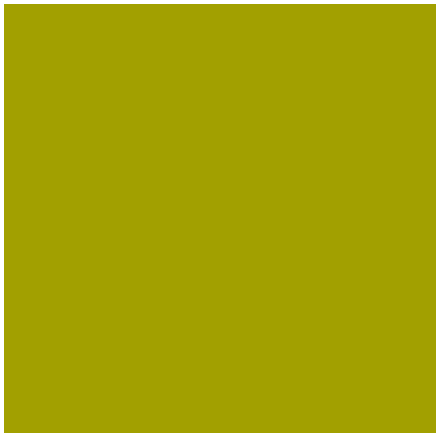
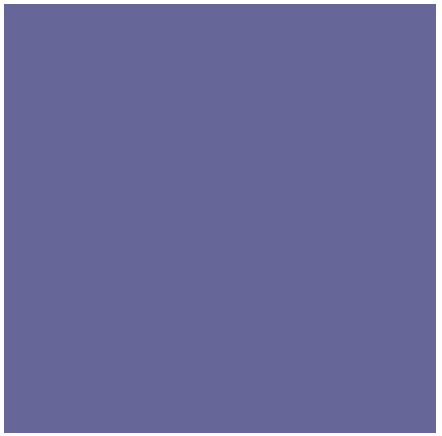
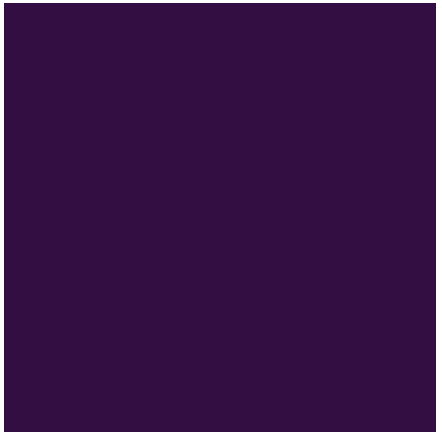
Определение тональности текста

Асирян Александр



Задача определения тональности текста

- Определение тональности всего текста
- Классификация только на положительное и отрицательное

Данные

Рецензии с https://www.kinopoisk.ru

- положительная рецензия с оценкой от 7 до 10 включительно: положительно
- отрицательная рецензия с оценкой от 1 до 4 включительно: отрицательно

По 5000 положительных и отрицательных примеров

Инструменты NLTK

NLTK - пакет библиотек и программ для символьной и статистической обработки естественного языка, написанных на языке программирования Python.

- сегментация на предложения
- токенизация
- морфологический анализ
- анализ тональности
- машинный перевод
- построение дерева зависимостей
- подсчет n-грамм

Инструменты scikit-learn

scikit-learn – это библиотека для машинного обучения на языке программирования Python.

Алгоритмы:

- классификация
- регрессия
- кластеризация
- …

Инструменты pymorphy2

pymorphy2 – морфологический анализатор, разработанный на языке программирования Python. Использует словарный подход (OpenCorpora). Кроме извлечения грамматической информации о слове, умеет ставить слово в нужную форму, например начальную.

Предобработка текста

- удаление ссылок
- удаление смайлов
- удаление переносов строк
- замена повторяющихся точек, пробелов или табуляций на один символ

Алгоритмы и признаки

Базовый
- SVC
- CountVectorizer для униграмм и биграмм слов

Остальные
- RandomForestClassifier
- PassiveAggressiveClassifier
- Perceptron
- LogisticRegression
- MultinomialNB
- CountVectorizer для униграмм и биграмм слов и букв
- pymorphy2 для морфологического анализа в CountVectorizer
- TfidfTransformer

Результаты

Классификатор	Точность	Полнота	F1
SVC	62.89	52.76	41.06
RandomForestClassifier	63.52	63.83	62.84
LogisticRegression	87.41	86.89	86.84
Perceptron	93.9	93.86	93.86
PassiveAggressiveClassifier	92.56	93.01	88.96
MultinomialNB	94.18	94.07	94.06