NLP in stock market prediction

by Степанов Семен

Описание задачи

Реализовать сервис, который может

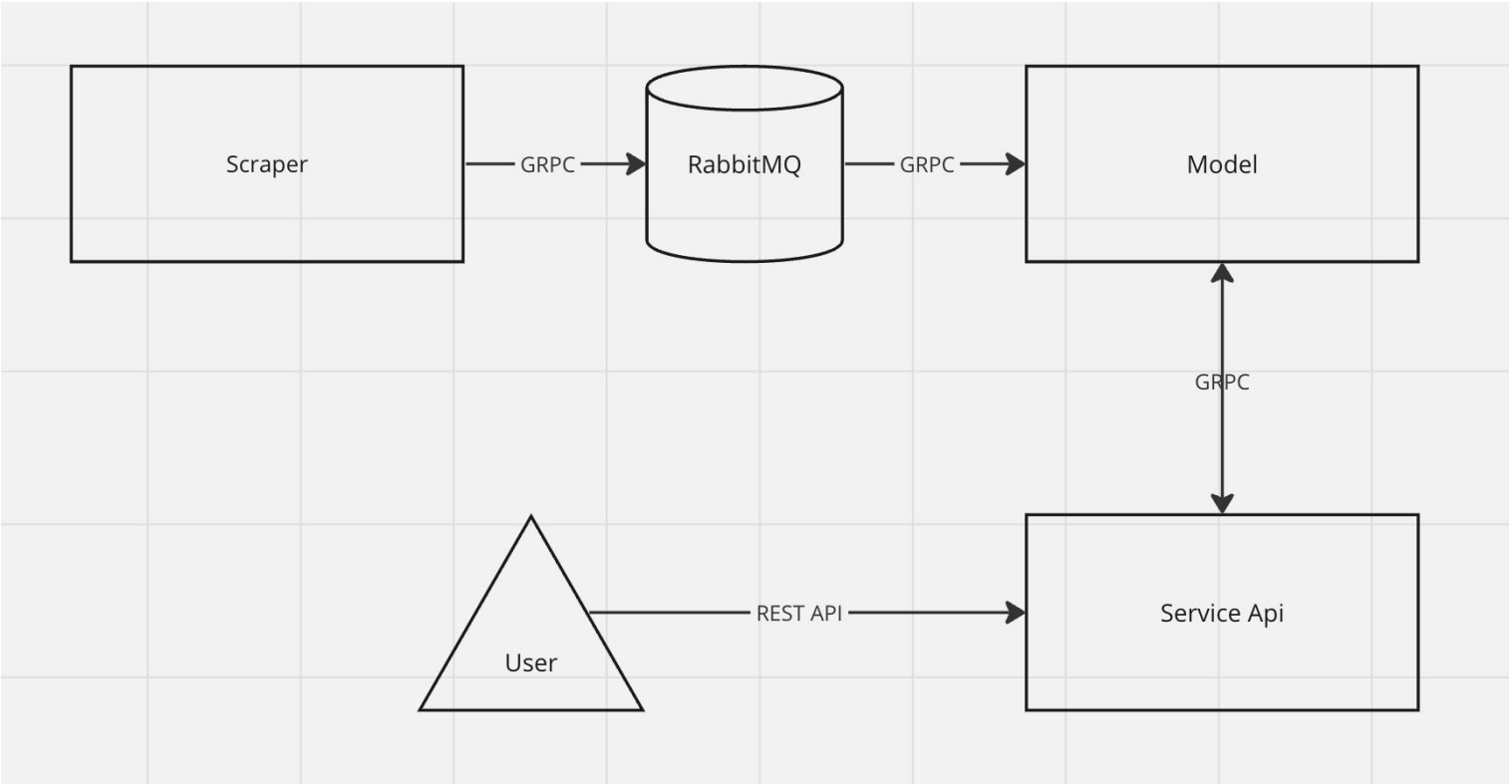
- Отвечать при помощи API на вопрос: вырастет/упадет/не изменится стоимости акции(индекса) для приложенной новости

- Держать модель задеплоиной а рамках Yandex.Cloud с 24/7 возможностью для доступа к API

- Уметь дообучать модель на батчах размеченных данных {Новость + Как повлияла на стоимости акции}

