§6 Программы

Импорт нужных модулей:

```
import pandas as pd
import re
import numpy as np
import matplotlib.pyplot as plt
import keras as kr
from gensim.models import Word2Vec
from gensim.test.utils import get_tmpfile
```

Группировка заголовков по дням:

```
data = pd.read_csv("data19992020.csv")
data = data.groupby('date')['title'].apply(lambda text: ' '.join(text))
data = pd.DataFrame(data, columns = ['date', 'title'])
data['date'] = data.index
data.index = range(data.shape[0])
```

Функция создания маркеров тренда:

```
def trend_marker(x):
    if x>0:
        return 1
    elif x<0:
        return 0
    else:
        return 2
```

Совмещение датасета новостей с датасетом курса доллара, создание маркеров для задачи классификации:

```
price = pd.read_csv('/home/bogdanrasengan/le_diploma/rubdoldata.csv')
data = data.merge(price, on='date')
data["trend"]= data["change%"].map(trend_marker)
data = data.drop(columns=['date', 'price', 'open', 'max', 'min', 'change%'])
data.trend = data.trend.shift(-1)
data = data.dropna()
data = data[data.trend != 2]
```

Функция токенизации текста:

```
def standartize(text):
    reg = re.compile('[^А-Яа-яа-zA-Z ]')
    text = reg.sub(' ', text).split()
    for i in range(len(text)):
        if len(text[i]) < 3:
            text[i] = 0
    while 0 in text:
        text.remove(0)
    return text
```

Токенизация и padding:

```
word_vecs = list(data.title.map(standartize))
lens = []
for i in word_vecs:
    lens.append(len(i))
for i in word_vecs:
    for j in range(max(lens) - len(i)):
        i.append('.')
```

Обучение Word2Vec модели:

```
model = Word2Vec(tokenized_titles, size=64, window=5, min_count=1, workers=4)
fname = get_tmpfile(f"/home/bogdanrasengan/le_diploma/beta0.04/w2v_64_5")
model.save(fname)
```

Замена слов на вектора в заголовках:

```
for title in word_vecs:
    for i in range(len(title)):
        title[i] = model.wv.get_vector(title[i])
```

Разбиение на тренировочную и тестовую часть:

```
X_train, y_train = np.array(word_vecs[:3600]), np.array(data.trend[:3600])
X_test, y_test = np.array(word_vecs[3600:]), np.array(data.trend[3600:])
```

Настройка и компиляция RNN модели:

```
model = kr.models.Sequential()
model.add(kr.layers.Bidirectional(kr.layers.recurrent.LSTM(16)))
model.add(kr.layers.core.Dense(1, activation='sigmoid'))
model.compile(optimizer='adam', loss='binary_crossentropy',
metrics=['accuracy'])
```

Запуск обучения:

```
history = model.fit(X_train, y_train, epochs=20, batch_size=200,
validation_data=(X_test, y_test))
```

Функция построения графиков обучения:

```
def plot_his(key, history):
    fig, (ax1, ax2) = plt.subplots(ncols=2, nrows = 1, figsize = (18, 6))
    ax1.plot(range(len(history.history[key])), history.history[key], c='r')
    ax1.set_xlabel('number of epochs')
    ax1.set_ylabel(key)
    ax2.plot(range(len(history.history[f'val_{key}'])),
history.history[f'val_{key}'], c='r')
    ax2.set_xlabel('number of epochs')
    ax2.set_ylabel(f'val_{key}')
    plt.savefig(f'{key}_rnn_w2v')
```

Построение графиков обучения(полные графики в приложении 1)

```
plot_his('accuracy',history)
plot_his('loss',history)
```

