**МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ**

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ

ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО»

(Университет ИТМО)

**Факультет** инфокоммуникационных технологий

**Образовательная программа** Интеллектуальный анализ данных

**Направление подготовки (специальность)** 45.03.04 – Интеллектуальные системы в гуманитарной сфере

**О Т Ч Е Т**

**об учебной, ознакомительной практике**

Тема задания: Анализ правок и замечаний экспертизы в конструкторе РПД и формулировка предложений по использованию полученных результатов для проверки на ошибки этих замечаний

Обучающийся: Лайок Олег Владимирович, К33421

Согласовано:

Руководитель практики от университета:Валитова Юлия Олеговна, доцент факультета инфокоммуникационных технологий, факультет

инфокоммуникационных технологий

Практика пройдена с оценкой **\_\_\_\_**

Дата **\_\_\_\_**

Санкт-Петербург 2022

**СОДЕРЖАНИЕ**

СОДЕРЖАНИЕ ...................................................................................................... 2

ВВЕДЕНИЕ ............................................................................................................. 3

1. ТЕОРЕТИЧЕСКИЙ ОБЗОР ................................................................................ 4
   1. Постановка проблемы, обоснование ее актуальности, изучение темы .... 4
   2. Методы исследования естественного языка и понятия, связанные с ними и используемые в данной работе……………………………………….... 5
2. ДАННЫЕ ............................................................................................................. 7
   1. Извлечение и подготовка данных для анализа…....................................... 7
   2. Начальный анализ…...................................................................................... 8
   3. Предварительная обработка текстовых данных………………….............. 9
   4. Нахождение основных тем комментариев ….............................................10
3. ОПИСАНИЕ РЕШЕНИЯ ПОСТАВЛЕННОЙ ЗАДАЧИ ............................... 11

ЗАКЛЮЧЕНИЕ .................................................................................................... 12

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ ........................................... 13

ПРИЛОЖЕНИЕ А................................................................................................. 14

ПРИЛОЖЕНИЕ Б. ................................................................................................14

ПРИЛОЖЕНИЕ В. ................................................................................................15

ПРИЛОЖЕНИЕ Г. .................................................................................................15

ПРИЛОЖЕНИЕ Д. ............................................................................................... 16

ПРИЛОЖЕНИЕ E. .................................................................................................17

# ВВЕДЕНИЕ

В качестве темы учебной ознакомительной практики был выбран анализ правок и замечаний экспертизы в конструкторе РПД и формулировка предложений по использованию полученных результатов для проверки на ошибки этих замечаний.

В процессе изучения теоретический материалов по выбранной теме, были поставлены задачи:

1. Ознакомится с сервисом конструктора РПД и найти необходимые разделы для исследования
2. Ознакомится с базой данных сервиса и извлечь из нее данные для анализа
3. Провести обработку полученных данных и выбрать метод их исследования
4. Проанализировать нужные данные используя выбранный метод
5. Проанализировать результаты машинного анализа, сформулировать предложения по проверкам заполняемых данных на сайте используя полученные результаты исследования

Работа с данным вопросом предоставляет возможность получить опыт в использовании методов машинного анализа данных, а именно естественного языка, на задаче из рабочей области, а также подготовить фундамент для последующего улучшения сервиса конструктора рпд.

# ТЕОРЕТИЧЕСКИЙ ОБЗОР

## Постановка проблемы, обоснование ее актуальности, изучение темы

В современном мире образовательный процесс в высших учебных заведениях находится в состоянии постоянного обновления и цифровизации. Преподавателям и учащимся предоставляется все больше возможностей и ресурсов для более комфортного обучения в университете с использованием современных информационных технологий. Проект, частью которого мне удалось стать, а именно конструктор рпд, позволяет преподавателям не только удобно заполнять и редактировать рабочие программы дисциплин, но и делать их более доступными и понятными для студентов, которые в будущем смогут формировать свой индивидуальных образовательный путь, выбирая дисциплины, которые максимально эффективно помогут им приобрести необходимые компетенции в их выбранной специальности. Именно поэтому так важно, чтобы представленные рабочие программы дисциплин были заполнены по всем нормам и шаблонам, для того чтобы как у студентов, так и у преподавателей формулировалось правильное понимание о той или иной дисциплины.

В процессе изучения материала, необходимого для качественного проведения исследования в рамках проходимой практики, было сделано:

* Изучен [сайт](https://op.itmo.ru) конструктора рпд, его разделы и возможности
* Изучены материалы по функционалу сайта, предоставленные его создателями
* Сайт был развернут локально на ПК, был получен доступ к базе данных проекта

## Методы исследования естественного языка и понятия связанные с ними используемые в данной работе

Для данной работы стоит выделить несколько терминов и понятий, связанных с компьютерной лингвистикой, которые использую при описании процесса исследования.

