ИТМО

Разработка инструмента для прогнозирования популярности постов в социальных сетях с применением методов машинного обучения

Строкова Анастасия Владиславовна, таб. 370088

Цели и задачи

Цель: повышение качества постов в социальных сетях за счет разработки интеллектуальной системы для анализа популярности контента в официальной группе университета

Задачи:

- проведение парсинга постов официальной группы ИТМО в социальной сети «ВКонтакте» с помощью API;
- проведение первоначального анализа датасета, построение визуализаций для определения критерия, по которому пост будет отнесен к популярным;
- предобработка и токенизация датасета с использованием библиотеки nltk (в том числе удаление пунктуации, понижение регистра слов, удаление стоп-слов);
- векторизация датасета с использованием TF-IDF и Bag of Words, определение тональности текстов;
- обучение моделей на основе некоторых подходов классического машинного обучения, сравнение полученных моделей на основе метрик качества;
- подбор гиперпараметров и демонстрация полученных результатов.

Структура разрабатываемого инструмента

Парсинг датасета

С помощью **API** социальной сети **«ВКонтакте»** был сделан **парсинг постов** в официальной группе ИТМО

15627
строк

9
столбцов

Рисунок 1. Фрагмент датасета

ИТМО
Университет
70 873 подписчика

	id_group	id_post	data	description	title	text	views	likes	reposts
0	-94	57387	1716805478	NaN	NaN	А что если бы существовал гайд по поступлению,...	13878.0	82	52
1	-94	57891	1725354045	Восходящая звезда телевидения — Антонина Итмош...	Клип @itmoru	NaN	3680.0	16	12
2	-94	57890	1725351793	Ловите ту самую атмосферу в видео 😉\n\nА инсай...	ITMO CONF 2024	Грандиозно, экспертно, футуристично! \n \nВот ...	2190.0	22	6
3	-94	57877	1725105600	🛫 Гоооооооооо \n\n...на ITMO GO!\n\nВ начале...	Промо ITMO GO! x ITMOTION | 07.09	🛫 Гоооооооооо \n \n...на ITMO GO! \n \nВ нач...	8106.0	90	36
4	-94	57876	1725094801	NaN	NaN	[https://vk.com/public105042669| Сборная ИТМО #...	11422.0	94	7
...	...	...	...	...	...	...	...	...	...
15622	-94	5	1165260027	NaN	NaN	Всем привет!	NaN	0	0
15623	-94	4	1165259762	NaN	NaN	и почему тут никто не пишет?	NaN	1	0
15624	-94	3	1165258871	NaN	NaN	ну вот, нас уже трое =)	NaN	0	0
15625	-94	2	1165257425	NaN	NaN	давай пытаться =)	NaN	2	0
15626	-94	1	1165242465	NaN	NaN	Алён, предлагаю пытаться подключать народ к гр...	NaN	20	1

15627 rows × 9 columns

Таблица 1. Описание полей датасета

№	Название столбца	Описание столбца
1	id_group	Номер группы vk (в нашем случае 94)
2	id_post	Номер поста
3	data	Дата поста
4	description	Описание видео, ссылки, изображения
5	title	Заголовок видео, ссылки, изображения
6	text	Текст поста
7	views	Кол-во просмотров поста
8	likes	Кол-во лайков поста
9	reports	Кол-во репостов поста

