Московский государственный технический университет им. Н.Э. Баумана
Кафедра «Системы обработки информации и управления»

Лабораторная работа №5
по дисциплине
«Методы машинного обучения»

Выполнил:
студент группы ИУ5-23М
Дин Но

Москва — 2021 г.

1. Цель лабораторной работы

Изучение методов предобработки текстов.

2. Задание

Для произвольного предложения или текста решите следующие задачи:
Токенизация.
Частеречная разметка.
Лемматизация.
Выделение (распознавание) именованных сущностей.
Разбор предложения.

3. Ход выполнения работы

```
import numpy as np
import pandas as pd
import os

df1=pd.read_csv('/content/drive/MyDrive/fake_and_real_news/Fake.csv')
df2=pd.read_csv('/content/drive/MyDrive/fake_and_real_news/True.csv')
df1['Target']=1
df2['Target']=0
df=pd.concat([df1,df2],axis=0)
df['original'] = df['text'] + ' ' + df['title']
df
```

	title	text	subject	date	Target	original
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	1	Donald Trump just couldn t wish all Americans ...
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	1	House Intelligence Committee Chairman Devin Nu...
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	1	On Friday, it was revealed that former Milwauk...
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	1	On Christmas day, Donald Trump announced that ...
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	1	Pope Francis used his annual Christmas Day mes...
...	...	...	...	...	...	...
21412	'Fully committed' NATO backs new U.S. approach...	BRUSSELS (Reuters) – NATO allies on Tuesday we...	worldnews	August 22, 2017	0	BRUSSELS (Reuters) – NATO allies on Tuesday we...
21413	LexisNexis withdrew two products from Chinese ...	LONDON (Reuters) – LexisNexis, a provider of l...	worldnews	August 22, 2017	0	LONDON (Reuters) – LexisNexis, a provider of l...
21414	Minsk cultural hub becomes haven from authorities	MINSK (Reuters) – In the shadow of disused Sov...	worldnews	August 22, 2017	0	MINSK (Reuters) – In the shadow of disused Sov...
21415	Vatican upbeat on possibility of Pope Francis ...	MOSCOW (Reuters) – Vatican Secretary of State ...	worldnews	August 22, 2017	0	MOSCOW (Reuters) – Vatican Secretary of State ...
21416	Indonesia to buy $1.14 billion worth of Russia...	JAKARTA (Reuters) – Indonesia will buy 11 Sukh...	worldnews	August 22, 2017	0	JAKARTA (Reuters) – Indonesia will buy 11 Sukh...

44898 rows × 6 columns

```
df=df.drop(['title','subject','date'],axis=1)
df
```

	text	Target	original
0	Donald Trump just couldn t wish all Americans ...	1	Donald Trump just couldn t wish all Americans ...
1	House Intelligence Committee Chairman Devin Nu...	1	House Intelligence Committee Chairman Devin Nu...
2	On Friday, it was revealed that former Milwauk...	1	On Friday, it was revealed that former Milwauk...
3	On Christmas day, Donald Trump announced that ...	1	On Christmas day, Donald Trump announced that ...
4	Pope Francis used his annual Christmas Day mes...	1	Pope Francis used his annual Christmas Day mes...
...	...	...	...
21412	BRUSSELS (Reuters) – NATO allies on Tuesday we...	0	BRUSSELS (Reuters) – NATO allies on Tuesday we...
21413	LONDON (Reuters) – LexisNexis, a provider of l...	0	LONDON (Reuters) – LexisNexis, a provider of l...
21414	MINSK (Reuters) – In the shadow of disused Sov...	0	MINSK (Reuters) – In the shadow of disused Sov...
21415	MOSCOW (Reuters) – Vatican Secretary of State ...	0	MOSCOW (Reuters) – Vatican Secretary of State ...
21416	JAKARTA (Reuters) – Indonesia will buy 11 Sukh...	0	JAKARTA (Reuters) – Indonesia will buy 11 Sukh...