Словарь – список всех слов, встречающихся в документе или коллекции документов

Корпус слов – список уникальных слов, из которых состоит документ или коллекция документов

Стоп-слова – слова, являющие ‘шумом’ при исследовании документа с использованием понятия ключевых слов. Данные слова никак не влияют на передачу основных тем документа и состоят обычно из служебных частей речи а так же часто встречающихся слов не имеющих смысловой нагрузки в конкретном наборе данных (например слова приветствия или просьбы).

Лемматизация – приведение слова к начальной словарной формы; обычно используется для того, чтобы алгоритм не различал несколько одинаковых слов в разных грамматических формах при нахождении ключевых слов и основных тем.

N-граммы – структуры естественного языка состоящие из n слов или символов стоящих рядом. В данной работе будут использоваться биграммы (два рядом стоящих слова) и триграммы (три рядом стоящих слова), чтобы искать не ключевые слова а ключевые словосочетания для каждого набора документов

Тематическое моделирование – построение тематическое модели каждого документа, т. е. определение темы к которой относится каждый документ коллекции. В данной работе документом является каждый отдельно взятый комментарий в своей группе комментариев с одинаковыми тэгами. С помощью данной модели можно улучшить поиск основных тем комментариев т. к. для каждой группы будет составляться не просто набор ключевых слов а набор основных тем этих комментариев, из которых уже и можно выявить, какие предложения по исправлению каких ошибок являются самыми частыми. В данной работе для решения задачи тематического моделирования и нахождения наиболее частотных тем используется алгоритм латентного размещения Дирихле (далее LDA)

LDA – алгоритм тематического моделирования, основан на вероятностной модели (с какой вероятностью тема встречается в документе и с какой вероятностью слово встречается в данной теме). На основе этих вычислений формируется вектор документов (какая тема с какой вероятностью встречается в документе) и вектор тем (с какой вероятностью слово встречается в теме) и на основе этих признаков можно выделить наиболее весомы темы в документе, т. е. которые встречаются в большом количестве документов. Замечу, что в данном алгоритме роль слов могу выполнять n-граммы (в данной работе в этом алгоритме как раз используются биграммы, а не просто слова)

Когерентность – или coherence, характеристика, показывающая насколько наиболее частые слова или n-граммы коллекции встречаются рядом в документах. Чем выше показатель когерентности, тем, тем лучше выполнено выделение тем. Именно он будет использоваться для нахождения оптимального количества основных тем коллекции документов.

# ДАННЫЕ

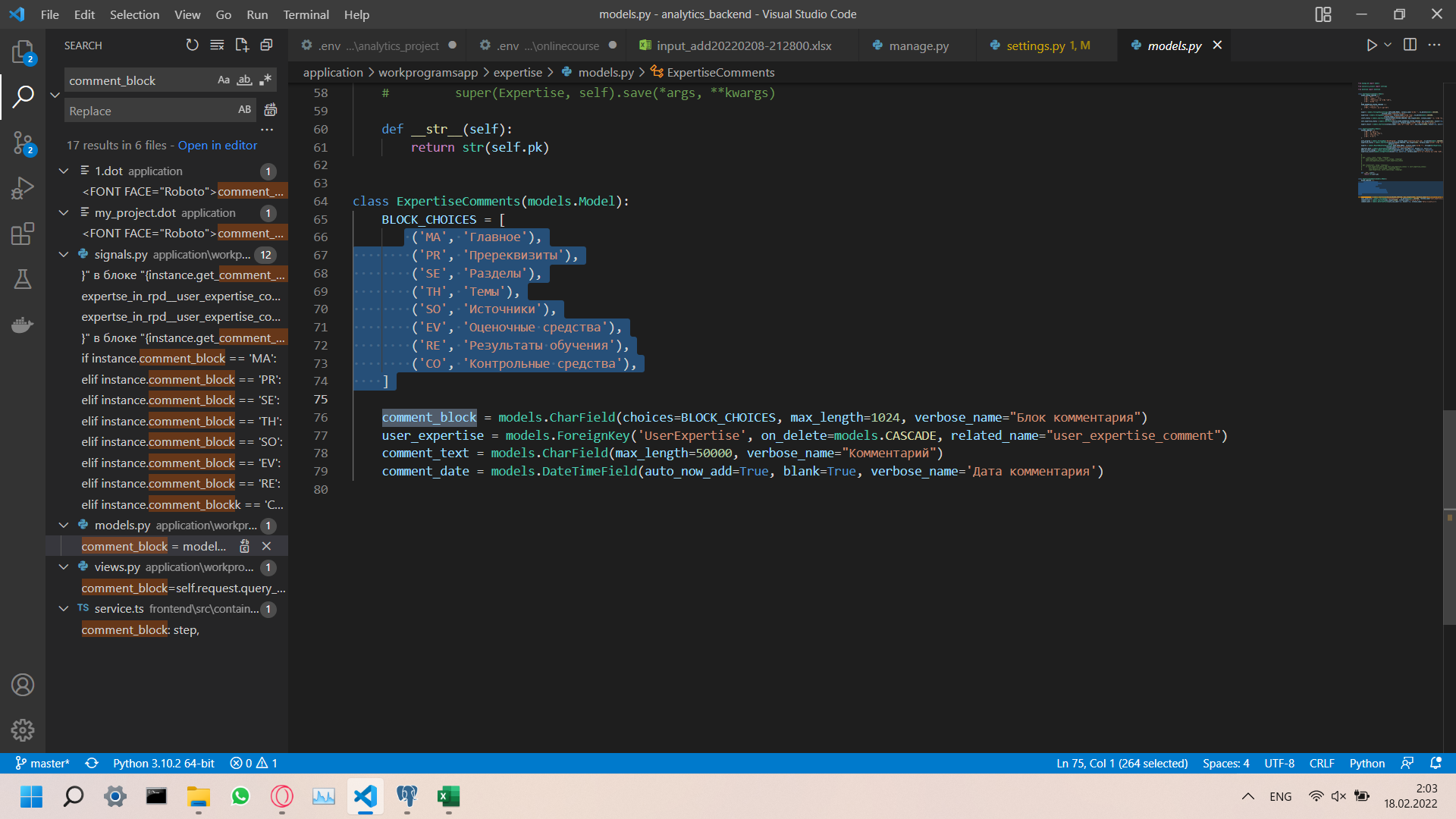
## Извлечение и подготовка данных для анализа

