Анализ тональности

Обработка естественного языка (Natural Language Processing) считается основополагающей для дальнейшего развития искусственного интеллекта.

Среди наиболее интересных и популярных методов этого широкого научного направления особняком стоит один, носящий название sentiment analysis, что в переводе на научный русский означает «анализ тональности текстов». Общее определение гласит, что анализ тональности текстов – это класс методов контент-анализа, предназначенный для автоматического выявления в тексте эмоционально окрашенной лексики, а также мнений (эмоциональных оценок) автора по поводу объектов, о которых идет речь в тексте.

Данные хранятся в файлах train.json (около 8000 записей с указанием правильной тональности) и test.json (около 2000 записей без указания правильной тональности), и доступны для скачивания.

Описание полей train.json файла:
text – текст, который будет использоваться для обучения модели;
id – уникальный идентификатор;
sentiment – тональность (нейтральная, позитивная, негативная) текста.

Описание полей test.json файла:
text – текст, тональность (нейтральная, позитивная, негативная) которого нужно определить;
id – уникальный идентификатор;

Задача состоит в построении модели, которая будет определять тональность (нейтральная, позитивная, негативная) текста (в файле **test.json**). Для этого нужно обучить модель на существующих данных (train.json). Стоить отметить, что по правилам, использовать сторонний корпус текстов с размеченной тональностью запрещено, но это не запрещает использование НЕ размеченных корпусов для обучения таких моделей как word2vec или для предобработки текста. Обучение модели производится при помощи алгоритмов машинного обучения. Полученная модель должна будет определить класс (нейтральный, позитивный, негативный) новых текстов (тестовых данных, которые не использовались для построения модели test.json) с максимальной точностью.

Процесс состоит из двух частей:

1. Построение модели на данных из файла train.json;
2. Предсказание тональности, используя ваш алгоритм, на данных из файла test.json. На последнем этапе вы должны отправить **результаты** анализа тональности на тестовых данных, а именно csv файл (пример на samples.csv) содержащий поля id (уникальный идентификатор документа, тональность которого вы определили), sentiment (тональность, предсказанная моделью: negative, positive, neutral), **также нужно отправить исходный код алгоритма, если таковой имеется, а для тех кто будет использовать готовые инструменты для построения алгоритма (Rapid Miner, SAS или другие SaaS), мы попросим наглядно продемонстрировать как работает модель**;

Ремарка: **ручная разметка данных из файла test.json и дальнейшая отправка как ответ - запрещена!**

3. Проверка результатов. Для проверки будет использоваться f1-score:

F-мера представляет собой гармоническое среднее между точностью и полнотой. Она стремится к нулю, если точность или полнота стремится к нулю.

$$F = 2\frac{Precision \times Recall}{Precision + Recall}$$

TP— истино-положительное решение;
TN— истино-отрицательное решение;
FP— ложно-положительное решение;
FN — ложно-отрицательное решение.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Примеры:

В 2012 году Армстронг по итогам расследования Американского антидопингового агентства был уличен в использовании запрещенных препаратов.

Тональность: негативная

Пока что я остаюсь при своем — Samsung Galaxy Note 3 — это лучший гаджет, что проходил через мои руки!

Тональность: позитивная

Казахстанцы должны платить за коммунальные услуги по месту временной регистрации.

Тональность: нейтральная

Обучающие материалы по обработке естественного языка и машинному обучению:

1. https://www.kaggle.com/c/word2vec-nlp-tutorial/details/part-1-for-beginners-bag-of-words
2. http://scikit-learn.org/
3. Andrew Ng Machine Learning coursera

4. http://www.nltk.org/