Наруцкий Владислав

Анализ эмоциональной окраски текста (<https://github.com/Verbon/SentimentAnalysis>)

Работа выполнена на основе отзывов о фильмах из пакета для Python - Natural Language Toolkit (<http://www.nltk.org/>). Данные содержат в себе 2 группы отзывов, разделенных на положительные и отрицательные. Цель работы - построение предиктивной модели для определения эмоциональной окраски произвольного текста.

В современных системах автоматического определения эмоциональной оценки текста чаще всего используется одномерное эмотивное пространство: позитив или негатив (хорошо или плохо). Поэтому в данной работе был применен вид классификации по бинарной шкале. В этом случае для определения полярности документа используется два класса оценок: позитивная или негативная. Одним из минусов данного подхода является то, что эмоциональную составляющую документа не всегда можно однозначно определить, т.е документ может содержать как признаки позитивной оценки, так и признаки негативной.

В данной работе в качестве метода классификации тональности используется машинное обучение с учителем. Сутью такого метода является то, что на первом этапе обучается машинный классификатор на заранее размеченных текстах, а затем используют полученную модель при анализе новых текстов.

Краткий алгоритм:

1. Подготовка выборки - нужно сформировать выборку и вычислить признаки текста. Для этого сначала из текста удаляются слишком редкие или популярные слова (стоп-слова). Затем вычисляются признаки. Это могут быть n-граммы, skip-граммы, на которых вычисляются счётчики или TF-IDF.
2. Обучение модели – обучать модель можно различными методами (ансамбли деревьев, градиентный бустинг, байесовские методы, линейные модели). Для данной работы был выбран метод линейной модели (основанной на методе опорных векторов), т.к. он хорошо масштабируется и может работать с большим количеством признаков, на очень больших выборках.
3. Определение эмоциональной тональности произвольного текста.

**Изначально поставленная задача**

Определить эмоциональную тональность высказываний "Amazing film! I will advice it to all my friends. Genious" и "Awful film! The man who advised me to watch it is really crazy idiot."

**Результат**

Построенная предиктивная модель корректно определила эмоциональную тональность вышеупомянутых высказываний:

"Amazing film! I will advice it to all my friends. Genious" – 1

"Awful film! The man who advised me to watch it is really crazy idiot." – 0,

Где 1 и 0 значат позитивную и негативную окраски соотвественно.

**Вывод**

В результате данной работы была изучена предметная область связанная с использованием машинного обучения для определения эмоциональной окраски текста, изучены основные виды и методы и классификации, а также успешно реализован собственный вариант предиктивной модели.