Введение

В настоящее время существует множество новостных сайтов, генерирующих содержимое разностороннего характера. Для объединения всей информации в одном удобном пользователю месте, были созданы новостные агрегаторы. Однако для разделения новостей по темам они либо используют ручной подход, либо ориентируются на то, к какой теме принадлежит новость в оригинальном источнике. В первом случае разметка будет весьма субъективна, к тому же могут допускаться ошибки. Также наличие большого объема источников пропорционально увеличивает необходимый штат сотрудников. Во втором случае необходимо настраивать точную сеть тематического соответствия между новостным агрегатором и каждым сайтом в отдельности. А также исключается возможность использования ресурсов, на которых отсутствует тематическая разметка.

Актуальность работы заключается в исследовании методов автоматического разделения коллекции новостей на заранее заданные тематики. Это поможет автоматизировать новостные агрегаторы и позволит им пользоваться новостными ресурсами без предварительной разметки.

Объект исследования– применение методов классификации для предоставления пользователю средств навигации по коллекции документов.

Предмет исследования– разбиение новостных документов на темы при помощи классификации и векторных моделей.

Цель работы– сравнение методов машинного обучения в задаче 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