Результаты наблюдений

Сухочев Александр и Балакший Андрей

4 мая 2015 г.

Гляньте на эту табличку!
Logistic Regression with 1 0
Accuracy: 0.663265306122
All test count: 98; TP: 7; TN: 58; FP: 10; FN: 23
и
Logistic Regression with tf idf
Accuracy: 0.704081632653
All test count: 98; TP: 1; TN: 68; FP: 0; FN: 29

С Тф идф выдает лучший результат, но посмотрите на ТП! Просто так совпало что у нас больше негативных в тестирующем множестве, поэтому мы не можем сказать что тф идф лучше работает, видно ведь что он хуже определяет Позитивные! Переделывать обучающее множество нельзя, т.е. у нас и при работе программы будет подобное соотношение позитивных/негативных, но спасибо за полезный опыт

1 Таблицы точностей различных машинных обучений и экстракторов.

1.1 Общая таблица

	SE	N(only)-gramm	N-gramm E	SE with not
MNB	0.6655			
GNB	0.6224			
SVM 1 0	0.6834			
SVM tf-idf	0.6789			
LogReg	0.7013			

1.2 Более подробно

1.2.1 Standart Extractor with mystem

	Correct	Total	Acc	TP	TN	FP	FN
MNB	742	1115	0.6655	61	618	51	322
GNB	694	1115	0.6224	159	535	197	224
SVM 1 0	762	1115	0.6834	160	602	130	223
SVM tf-idf	757	1115	0.6789	71	686	46	312
Log Reg count	782	1115	0.7013	134	648	84	249

1.2.2 Standart Extractor without mystem

	Correct	Total	Acc	TP	TN	FP	FN
MNB	668	1115	0.5991	149	519	213	234
GNB	694	1115	0.6224	159	535	197	224
SVM 1 0	729	1115	0.6538	120	609	123	263
SVM tf-idf	757	1115	0.6789	48	709	23	335
Log Reg count	744	1115	0.6673	96	648	84	287

Вывод: гляньте как плохо всё без майстема! Вывод: а что поменялось ?

2 Фичи

2.1 Standard Extractor with mystem

2.1.1 Склеивать не + слово

2.1.2 Количество строк

SVM 1 0

	Correct	Total	Acc	TP	TN	FP	FN
0	668	1115	0.5991	149	519	213	234
1	668	1115	0.5991	149	519	213	234
2	694	1115	0.6224	159	535	197	224
3	729	1115	0.6538	120	609	123	263
4	757	1115	0.6789	48	709	23	335
5	744	1115	0.6673	96	648	84	287