Приложения

Приложение 1 – Изначальный датасет новостей

	url	title	topic	tags	date
0	https://lenta.ru/news/1914/09/16/hungarnn/	1914. Русские войска вступили в пределы Венгрии	Библиотека	Первая мировая	1914/09/16
1	https://lenta.ru/news/1914/09/16/lermontov/	1914. Празднование столетия М.Ю. Лермонтова от...	Библиотека	Первая мировая	1914/09/16
2	https://lenta.ru/news/1914/09/17/nesteroff/	1914. Das ist Nesteroff!	Библиотека	Первая мировая	1914/09/17
3	https://lenta.ru/news/1914/09/17/bulldogn/	1914. Бульдог-гонец под Льежем	Библиотека	Первая мировая	1914/09/17
4	https://lenta.ru/news/1914/09/18/zver/	1914. Под Люблином пойман швабский зверь	Библиотека	Первая мировая	1914/09/18
...	...	...	...	...	...
802652	https://lenta.ru/news/2020/04/23/temnotaktar/	Тактаров захотел разобраться с Емельяненко в т...	NaN	Бокс и ММА	2020/04/23
802653	https://lenta.ru/news/2020/04/23/otvetil/	Российский губернатор ответил на вопросы о пое...	NaN	Политика	2020/04/23
802654	https://lenta.ru/news/2020/04/23/danger/	Названы самые опасные гаджеты на карантине	NaN	Гаджеты	2020/04/23
802655	https://lenta.ru/news/2020/04/23/hunger/	Рекордная безработица обернулась массовым голо...	NaN	Госэкономика	2020/04/23
802656	https://lenta.ru/news/2020/04/23/no_help/	Помощь по зарплатам для россиян оказалась под ...	NaN	Госэкономика	2020/04/23

802657 rows × 5 columns

Приложение 2 – Изначальный датасет курса доллар-рубль

	date	price	open	max	min	change%
0	2001/09/14	29.4730	29.3500	29.4800	29.3500	-0.00
1	2001/09/17	29.4610	29.4600	29.4730	29.4400	-0.04
2	2001/09/18	29.4600	29.4450	29.4750	29.4400	-0.00
3	2001/09/19	29.4530	29.4600	29.4810	29.4300	-0.02
4	2001/09/20	29.4300	29.4550	29.4760	29.4030	-0.08
...	...	...	...	...	...	...
4790	2020/03/17	75.4362	74.6596	75.4751	73.4589	1.04
4791	2020/03/18	80.8692	75.4362	80.8700	75.4362	7.20
4792	2020/03/19	79.1566	80.8692	81.6615	78.6405	-2.12
4793	2020/03/20	79.9236	79.1566	80.1497	77.4478	0.97
4794	2020/03/23	79.5936	79.9236	81.2429	79.3858	-0.41

4795 rows × 6 columns

Приложение 3 – Датасет сгруппированных по датам новстей (с помощью конкатенации).

	date	title
0	1999/08/31	Космонавты сомневаются в надежности "Мира" Взр...
1	1999/09/01	США заплатили Китаю 4,5 миллиона долларов за "...
2	1999/09/02	Боевики вытеснены из дагестанского села Карама...
3	1999/09/03	В Гонконге разгромлен синдикат по производству...
4	1999/09/06	"Mabetex" подает в суд на "Corrierre della Ser...
...	...	...
7510	2020/04/19	В США рассказали о головной боли из-за потерян...
7511	2020/04/20	Создана модель развития коронавируса Более 10 ...
7512	2020/04/21	Раскрыта опасность средств защиты от коронавир...
7513	2020/04/22	В Москве умерли 28 пациентов с коронавирусом Т...
7514	2020/04/23	Оценены шансы коронавируса стать худшей пандем...