- Реализовать сервис для дообучения. Некий realtime Scraper новостных сайтов + stock price для создания размеченных данных

Визуал архитектуры



Текущая подзадача

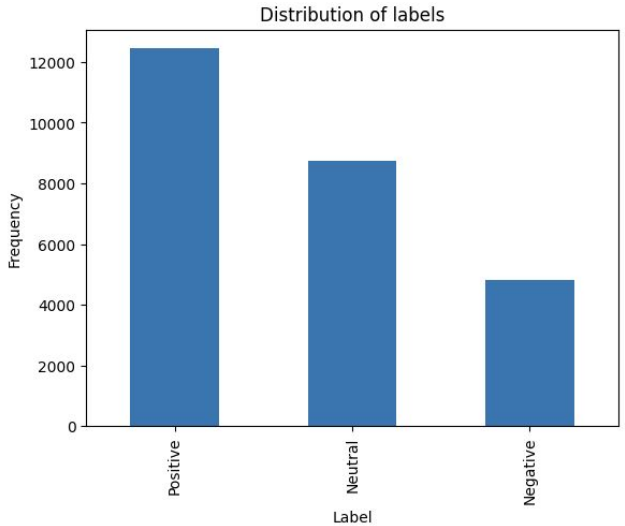
Реализовать классификатор для новостей, который по headline предсказывает класс (Positive, Neutral, Negative)

- Найти данные и произвести EDA
- Обучить модель

Данные

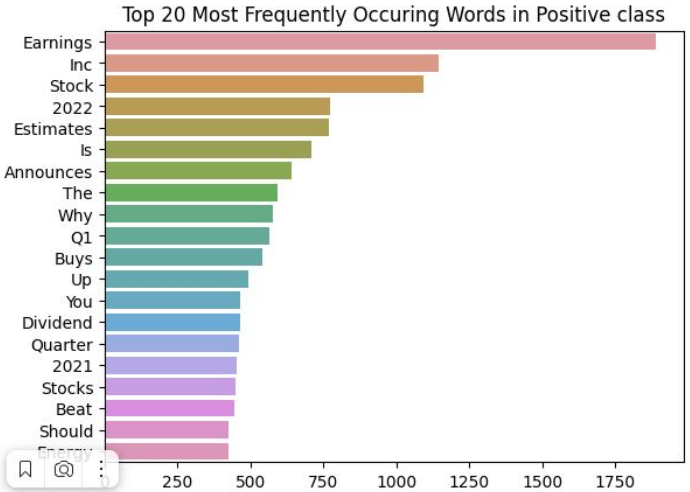
В качестве датасета для обучения был взят датасет с kaggle, содержащий новостные заголовки и “лэйблы” классов

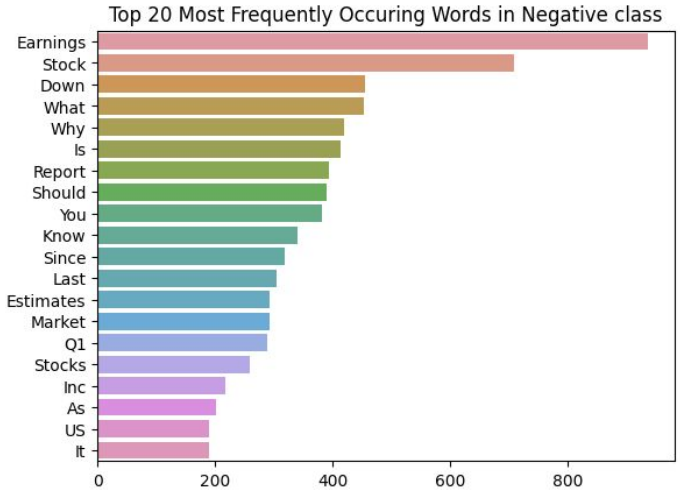
```
Data columns (total 3 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Unnamed: 0   26000 non-null  int64
 1   headline     26000 non-null  object
 2   label        26000 non-null  object
```



