых структур отличаются большим разнообразием, диктуемым обилием и важностью задач NLP. Подходы можно разделить на два вида: четкое и нечеткое моделирование [1].

Четкое моделирование может производиться с использованием аппаратов:

* дискретной математики [8],
* линейной алгебры [9],
* теории графов [Бакиева 2018]
  + семантические сети [Усталов 2017].

Нечеткое моделирование объединяет в себе большое количество различных методов нечеткой математики, в том числе следующиt:

* вероятностный (статистический) анализ [Hastie 2009, Imai 2008, Карпович 2017],
* деревья решений,
* словари и синтаксические правила (экспертные системы)
* онтологический подход
* цепи Маркова:
  + скрытые марковские модели [10],
  + n-граммы
* вероятностные графические модели [Koller 2009]
* нейросетевые технологии [2]:
* рекуррентные нейросети [Грачев 2019],
* LSTM-решения [10, 11],
* CRF-модели [11],
* техники интеллектуального анализа данных:
  + метод опорных векторов,
  + модель максимальной энтропии [10, Ratnaparkhi 1996, Чернявский 2012]
  + наивный байесовский классификатор,
  + эволюционное и генетическое моделирование [Бондарчук 2016].

В наиболее простых случаях модель строится на основе одного поданного на вход текста [6][8], но часто бывает необходимо также учитывать контекст [7].

## Задача классификации текстов: описание

Задача классификации текстов формулируется следующим образом. Существуют заранее заданные категории, и необходимо отнести документ к одной из них. Следует отличать эту задачу от задачи кластеризации текстов, где категории заранее не заданы, и генерируются в процессе как группы семантически похожих документов.

Также классификацию текстов не следует путать с тематическим моделированием (построением тематической модели). Эта задача близка к кластеризации текстов, но является более сложной, решение которой устойчиво к синонимии (синонимы попадут в один класс) и омонимии (омонимы попадут в разные классы) [Батура Стрекалова 2018]. Сходство между текстами определяется с помощью функции сходства [Zhao Gray 2020, 170].

Строгая постановка задачи автоматической классификации текстов выглядит следующим образом. Есть множество категорий , множество документов и неизвестная целевая функция . Требуется построить классификатор , максимально близкий к , имея начальную коллекцию размеченных документов .

Классификатор может выдавать точное решение или степень близости . Категории могут пересекаться или не пересекаться.

Классификация текстов может применяться для решения следующих задач.

* Фильтрация документов, распознавание спама
* Автоматическая аннотирование
* Снятие неоднозначности (автоматические переводчики)
* Составление интернет-каталогов
* Классификация новостей
* Распределение рекламы
* Персонализация контента.

При классификации текстов выполняются следующие этапы.

1. Индексация документов.
2. Обучение классификатора.
   1. Выбор модели.
   2. Настройка параметров модели.
3. Оценка качества классификатора.

## Задача классификации текстов: индексация документов

Текст зачастую представляется как неупорядоченный набор слов (*bag of words*). Тексту сопоставляется разреженный числовой вектор, на каждый терм (слово) из лексикона корпуса текстов в нем выделяется позиция, значение – количество вхождений данного терма в текст.

Способы уменьшения размерности вектора:

* Улучшение алгоритма лемматизации
* Добавить такие данные, как положение слова в тексте, длина текста
* Не учитывать слишком редкие слова («шум») и слишком частые слова (предлоги, артикли).
* Работа с термами
  + Учет «важных» словосочетаний (с помощью вычисления корреляции между встречаемостью словосочетания в тексте и принадлежностью текста категории)
  + Объединение синонимов в единый терм
  + Создание новых термов, исходя из их совстречаемости (cooccurence): объединение слов, которые часто встречаются рядом.
* Использование SVD-разложения. Это позволяет уменьшить базис векторного пространства, оставив в нем всего лишь размерностей (точно или приближенно, в зависимости от данных).