7515 rows × 2 columns

Приложение 4 – Объединенный датасет новостей и курса

	date	title	price	open	max	min	change%
0	2001/09/14	Конгресс увеличил сумму, выделяемую на ликвида...	29.4730	29.3500	29.4800	29.3500	-0.00
1	2001/09/17	Итальянские войска не будут помогать США в бор...	29.4610	29.4600	29.4730	29.4400	-0.04
2	2001/09/18	Дезертир продал автомат раньше, чем его поймал...	29.4600	29.4450	29.4750	29.4400	-0.00
3	2001/09/19	Правоохранительные органы США создают общенаци...	29.4530	29.4600	29.4810	29.4300	-0.02
4	2001/09/20	Телекомпанию BBC обвиняют в антиамериканских н...	29.4300	29.4550	29.4760	29.4030	-0.08
...	...	...	...	...	...	...	...
4784	2020/03/17	Дети оказались скрытыми разносчиками коронавир...	75.4362	74.6596	75.4751	73.4589	1.04
4785	2020/03/18	Названо слабое место коронавируса Россия модер...	80.8692	75.4362	80.8700	75.4362	7.20
4786	2020/03/19	В Италии зафиксировали рекордное число жертв к...	79.1566	80.8692	81.6615	78.6405	-2.12
4787	2020/03/20	Названа вероятность заразиться коронавирусом о...	79.9236	79.1566	80.1497	77.4478	0.97
4788	2020/03/23	В США рассказали о превосходстве украинских Т-...	79.5936	79.9236	81.2429	79.3858	-0.41

4789 rows × 7 columns

Приложение 5 – Датасет после удаления ненужных признаков и составления нового признака основанного на изменении курса

	title	trend
0	Конгресс увеличил сумму, выделяемую на ликвида...	0.0
2	Дезертир продал автомат раньше, чем его поймал...	0.0
3	Правоохранительные органы США создают общенаци...	0.0
4	Телекомпанию BBC обвиняют в антиамериканских н...	0.0
5	Президентом Эстонии стал Арнольд Рюйтель В Бел...	1.0
...	...	...
4783	В США объяснили создание российского «Бурака-М...	1.0
4784	Дети оказались скрытыми разносчиками коронавир...	1.0
4785	Названо слабое место коронавируса Россия модер...	0.0
4786	В Италии зафиксировали рекордное число жертв к...	1.0
4787	Названа вероятность заразиться коронавирусом о...	0.0