44898 rows × 3 columns

```
import nltk
nltk.download('wordnet')
from nltk.stem import PorterStemmer
from nltk.tokenize import sent_tokenize, word_tokenize
import re
import string
from keras.preprocessing.text import Tokenizer
from nltk.stem import WordNetLemmatizer

def custom_preprocessor(text):
    text = text.lower()
    text = re.sub('\[.*?\]', '', text)
    text = re.sub("\\W"," ",text)
    text = re.sub('https?://\S+|www\.\S+', '', text)
    text = re.sub('<.*?>+', '', text)
    text = re.sub('[%s]' % re.escape(string.punctuation), '', text)
    text = re.sub('\n', '', text)
    text = re.sub('\w*\d\w*', '', text)
    return text

df
```

```
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data]   Unzipping corpora/wordnet.zip.
```

	text	Target	original
0	Donald Trump just couldn t wish all Americans ...	1	Donald Trump just couldn t wish all Americans ...
1	House Intelligence Committee Chairman Devin Nu...	1	House Intelligence Committee Chairman Devin Nu...
2	On Friday, it was revealed that former Milwauk...	1	On Friday, it was revealed that former Milwauk...
3	On Christmas day, Donald Trump announced that ...	1	On Christmas day, Donald Trump announced that ...
4	Pope Francis used his annual Christmas Day mes...	1	Pope Francis used his annual Christmas Day mes...
...	...	...	...
21412	BRUSSELS (Reuters) – NATO allies on Tuesday we...	0	BRUSSELS (Reuters) – NATO allies on Tuesday we...
21413	LONDON (Reuters) – LexisNexis, a provider of l...	0	LONDON (Reuters) – LexisNexis, a provider of l...
21414	MINSK (Reuters) – In the shadow of disused Sov...	0	MINSK (Reuters) – In the shadow of disused Sov...
21415	MOSCOW (Reuters) – Vatican Secretary of State ...	0	MOSCOW (Reuters) – Vatican Secretary of State ...
21416	JAKARTA (Reuters) – Indonesia will buy 11 Sukh...	0	JAKARTA (Reuters) – Indonesia will buy 11 Sukh...

44898 rows × 3 columns

```
lemmatizer = WordNetLemmatizer()

def lemmatization(text):
    lemmas = []
    for word in text.split():
        lemmas.append(lemmatizer.lemmatize(word))
    return " ".join(lemmas)

df['text']=df['text'].apply(custom_preprocessor)
df['original'].apply(lemmatization)
```

```
0         Donald Trump just couldn t wish all Americans ...
1         House Intelligence Committee Chairman Devin Nu...
2         On Friday, it wa revealed that former Milwauke...
3         On Christmas day, Donald Trump announced that ...
4         Pope Francis used his annual Christmas Day mes...
                                ...
21412     BRUSSELS (Reuters) - NATO ally on Tuesday welc...
21413     LONDON (Reuters) - LexisNexis, a provider of l...
21414     MINSK (Reuters) - In the shadow of disused Sov...
21415     MOSCOW (Reuters) - Vatican Secretary of State ...
21416     JAKARTA (Reuters) - Indonesia will buy 11 Sukh...
Name: original, Length: 44898, dtype: object
```

Список литературы

[1] Гапанюк Ю. Е. Лабораторная работа «Линейные модели, SVM и деревья решений»[Электронный ресурс] // GitHub. — 2019. — Режим доступа: https://github.com/ugapanyuk/ml_course/wiki/LAB_TREES (дата обращения: 19.04.2019).
[2] Team The IPython Development. IPython 7.3.0 Documentation [Electronic resource] //Read the Docs. — 2019. — Access mode: https://ipython.readthedocs.io/en/stable/ (online; accessed: 20.02.2019).
[3] Waskom M. seaborn 0.9.0 documentation [Electronic resource] // PyData. — 2018. —Access mode: https://seaborn.pydata.org/ (online; accessed: 20.02.2019).
[4] pandas 0.24.1 documentation [Electronic resource] // PyData. — 2019. — Access mode:http://pandas.pydata.org/pandas-docs/stable/ (online; accessed: 20.02.2019).
[5] dronio. Solar Radiation Prediction [Electronic resource] // Kaggle. — 2017. — Accessmode: https://www.kaggle.com/dronio/SolarEnergy (online; accessed: 18.02.2019).
[6] Chrétien M. Convert datetime.time to seconds [Electronic resource] // Stack Overflow.— 2017. — Access mode: https://stackoverflow.com/a/44823381 (online; accessed:20.02.2019).
[7] scikit-learn 0.20.3 documentation [Electronic resource]. — 2019. — Access mode: https://scikit-learn.org/ (online; accessed: 05.04.2019).