## Задача классификации текстов: построение и обучение классификатора

Итак, задача состоит в том, чтобы выбрать классификатор, принимающий на вход вектор, отождествляющего текст, и определяющего к каким категориям относится текст. Решения данной творческой задачи отличаются большим разнообразием. Приведем некоторые из них.

### Пороговые классификаторы

При использовании этого подхода строится классификатор , который для каждого документа возвращает значение принадлежности к категории . После чего выбирается порог , и результат классификации трактуется как

При выборе порога применяются следующие методы:

* Пропорциональный метод: для каждой категории порог выбирается так, чтобы доля документов из обучающей выборки, отнесенных к категории, была равна доле документов тестовой выборки, отнесенных к этой категории. Однако этот метод опасен переобучением под конкретные выборки.
* Метод ближайших категорий. Каждый документ считается принадлежащим к ближайшим категориям, и пороговые значения выбираются соответствующим образом. Этот метод используется при автоматическом подборе ключевых слов.

### Линейный потоковый классификатор

Выход классификатора вычисляется как скалярное произведение. Категории сопоставляется вектор коэффициентов , его размерность совпадает с размерностью векторного пространства документов. Тогда значение классификатора вычисляется по следующей формуле.

Обычно проводится нормализация, так что в качестве выхода используется косинус угла между векторами и .

Координаты векторов строятся потоковым образом, по каждой категории в отдельности. Вначале выбирается какое-нибудь начальное значение, скажем, вектор . Затем начинаем идти по документам обучающего множества. Если встречается документ из категории , и он неверно классифицируется, в вектор вносится покоординатная поправка (прибавление числа либо умножение на число). Если вектор стал лучше классифицировать документы категории , поправка принимается. Для уменьшения влияния порядка обработки документов, можно использовать более сложный способ построения коэффициентов, наподобие генетических алгоритмов.

Так как косинус принимает значения в интервале , веса малозначимых для классификации слов окажутся близкими к нулю. Это позволяет обоснованно их выкинуть.

### Линейная регрессионная модель

Основная идея данного метода состоит в постоянном уменьшении ошибки модели. Имеется три матрицы.

1. Матрица данных . Каждая ее строка – это вектор-документ.
2. Матрица ответов . Каждая ее строка – правильный выход классификатора по соответствующему документу. Длина строки – количество категорий.
3. Матрица линейных правил .

Ставится задача оптимизации: . Столбцы матрицы – те же , что в линейном классификаторе, но строятся по другому алгоритму. Алгоритм построения матрицы .

1. Для каждой категории взять вектор ответов и спроецировать на линейную оболочку столбцов матрицы .
2. Разложить полученную проекцию по базису столбцов .
3. Это разложение и будет вектором , характеризующим данную категорию.

### ДНФ-классификатор

Данный подход очень похож на механизм нечеткого логического вывода. Классификатор состоит из множества правил вида

Документ принадлежит к категории, если он удовлетворяет ДНФ-формуле, то есть выполняется хотя бы один конъюнкт, входящий в ДНФ.

На первом этапе формируются тривиальные правила, верные для всех документов обучающей выборки и только для них: для каждой категории , включающей документы , строится описывающая её ДНФ:

На данном этапе классификатор переобучен: идеально классифицирует на учебной выборке и бесполезен вне нее, а также слишком сложный. На дальнейших этапах проводится серия упрощений (склеивание скобок, удаление лишних посылок и т.п.) до получения классификатора, способного работать вне учебной выборки.

### Нейросетевой классификатор

Классификатор представляет собой искусственную нейронную сеть, входы которой соответствуют термам, а выходы — категориям. Для того, чтобы классифицировать документ, веса его термов подаются на соответствующие входы сети; активация распространяется по сети, и значения, получившиеся на выходах, и есть результат классификации. Типичный метод обучения такой сети — обратное распространение ошибки (*back propagation*). Если на одном из тренировочных документов получен неправильный ответ на одном из выходов, то ошибка распространяется обратно по сети, и веса рёбер меняются так, чтобы ошибку уменьшить.

