Введение

Классификация документов — одна из задач информационного поиска, заключающаяся в отнесении документа к одной из нескольких категорий на основании содержания документа. Информационный поиск до некоторого времени оставался скромной научной и прикладной областью, в которой работало относительно небольшое количество ученых. Однако в последние годы эта область получила сильное развитие. Современный информационный поиск позволяет эффективно работать с большими объемами данных, которые генерируются ежедневно.

В машинном обучении задача классификации относится к разделу обучения с учителем. Для эффективной работы алгоритма классификации необходима обучающая выборка- некоторое количество хороших образцов документов из каждого класса. Однако в обучении с учителем сохраняется необходимость предварительной подготовки (ручной разметки) данных.

В настоящее время существует множество новостных сайтов, предоставляющих разнообразный контент. Для удобства пользователей были созданы новостные агрегаторы, которые собирают информацию в одном месте. Однако для классификации новостей по темам они либо используют ручной подход, который может быть субъективным и подвержен ошибкам, либо ориентируются на темы оригинальных источников. В первом случае это требует значительных ресурсов, так как увеличивается количество сотрудников, а во втором — необходима точная настройка тематического соответствия между агрегатором и каждым отдельным сайтом. Кроме того, это ограничивает возможность работы с ресурсами, у которых нет тематической разметки.

Актуальность данного исследования заключается в разработке методов автоматического разделения новостей на заранее определенные темы. Это позволит автоматизировать работу новостных агрегаторов и использовать новостные ресурсы без необходимости предварительной разметки.

Объектом исследования является применение методов классификации для создания навигационных инструментов по коллекции документов. Предметом исследования выступает процесс разбиения новостных материалов на темы с использованием классификации и векторных моделей. Цель работы заключается в сравнении методов машинного обучения для классификации и векторизации новостных статей.

Постановка задачи

Целью данной работы является построение автоматических классификаторов и векторизаторов новостных документов и сравнение методов их работы на заданном корпусе.

Поставленную задачу можно разбить на следующие подзадачи:

- сбор базы данных новостных статей
- предварительная обработка текстов коллекции
- выбор, разработка и применение методов векторизации документов
- описание классов и разбиение данных при помощи тематической разметки
- выбор, разработка и применение методов классификации данных
- оценка качества классификации
- проведение сравнительного анализа полученных результатов
- выбор лучшего решения
- разработка telegram-бота