С помощью средства СУБД pgAdmin было выполнено подключение к базе данных проекта. Была найдена конкретная таблица с данными комментарием экспертизы и экспортирована в формат .xlsx для более удобного анализа.

Изображение выглядит как текст

Автоматически созданное описание

Анализ данных производился с помощью языка Python поэтому было принято решение использовать среду разработки Google Collab для более наглядного представления результатов. При загрузке данной таблицы замечено поле comment\_block которое представляет собой конкретный тэг для каждого комментария экспертизы. Из кода самого проекта было выяснено, что эти тэги показывают к какому разделу заполнения рпд на сайт относится каждый комментарий, здесь представляю список этих тэгов:



Логически был сделан вывод что имеет смысл анализировать комментарии для каждого из этих тэгов, поэтому данные были разбиты на группы по этим тэгам. Для данной работы единственное поле подлежащее анализу это поле comment\_text, где находится само замечание эксперта, поэтому все последующие исследования проводились именно с ним

## Начальный анализ

Было принято решение посмотреть какое количество комментариев есть по каждому тэгу, чтобы найти раздел, в котором совершается больше всего ошибок. Было выявлено, что это разделы Оценочные средства, Контрольные средства и Результаты обучения, заначит именно их стоит рассматривать в первую очередь при формулировке предложений по автоматической проверке на ошибке, однако анализ был проведен на каждом разделе, т. к. может выясниться, что и в разделах с малым количеством комментариев будут выявлены однотипные ошибки

Изображение выглядит как текст, снимок экрана, монитор, внутренний

Автоматически созданное описание

## Предварительная обработка текстовых данных

При анализе каждого раздела с данными производятся следующие преобразования: разбиение каждого комментария на токены, удаление стоп слов и пунктуации. Для составления списка стоп слов была создана функция, которая находит самые часто встречающиеся слова в каждом разделе, чтобы вручную добавить нужные стоп слова, например такие как ‘добрый день’, ‘здравствуйте’, ‘пожалуйста’ и т. д., чтобы алгоритм выделял самые часто встречающиеся темы комментариев именно по значимым словам. После применения данной функции к каждому разделу в список стоп слов были добавлены несколько таких характерных для данных документов. Так же было принято решение не приводить слова к начальной форме, т. к. из-за специфики данных грамматическая форма у слов похожая (повелительное наклонение у глаголов и одинаковые падежи у сущ.), а значит алгоритму не нужно решать проблему одинаковых слов в разных формах, более того лемматизация слов затруднила бы анализ результатов.