## Задача классификации текстов: оценка классификатора

Существует 4 возможных исходов применения классификатора по конкретному входу: два верных результата и два ошибочных результата. Они приводятся в таблице 3, называемой матрицей несоответствия (confusion matrix). Исходы обозначаются TP, FP, FN, TN, и так же обозначается количество входов с соответствующим результатом.

Таблица 3. Матрица несоответствия.

|  |  |  |
| --- | --- | --- |
|  | 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) |

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.

## Заключение

В ходе выполнения работы, поставленные планом задачи были выполнены в полном объеме.

## Список использованной литературы

[1] Заболеева-Зотова, А.В. Использование естественного языка при автоматизации слабо структурируемых процессов в проектировании технических, программных и информационных систем. [Текст]: дис. ... доктор техн. наук: 05.13.01: защищена 23.12.04 / Заболеева-Зотова Алла Викторовна. - Волгоград, 2004. - 378 с. - Введение: с.1.

[2] Гольдберг Й. Нейросетевые методы в обработке естественного языка / Й. Гольдберг; пер. с англ. А. Слинкин. – ДМК-Пресс, 2019. – 282 с.

[3] Большакова Е.И. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие. М.: МИЭМ, 2011. 269 с.

[4] Смазневич И. NLP: как стать специалистом по обработке естественного языка [Электронный ресурс] // Tproger — издание о разработке и обо всём, что с ней связано. URL: https://tproger.ru/blogs/nlp-professional-howto/ (дата обращения: 29.09.2020).

[5] Обработка естественного языка [Электронный ресурс] // Викиконспекты. URL: http://neerc.ifmo.ru/wiki/index.php?title=Обработка\_естественного\_языка (дата обращения: 29.09.2020).

[6] Adesina A., Azeez N., Abidoye P. Context-Aware Stemming Algorithm for Semantically Related Root Words. African Journal of Computing and ICTs, 2012, vol. 5.

[7] Mohler T. Context Analysis in NLP: Why It’s Valuable and How It’s Done [Электронный ресурс] // Lexalytics. Tell powerful stories from complex text data. URL: https://www.lexalytics.com/lexablog/context-analysis-nlp/ (дата обращения: 29.09.2020).

[8] Mou L., Zhou H., Li L. Discreteness in Neural Natural Language Processing [Электронный ресурс] // URL: https://lili-mou.github.io/resource/emnlp19-1.pdf (дата обращения: 29.09.2020).

[9] Brownlee J. 10 Examples of Linear Algebra in Machine Learning [Электронный ресурс] // Machine Learning Mastery: Making Developers Awesome at Machine Learning. URL: https://machinelearningmastery.com/examples-of-linear-algebra-in-machine-learning/ (дата обращения: 29.09.2020).

[10] Кульневич А.Д. Интеллектуальный анализ текстовых данных с применением методов машинного обучения: дис. ... магистра. Томский политехнический университет, Томск, 2019.

[11] Коноплич Г.В. Распознавание именованных сущностей в русском языке с использованием глубоких нейронных сетей: дис. ... магистра. СПбНИУ ИТМО, Санкт-Петербург, 2018.

## Приложение 1. Список использованных терминов из лингвистики

Фонема – звук.

Графема – буква.

Сверхфразовое единство – примерно равно абзацу.

Лексикон – словарь.

Анафора – повторение начальных созвучий, частей слов, словосочетаний, синтаксических конструкций.

Кореференция – обозначение одной сущности (референта) в разных частях текста с помощью различных обозначений. Пример: *Иван Алексеевич Бунин* — выдающийся русский *писатель и поэт*. *Он* является первым *лауреатом Нобелевской премии* по литературе из России.