Отчет по заданию «Сентимент анализ»
студента 620 группы Бугаевского Владимира

Все документы подвергались морфологической предъобработке. Количество примеров каждого из классов в обучающей выборке похоже на количество примеров в тестовой выборке:

- обучающая выборка – `{-1: 1864, 0: 914, 1: 1115}`;
- тестовая выборка – `{-1: 1890, 0: 1235, 1: 1448}`.

Были попробованы следующие типы векторов:

- binary – бинарный вектор вхождений термов (tf > 0);
- count – вектор частот вхождений термов (tf);
- tfidf – Tf-Idf вектор термов.

Также дополнительно строились две пары векторов: со стоп-словами (stopwords) и без них.

Были попробованы следующие классификаторы:

- логистическая регрессия (LogReg);
- наивный Байесовский классификатор (Naive Bayes);
- линейный SVM (Linear SVM);
- нейронная сеть с двумя скрытыми слоями (NN).

Замечание: Naive Bayes нельзя использовать с векторами tfidf.

Замечание: вектора для линейных моделей: логистической регрессии, SVM, нейронной сети – дополнительно нормировались по l2-мере.

Подробные результаты для всех возможных пар можно увидеть в прикрепленном к отчету jupyter-ноутбуке. В отчете покажем лишь сводную таблицу с наилучшими результатами для каждого из классификаторов.

cls	vector	stopwords	f1-measure	
			micro	macro
LogReg	binary	0	0.617538	0.581217
Naive Bayes	count	0	0.637656	0.591538
Linear SVM	tfidf	0	0.597638	0.570190
Linear SVM	count	0	0.596764	0.571277
NN	binary	0	0.636781	0.593978
NN	tfidf	0	0.632845	0.595804

Отметим сразу несколько любопытных фактов:

- стоп-слова только ухудшали результат: в большинстве случаев модель на векторах со стоп-словами показывала лучший результат, чем та же модель на векторах без стоп-слов;
- простейшая логрегрессия уже дает неплохой результат (3-е место по F1-мере);
- лучшие результаты показали многослойный персептрон (нейронная сеть) и наивный Байесовский классификатор, неожиданно просел линейный SVM.