https://vk.com/itmoru https://dev.vk.com/ru/method

Визуализация данных

Считаем пост популярным, если набрал **более 30-ти лайков**

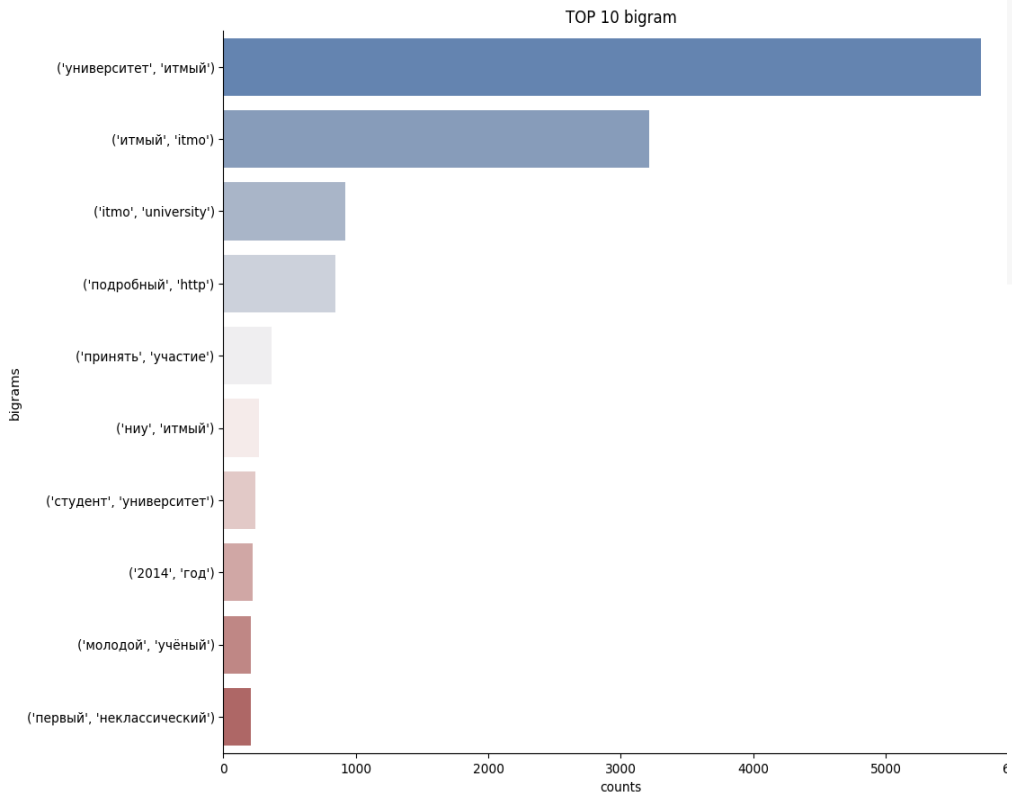


Рисунок 2. Популярные биграмы

```
# зависимость лайков от просмотров
x = data_posts['views']
y = data_posts['likes']

fig, ax = plt.subplots()
plt.scatter(x, y, label = 'views')

plt.legend()
plt.ylabel('likes')
plt.xlabel('views')
plt.title('Диаграмма рассеивания')
plt.show()
```

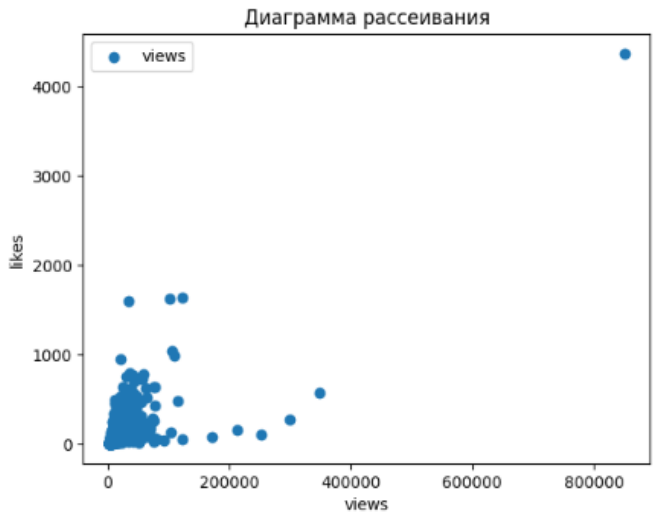


Рисунок 3. Зависимость лайков от просмотров

```
# зависимость репостов от лайков
x = data_posts['reposts']
y = data_posts['likes']

fig, ax = plt.subplots()
plt.scatter(x, y, label = 'reposts')

plt.legend()
plt.ylabel('likes')
plt.xlabel('reposts')
plt.title('Диаграмма рассеивания')
plt.show()
```