4546 rows × 2 columns

Приложение 6 - Датасет после применения алгоритма padding и word2vec

	0	1	2	3	4	...	1194	1195	1196	1197	trend
0	[-0.8669171, 0.8682023, 0.2507141, 0.36427072,...	[-1.4618559, 0.5298251, 0.44858122, -1.4948398...	[-0.65548396, 0.14827566, 0.2373754, -0.984580...	[-0.005379057, -0.010677823, -0.012190706, 0.0...	[0.06222183, 0.5415078, -0.21569674, -0.301038...	...	[-1.3246988, 2.5479429, -0.2338264, -1.1839666...	[-1.3246988, 2.5479429, -0.2338264, -1.1839666...	[-1.3246988, 2.5479429, -0.2338264, -1.1839666...	[-1.3246988, 2.5479429, -0.2338264, -1.1839666...	0.0
1	[0.21244618, 0.22267425, -0.09020767, 0.067462...	[-1.0552956, 0.668358, 0.51148856, -0.7702768,...	[0.15998423, 0.34161103, 0.06646128, -0.208416...	[0.2956615, 0.48339683, -0.13032489, -0.886134...	[-1.3278818, 0.62243444, 0.17482728, -0.660720...	...	[-1.3246988, 2.5479429, -0.2338264, -1.1839666...	[-1.3246988, 2.5479429, -0.2338264, -1.1839666...	[-1.3246988, 2.5479429, -0.2338264, -1.1839666...	[-1.3246988, 2.5479429, -0.2338264, -1.1839666...	0.0
2	[0.023477033, 0.06973852, 0.00039949044, -0.01...	[0.17222373, 0.4347452, 0.09182871, -0.1341721...	[-1.6987976, -0.11039375, -3.4531517, -3.23284...	[-0.11533699, 0.30058292, 0.021350592, -0.5064...	[0.09124328, 0.07900997, 0.006078894, -0.03068...	...	[-1.3246988, 2.5479429, -0.2338264, -1.1839666...	[-1.3246988, 2.5479429, -0.2338264, -1.1839666...	[-1.3246988, 2.5479429, -0.2338264, -1.1839666...	[-1.3246988, 2.5479429, -0.2338264, -1.1839666...	0.0
...	...	...	...	...	...	...	...	...	...	...	...
4543	[-0.029249595, -0.1318893, -0.27465707, -2.106...	[0.026732022, 0.0014417237, -0.023666881, -0.0...	[0.682403, 0.59616715, -0.44613975, -1.0629076...	[0.027396997, 0.05035559, -0.0302479, -0.09855...	[2.1295195, -0.81480193, -0.74341094, -4.28264...	...	[-1.3246988, 2.5479429, -0.2338264, -1.1839666...	[-1.3246988, 2.5479429, -0.2338264, -1.1839666...	[-1.3246988, 2.5479429, -0.2338264, -1.1839666...	[-1.3246988, 2.5479429, -0.2338264, -1.1839666...	0.0
4544	[-0.15648688, 1.4578235, -1.5931102, 0.4908076...	[-0.39107972, 0.0969286, 0.34740672, -0.804135...	[-0.9091846, 0.011080605, 0.18596154, -1.45695...	[-1.2893343, -0.09006908, -0.009314546, -2.595...	[-0.30793485, 1.1936431, -2.0125093, -0.625150...	...	[-1.3246988, 2.5479429, -0.2338264, -1.1839666...	[-1.3246988, 2.5479429, -0.2338264, -1.1839666...	[-1.3246988, 2.5479429, -0.2338264, -1.1839666...	[-1.3246988, 2.5479429, -0.2338264, -1.1839666...	1.0
4545	[0.47848737, -1.9019147, -0.17539033, -4.53218...	[-0.064552136, 0.19435416, -0.07428323, -0.371...	[0.007315976, 0.0291767, 0.0021893813, -0.0216...	[0.013187584, 0.05945578, -0.025966123, -0.033...	[-0.5837957, 0.9114325, 0.9087801, -0.4013997,...	...	[-1.3246988, 2.5479429, -0.2338264, -1.1839666...	[-1.3246988, 2.5479429, -0.2338264, -1.1839666...	[-1.3246988, 2.5479429, -0.2338264, -1.1839666...	[-1.3246988, 2.5479429, -0.2338264, -1.1839666...	0.0

4546 rows × 1199 columns

Приложение 7 – Пример вектора отдельного слова из датасета векторов новостей.

```
array([-0.8669171 ,  0.8682023 ,  0.2507141 ,  0.36427072, -0.18893325,
       -1.8522432 , -0.8941062 ,  1.0795678 , -0.30273834,  0.5011146 ,
       -0.6751242 ,  1.1204183 , -0.53109866, -0.15054038,  0.9508042 ,
       -1.9429744 , -2.7447126 , -0.7497335 ,  0.4777511 ,  0.24064848,
       -2.3295817 , -1.6252114 ,  0.06299474,  0.38160607, -1.0002854 ,
       -1.3564508 , -0.92572844, -0.07960084,  0.51667005, -0.905476  ,
       -0.10368816, -0.564028  , -0.15176238, -0.5630232 , -1.5366814 ,
       -0.53910893,  0.13967085,  0.28928864,  0.10385697, -1.6256838 ,
        0.9061764 ,  0.9676117 ,  0.5307054 , -1.3096952 , -0.43073317,
       -0.54612714,  0.01183307,  0.5197771 , -0.9978606 , -1.1482041 ,
       -0.18571675,  1.2037708 , -0.32009634, -0.5649487 , -1.502912  ,
       -1.0669935 , -0.6776898 ,  0.1786546 ,  0.16513363,  1.9141737 ,
        0.06799839,  1.0516206 ,  0.2851056 ,  0.21277285], dtype=float32)
```