## Нахождение основных тем комментариев

Для нахождение основных тем документов было использовано два подхода: выделение наиболее часто встречающихся биграмм и триграмм, и сравнение с результатами выделения основных тем с помощью алгоритма LDA. Замечу, что для алгоритма LDA использовалось выделение тем по биграммам, т. к. для данной коллекции документов логичнее делать анализ не по ключевом словам а словосочетаниям, чтобы легче определить основные темы комментариев. Таким образом выводя не только результаты работы LDA, мы можем понять по найденным частым n-граммам какие термины формируют каждую тему. Это поможет нам лучше проанализировать результаты работы алгоритма и снизить ручной анализ комментариев к минимуму. Для определения оптимального количества тем использовался параметр когерентности, которых высчитывал при скольких темах (от 2 до 30) значение этого параметра будет максимальным. Больше 30 тем для данного исследования нигде не нужно, т. к. количество и разнообразие комментарии даже в категории с максимальным количеством комментариев не подразумевает такого колличества тем.

# ОПИСАНИЕ РЕШЕНИЯ ПОСТАВЛЕННОЙ ЗАДАЧИ

В ходе анализа полученных результатов были выделены следующие самые распространенные ошибки в данных категориях:

1. Главное: необходимо добавить описание дисциплины, если дисциплина реализуется на английском языке добавить описание на английском, добавить авторский состав
2. Пререквизиты: добавить или дополнить пререквизиты (если дисциплина на английском, то и пререквизиты должны быть на английском), добавить учебные сущности
3. Темы: недостаточное кол-во тем, несоответствие выделенных часов на раздел с теоретическим объемом раздела
4. Источники: добавить источники (аналогично предыдущим пунктам если необходимо, то на английском), добавить ссылки на электронную библиотечную систему
5. Оценочные средства: добавить шкалу и критерии оценивания, шаблон отчета и пример выполнения лабораторной работы, форму (технологию проведения работы
6. Результаты обучения: добавить какие оценочные средства развивают каждый пререквизит (постреквизит), добавить оценочные средства к каждой сущности
7. Контрольные средства: добавить список вопросов к зачету/экзамену, шкалу(критерии) оценивания зачета/экзамена

После выделения главных комментариев экспертизы можно выделить следующие предложения по автоматической проверке на данные замечания:

1. Проверять язык дисциплины и язык описания дисциплины, пререквизитов и источников на соответствие
2. Определить и установить минимальное количество тем и пререквизитов необходимых для корректного оформления рпд
3. Проверять на наличие шкалы и критерий оценивания в оценочных средствах, а также сделать обязательным поле шаблонов отчета и примеров лабораторных и практических работ, проектов
4. Сделать обязательное поле для добавления списка вопросов к экзамену зачету, а также сделать обязательной для заполнения шкалу оценивания при добавлении оценочного средства промежуточной аттестации (контрольного средства)

# ЗАКЛЮЧЕНИЕ

В результате прохождения практики были получены знания в сфере обработки естественного языка, а также получен опыт работы над реальным проектом. Полученные данные и результаты их анализа в дальнейшем планируется применить в проекте конструктора рпд, чтобы упростить труд экспертов и сделать систему более удобной и понятной для преподавателей и студентов.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. <https://op.itmo.ru>
2. Воронцов К. В. Вероятностное тематическое моделирование: теория, модели, алгоритмы и проект BigARTM. 2020.
3. Воронцов К. В., Потапенко А.А. Регуляризация вероятностных тематических моделей для повышения интерпретируемости и определения числа тем

# ПРИЛОЖЕНИЕ А

Вид рабочего проекта дисциплины на сайте конструктора рпд

Изображение выглядит как текст, снимок экрана, монитор

Автоматически созданное описание

# ПРИЛОЖЕНИЕ Б

Функция для нахождения самых частых слов

Изображение выглядит как текст

Автоматически созданное описание

# ПРИЛОЖЕНИЕ В

Функция нахождения самых частых тем комментариев

Изображение выглядит как текст, снимок экрана, монитор, внутренний

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, монитор, электроника

Автоматически созданное описание

# ПРИЛОЖЕНИЕ Г

Функция для вычисления когерентности

Изображение выглядит как текст, снимок экрана, монитор, внутренний

Автоматически созданное описание

# ПРИЛОЖЕНИЕ Д

Пример результатов нахождения основных тем комментариев

Изображение выглядит как текст, снимок экрана, монитор, компьютер

Автоматически созданное описание

Изображение выглядит как текст

Автоматически созданное описание

Изображение выглядит как текст, компьютер, монитор, внутренний

Автоматически созданное описание

# ПРИЛОЖЕНИЕ E

[Ссылка](https://colab.research.google.com/drive/1KxPY6rWVdwHbfEbbTRR-dTCtEW8aHy58?usp=sharing) на интерактивные результаты нахождения основных тем комментариев и код работы.