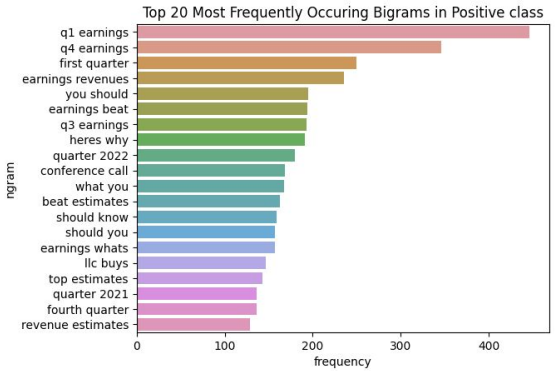
EDA

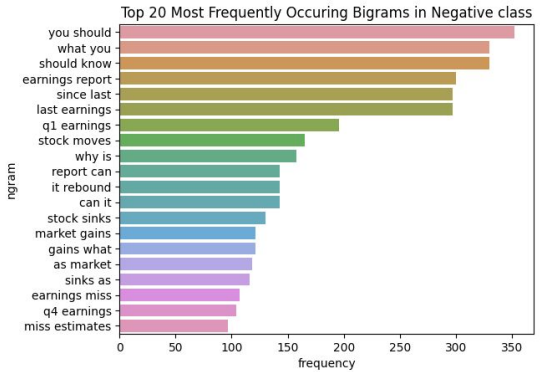
- Провалидировать что датасет пригоден для обучения модели
- Выявить тривиальные зависимости
- Подготовить датасет к использованию моделью

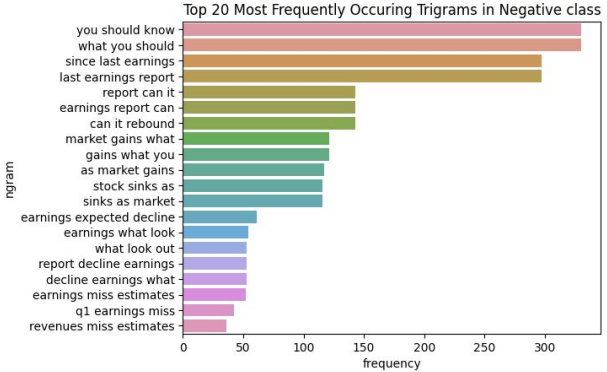




Биграммы и триграммы







Модель

Baseline - классификатор, возвращающий всегда Positive класс

```
Accuracy:  0.4626923076923077
Precision:  0.21408417159763313
Recall:  0.4626923076923077
```

LogisticRegression из sklearn

- TfidfVectorizer для предобработки данных (без учета н-грамм)

```
Accuracy:  0.7240384615384615
Precision:  0.7305690810198644
Recall:  0.7240384615384615
```

```
Precision:  [0.80136054 0.73533163 0.6983086 ]
Recall:  [0.60472279 0.63351648 0.84081463]
```

```
F2 Score:  0.7209434578079487
ROC AUC:  0.8655062270037884
```

Тюнинг данных

- Баланс классов
- Нормальная форма текстов
- Андерсемплинг и оверсемплинг

Андерсемплинг + нормальная форма текста

```
Accuracy:  0.7266736038848421
Precision:  0.7262862094077107
Recall:  0.7266736038848421
```

```
Precision:  [0.79452055 0.70974808 0.66983122]
Recall:  [0.80715706 0.69304813 0.67409766]
```

```
F2 Score:  0.7265630347217633
ROC AUC:  0.8877699404046062
```

Оверсемплинг + нормальная форма текста + нграммы

```
Accuracy:  0.8177801579863435
Precision:  0.8277837389694488
Recall:  0.8177801579863435
```

```
Precision:  [0.94465097 0.83285714 0.70284939]
Recall:  [0.90646056 0.70382294 0.84193417]
```

```
F2 Score:  0.8169326233043109
ROC AUC:  0.919675010666111
```

Пример ответов модели

Модель задеплоина на fastapi сервере

curl -X POST "http://158.160.28.16:8000/predict" -H "accept: application/json" -H "Content-Type: application/json" -d "{\"title\":\"US government approves bitcoin etf\"}"

{"prediction":"Positive"}

Планы на будущее

- Провести валидацию на обновленных данных
- Бустинг в классификаторе
- DL для классификатора
- Решать задачу классификации для батча новостей
- Написать инфраструктуру
- Реализовать scraper и систему дообучения