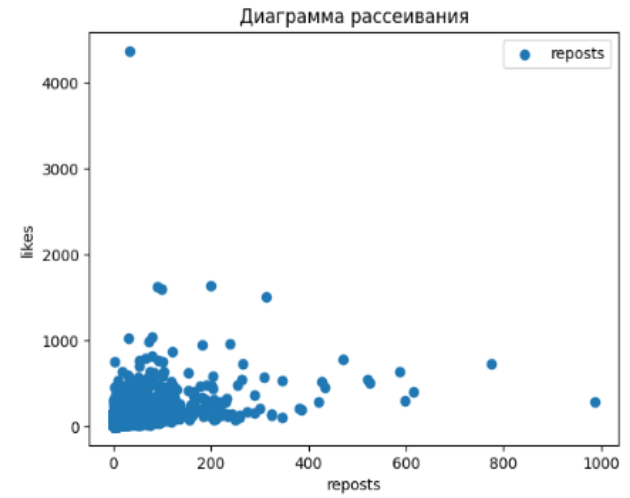


Рисунок 4. Зависимость лайков от репостов

Предобработка текста

Токенищация
Разделение текста на значимые единицы

Лемматизация
Приведение слова к начальной форме

Стемминг
Выделение основы слова

	text	tokenized_text	lemmatized_text	stemmed_text	popular
0	А что если бы существовал гайд по поступлению,...	существовал гайд поступлению родителей... итмо с...	существовать гайда поступление родителей... итмы...	существова гайд поступлен родителей... итм сбер ...	1
1	NaN	NaN	NaN	NaN	0
2	Грандиозно, экспертно, футуристично! Вот ка...	грандиозно экспертно футуристично прошел itmo ...	грандиозно экспертный футуристичный пройти itm...	грандиозн экспертн футуристичн прошел itmo con...	0
3	Гоооооооооо ...на ITMO GO! В начале т...	гоооооооооо ... itmo go начале тусовки 7 сен...	гоооооооооо ... itmo go начало тусовка 7 сен...	гоооооооооо ... itmo go начал тусовк 7 сентяб...	1
4	[ИТМО #KRONBARS] заняла третье место в Чемпио...	итмо kronbars заняла третье место чемпионате р...	итмый kronbars занять третий место чемпионат р...	итм kronbars заня трет мест чемпионат россии-2...	1
...	...	...	...	...	...
15622	Всем привет!	всем привет	весь привет	всем привет	0

Рисунок 5. Результат предобработки текстов постов

- Для продолжения работ был выбран столбец с **лемматизацией**, так как с этим подходом сохраняется наибольшее количество значимых признаков
- Текст был векторизирован на основе подходов **Bag of Words** и **TF-IDF**

