



Радар тенденций новостных статей

Цифровой прорыв 2022

Александр Широков

Задача

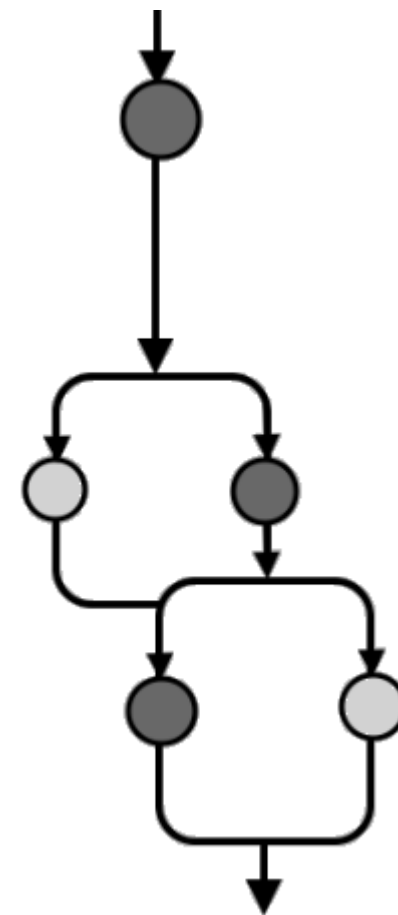
Предсказать 3 численные характеристики, которые в полной мере показывают популярность статьи:

- **Views** – количество просмотров
- **Full reads percent** – процент читателей полностью прочитавших статью
- **Depth** – объём прочитанного материала

$$\text{Total} = 0.4 \cdot R2_{\text{views}} + 0.3 \cdot R2_{\text{frp}} + 0.3 \cdot R2_{\text{depth}}$$

Этапы решения

1. Парсинг дополнительных данных с сайта РБК
2. Препроцессинг данных
3. Генерация признаков
4. Обучение модели
5. Предсказание



Парсинг

■ Дополнительные признаки

1. Категория новости (текст)
2. Краткая выжимка новости
3. Есть ли у новости картинка
4. Авторы новости (имена)
5. Тэги новости (текст)

1
Военная операция на Украине, 24 мая, 03:50 | 15 954 | Поделиться

Власти Херсонской области пообещали сделать русский язык государственным

Русский станет основным языком для «всех вопросов государственного значения», заявил назначенный в апреле замглавы военно-гражданской администрации Херсонской области. Украинский, по его словам, запрещать не будут 2



Фото: РИА Новости

4 Авторы Теги 5



Наталья Анисимова

Парсинг

- Дополнительные признаки
 6. Количество параграфов в тексте
 7. Количество ссылок на другие новости в тексте новости
 8. Текст новости
- Scrapy Parser

6 республики вправе устанавливать свои государственные языки, они употребляются наряду с государственным языком Российской Федерации).

7 **Власти Херсонской области объявили о введении бивалютной зоны**

Политика



8 Российское Минобороны отчиталось о взятии Херсона под контроль через несколько дней после начала военной операции на Украине. 15 марта был установлен контроль над всей территорией области. Позднее там создали военно-гражданскую администрацию; Стремоусова назначили замглавы в конце апреля.

document_id = <page_id> (628c22b89a79470e553f594b) + <session_id>

<https://www.rbc.ru/rbcfreenews/628c22b89a79470e553f594b>

Препроцессинг данных

- Поиск именованных сущностей в title с помощью **Natasha**
- Препроцессинг тэгов и авторов из строк в список
- Выделил категорию из title (где это было возможно)
- Заменял пустые значения на **Nan**
- Перевёл в **json**

```
{
  "document_id": "628c22b89a79470e553f594bQS5CqzXYRnmDdR2LaSreEw",
  "page_id": "628c22b89a79470e553f594b",
  "session": "QS5CqzXYRnmDdR2LaSreEw",
  "ctr": 1.598,
  "publish_date": "2022-05-24 00:50:55",
  "title": "Власти Херсонской области пообещали сделать русский язык государственным",
  "title_parsed": "Власти Херсонской области пообещали сделать русский язык государственным",
  "title_preprocessed": "Власти Херсонской области пообещали сделать русский язык государственным",
  "authors": null,
  "authors_parsed": ["Наталья Анисимова"],
  "tags": ["5433603acbb20f6e5def0cc5", "5409f420e063daa0f408b5a5", "545f38dacbb20fe3c1b1fcce", "621a3d0c9a794728d449ae5e"],
  "tags_parsed": ["Херсонская область", "русский язык", "украинский язык", "Военная операция на Украине"],
  "category": "5409f11ce063da9c8b588a12",
  "category_parsed": "Военная операция на Украине",
  "category_from_title": null,
  "news_text_parsed": "Русский язык наравне ...",
  "news_text_overview_parsed": "Русский станет основным языком ...",
  "news_amount_of_paragraphs_parsed": 9,
  "news_amount_of_inline_items_parsed": 2,
  "news_inline_titles_parsed": [
    "Власти Херсонской области объявили о введении бивалютной зоны",
    "Хуснуллин заявил о «большой перспективе Херсона в российской семье»"
  ],
  "news_has_image_parsed": 1,
  "news_image_title_parsed": "Фото: РИА Новости",
  "keywords_parsed": ["Российский", "Херсонский"],
  "PER_TITLE": null,
  "ORG_TITLE": ["Херсонская область"],
  "LOC_TITLE": null,
}
```

Генерация признаков

- **TF-IDF** для авторов, тэгов и title
- Временные признаки
- **Длина текста**, количество слов в тексте, обработка **ctr**
- ... Много-много других гипотез, группировок, которые почему-то не работали
- Зато написал удобный **Pipeline**

```
FeatureGenerator = Compose(  
    transforms=[  
        DatetimeTransformer(),  
        TitleTransformer(),  
        CTRTransformer(),  
        TagsTransformer(),  
        AuthorsTransformer(),  
        TextTransformer(),  
        NatashaTextTransformer(),  
        CategoryTransformer(),  
        FeatureSelector()  
    ]  
)  
train_features = FeatureGenerator(data=train_data, mode='train')  
test_features = FeatureGenerator(data=test_data, mode='test')
```

Обучение модели

- Обработал выбросы в таргете
- Логарифмировал таргет
- **LightGBM Regressor**
- Валидация: KFold – 3 splits
- Усреднение в ответе

```
lgbm_regressor = lgb.LGBMRegressor(  
    objective='rmse',  
    random_state=33,  
    early_stopping_round=3000,  
    n_estimators=500,  
    subsample=1,  
    colsample_bytree=0.95,  
    learning_rate=0.09,  
    max_depth=-1,  
    verbose=-1  
)
```



LightGBM

Предсказание

- Медианное усреднение ответов трех регрессоров
- Специальная обработка для тех page_id, которые есть **и в train и в test**

Итоги

- Public: 0.7619
- Валидация
 - **Views** – 71.3 ± 0.3
 - **Depth** – 55.1 ± 0.4
 - **Full reads percent** – 84.2 ± 0.3

Спасибо за внимание!

[Ссылка на репозиторий](#)