Бинарная классификация научно-популярных текстов для различных жанров

Ю.Коломенская

Постановка задачи

Проверить, насколько сложно классификаторам отличать друг от друга тексты различных жанров (бинарная классификация: новости vs статьи, статьи vs лекции).

Какие из жанров классификатору разделить сложнее, какие проще?

Гипотеза: расшифровки лекций проще отделять от новостей или статей,

чем новости от статей

Данные

Материалы с научно-популярных русскоязычных ресурсов:

- Indicator.ru
- Politru.lectures
- Arzamas
- N+1 новости

Классификаторы

- MultinomialNB
- LogisticRegression
- LinearSV
- DecisionTreeClassifier

Обработка

Тексты с каждого ресурса оформлялись в pandas (текст + метка класса):

```
indicator.head()
```

	text	label
0	Утилизация парникового газа, гибель мозаичнох...	0
1	О том, как механическое пианино прошло путь и...	0
2	1 октября 1881 года родился человек, фамилию ...	0
3	Майкл Курц — астрофизик из Гарвардского униве...	0
4	Информационно-сервисный портал Indicator.Ru с...	0

```
politru.head()
```

	text	label
0	Мы публикуем расшифровку лекции доктора биоло...	1
1	14 июня 2012 г. (четверг) в рамках проекта «Пу...	1
2	Мы публикуем расшифровку лекции одного из круп...	1
3	Мы публикуем авторизованную расшифровку публич...	1
4	Уважаемые коллеги!Приглашаем вас принять участ...	1

Обработка

- Токенизация
- Удаление стоп-слов (опционально)
- tf*idf векторизация (словные n-граммы (1,2))

Лекции vs статьи

	TF*IDF без стоп-слов, униграммы	TF*IDF без стоп-слов, униграммы + биграммы	TF*IDF, со стоп-словами, униграммы	TF*IDF, со стоп-словами, униграммы + биграммы
MultinomialNB	Acc: 0.8031 F: 0.7154	Acc: 0.9273 F: 0.7783	Acc: 0.8581 F: 0.5505	Acc: 0.9148 F: 0.7209
LogisticRegression	Acc: 0.9716 F: 0.9333	Acc: 0.9609 F: 0.8962	Acc: 0.9414 F: 0.8533	Acc: 0.9574 F: 0.8775
LinearSVC	Acc: 0.9840 F: 0.9626	Acc: 0.9840 F: 0.9592	Acc: 0.9822 F: 0.9599	Acc: 0.9822 F: 0.9523
DecisionTreeClassifier	Acc: 0.9822 F: 0.9593	Acc: 0.9911 F: 0.9777	Acc: 0.9609 F: 0.9197	Acc: 0.9840 F: 0.9596

Статьи vs новости

	TF*IDF без стоп-слов, униграммы	**TF*IDF без стоп-слов, униграммы + биграммы**	**TF*IDF, со стоп-словами, униграммы**	**TF*IDF, со стоп-словами, униграммы + биграммы**
MultinomialNB	Acc: 0.4867 F: 0.5889	Acc: 0.5079 F: 0.6736	Acc: 0.5079 F: 0.6736	Acc: 0.5052 F: 0.6701
LogisticRegression	Acc: 0.4153 F: 0.4622	Acc: 0.5 F: 0.6666	Acc: 0.4761 F: 0.6278	Acc: 0.4550 F: 0.5296
LinearSVC	Acc: 0.4179 F: 0.4622	Acc: 0.4682 F: 0.6666	Acc: 0.4682 F: 0.3888	Acc: 0.4444 F: 0.5296
DecisionTreeClassifier	Acc: 0.4973 F: 0.5410	Acc: 0.4682 F: 0.5037	Acc: 0.4894 F: 0.5415	Acc: 0.5079 F: 0.5303

Лекции vs новости

	TF*IDF без стоп-слов, униграммы	TF*IDF без стоп-слов, униграммы + биграммы	TF*IDF, со стоп-словами, униграммы	TF*IDF, со стоп-словами, униграммы + биграммы
MultinomialNB	Acc: 0.9723 F: 0.9545	Acc: 0.9600 F: 0.9392	Acc: 0.9585 F: 0.9364	Acc: 0.9539 F: 0.9278
LogisticRegression	Acc: 0.9585 F: 0.9298	Acc: 0.9600 F: 0.9346	Acc: 0.9754 F: 0.9626	Acc: 0.9508 F: 0.9215
LinearSVC	Acc: 0.9877 F: 0.9798	Acc: 0.9846 F: 0.9754	Acc: 0.9861 F: 0.9793	Acc: 0.9738 F: 0.9599
DecisionTreeClassifier	Acc: 0.9846 F: 0.9846	Acc: 0.9738 F: 0.9590	Acc: 0.9723 F: 0.9606	Acc: 0.9815 F: 0.9733

Лекции Arzamas vs лекции polit.ru

	TF*IDF без стоп-слов, униграммы	**TF*IDF без стоп-слов, униграммы + биграммы**	**TF*IDF, со стоп-словами, униграммы**	**TF*IDF, со стоп-словами, униграммы + биграммы**
MultinomialNB	Acc: 0.7431 F: 0.3529	Acc: 0.8171 F: 0.6299	Acc: 0.9455 F: 0.9066	Acc: 0.9727 F: 0.9580
LogisticRegression	Acc: 0.9883 F: 0.9818	Acc: 1.0 F: 1.0	Acc: 1.0 F: 1.0	Acc: 0.9727 F: 0.9580
LinearSVC	Acc: 1.0 F: 1.0	Acc: 1.0 F: 1.0	Acc: 1.0 F: 1.0	Acc: 0.9844 F: 0.9764
DecisionTreeClassifier	Acc: 0.9922 F: 0.9882	Acc: 0.9766 F: 0.9666	Acc: 0.9922 F: 0.9879	Acc: 0.9961 F: 0.9942