Классификация текстов

Модели машинного обучения

MultinomialNB
Наивный Байес
sklearn.naive_bayes

RandomForestClassifier
Случайный лес
sklearn.ensemble

DecisionTreeClassifier
Случайный лес
sklearn.tree

KNeighborsClassifier
Метод ближайших соседей
sklearn.neighbors

LinearSVC
Метод опорных векторов
sklearn.svm

LogisticRegression
Логистическая регрессия
sklearn.linear_model

Метрики качества

- Accuracy
- Precision
- Recall
- F1

Размер выборок

- Обучающая выборка – 12501 строка
- Тестовая выборка – 3126 строк

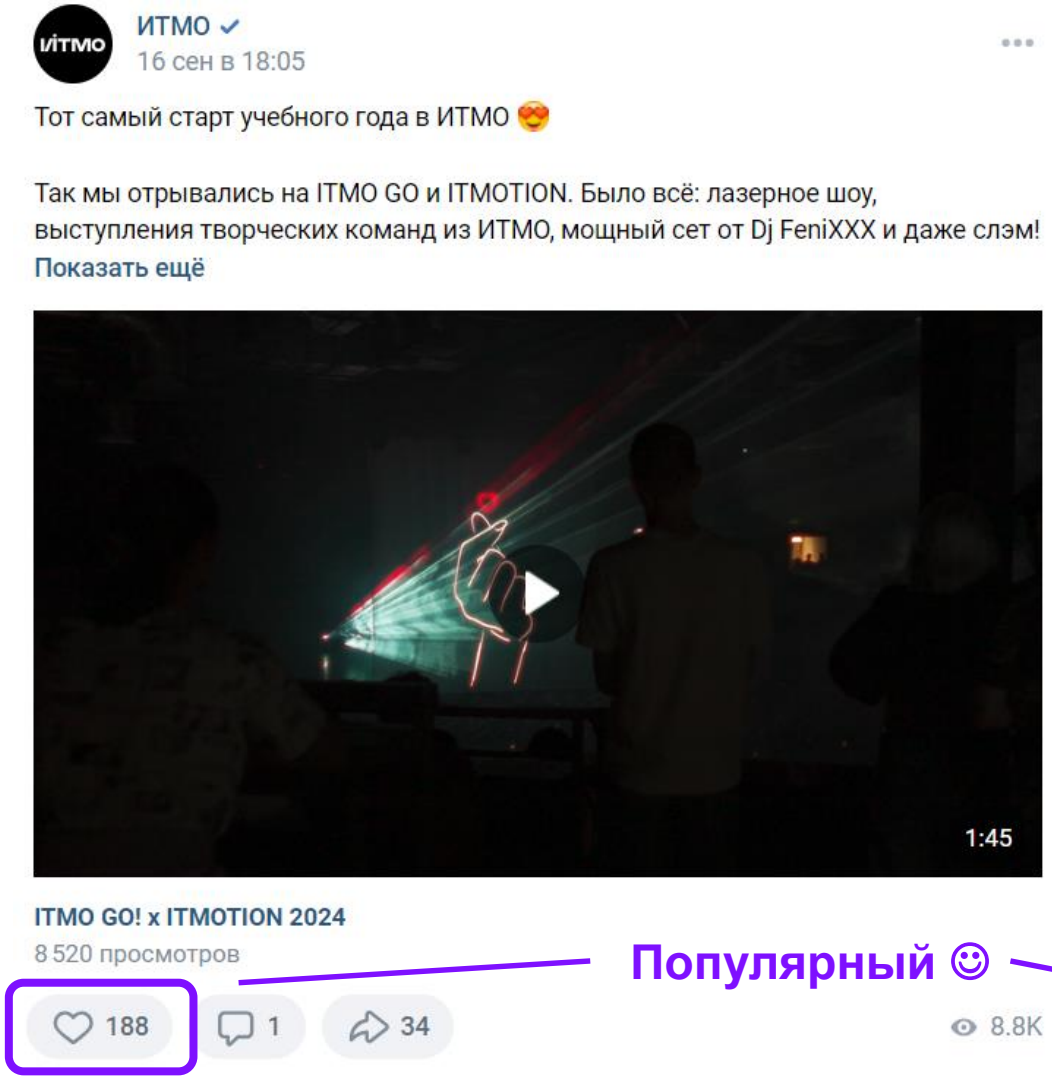
Результаты экспериментов

Модель	Accuracy	Precision	Recall	F1
Векторизация TF-IDF				
Наивный Байес	0.819578	0.576689	0.500498	0.452171
Случайный лес	**0.842290**	**0.815713**	**0.580185**	**0.595867**
Деревья решений	0.771913	0.630408	0.646763	0.637166
K-ближайших соседей	0.822457	0.752636	0.511956	0.476673
Метод опорных векторов	0.809661	0.701315	0.759182	0.720585
Логистическая регрессия	0.768074	0.672509	0.751147	0.687816
Векторизация Bag of Words				
Наивный Байес	0.821817	0.700148	0.706996	0.703439
Случайный лес	0.844210	0.817286	0.586900	0.605822
Деревья решений	0.786948	0.648707	0.661476	0.654384
K-ближайших соседей	0.811900	0.637709	0.556108	0.561612
Метод опорных векторов	0.823417	0.702139	0.706586	0.704307
Логистическая регрессия	**0.819898**	**0.705423**	**0.739783**	**0.719363**

Таблица 2. Результаты экспериментов со моделями машинного обучения

Демонстрация результата

ИТМО



Популярный ☺

1

```
[ ]  text = 'Тот самый старт учебного года в ИТМО 😍 Так мы отрывались на ITMO GO и ITMOTION.
```

2

```
# вывод на экран токенизированного текста
tokenized_text = tokenize_text(text)
print(tokenized_text)
```

```
самый старт учебного года итмо отрывались itmo go itmotion всё лазерное шоу выступления
```

3

```
# лемматизация текста
lemmatized_text = lemmatize_text(tokenized_text)
print(lemmatized_text)
```

```
самый старт учебный год итмый отрываться itmo go itmotion всё лазерный шоу выступление
```

4

```
# векторизация текста
X_new = vectorizer.transform([lemmatized_text])
print(X_new.toarray())
```

```
[[0 0 0 ... 0 0 0]]
```

5

```
# классификация нового текста
y_pred = clf_lr.predict(X_new)
print(f"Предсказанный класс: {y_pred}")
```

```
Предсказанный класс: [1]
```

Спасибо
за внимание!

ITMO more than a UNIVERSITY

stnastyast@yandex.ru
Telegram: @stnastyast