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

Лабораторная работа №6

по дисциплине «Методы машинного обучения»

на тему«Классификация текста»

Выполнил: Ван Пэй

Группа：ИУ5-22М

Москва — 2021 г.

Цель лабораторной работы:

Изучение методов классификации текстов.

Задание:

Для произвольного набора данных, предназначенного для классификации текстов, решите задачу классификации текста двумя способами:

Способ 1. На основе CountVectorizer или TfidfVectorizer.

Способ 2. На основе моделей word2vec или Glove или fastText. Сравните качество полученных моделей.

Для поиска наборов данных в поисковой системе можно использовать ключевые слова "datasets for text classification".

Описание решения:

Использован способ 2: На основе моделей. Начало модели соответствует Лаб 5.

3. Базовые векторные представления слов и текстов

Мешок слов aka Bag of words aka BOW)

```
]: with open('../input/preprocessed1/lemmatized_tokens (1).txt', 'r') as
       data = f.read().split('\n')

   tokenized_lemmatized_words = [i.split(',') for i in data]
   tokenized_lemmatized_words[:5]
```

```
[['д ж о а н',
  'к э т л и н',
  'р о у л и н г',
  'г а р р и',
  'п о т т е р',
  'ф и л о с о ф с к и й',
  'к а м е н ь',
  'г л а в а',
  'п е р в а я',
  'м а л ь ч и к',
  'к о т о р ы й',
  'в ы ж и л',
  'м и с т е р',
  'м и с с и с',
  'д а р с л и',
  'д о м а',
  'н о м е р',
  'ч е т ы р е',
  'б и р ю ч и н н о м у',
  'п р о е з д у',
  'м о г л и',
  'г о р д о с т ь ю',
  'с к а з а т ь',
  'с л а в а',
  'б о г у',
  'с о в е р ш е н н о',
  'н о р м а л ь н ы е',
  'л ю д и'],
 ['п о с л е д н и м и',
  'к о г о',
  'м о г л и',
  'о ж и д а т ь',
  '
```

```
import gensim
from gensim import corpora

dictionary = corpora.Dictionary()

corpus = [dictionary.doc2bow(sent, allow_update=True) for sent in toke
```

```
# тут можно посмотреть что лежит в слов
# dictionary.token2id
```

```
for doc in corpus[:10]:
    print([[dictionary[id], freq] for id, freq in doc])
```

```
[['б и р ю ч и н н о м у', 1], ['б о г у', 1], ['в ы ж и л', 1],
['г а р р и', 1], ['г л а в а', 1], ['г о р д о с т ь ю', 1],
['д а р с л и', 1], ['д ж о а н', 1], ['д о м а', 1], ['к а м е
н ь', 1], ['к о т о р ы й', 1], ['к э т л и н', 1], ['л ю д и',
1], ['м а л ь ч и к', 1], ['м и с с и с', 1], ['м и с т е р', 1],
['м о г л и', 1], ['н о м е р', 1], ['н о р м а л ь н ы е', 1],
['п е р в а я', 1], ['п о т т е р', 1], ['п р о е з д у', 1],
['р о у л и н г', 1], ['с к а з а т ь', 1], ['с л а в а', 1],
['с о в е р ш е н н о', 1], ['ф и л о с о ф с к и й', 1], ['ч е т
ы р е', 1]]
[['м о г л и', 1], ['с о в е р ш е н н о', 1], ['к о г о', 1],
['о д о б р я л и', 1], ['о ж и д а т ь', 1], ['п о д о б н о й',
1], ['п о с л е д н и м и', 1], ['с т р а н н о м', 1], ['т а и н
с т в е н н о м', 1], ['у ч а с т и я', 1], ['ч е м н и б у д ь',
1], ['ч е п у х и', 1]]
[['д а р с л и', 1], ['м и с т е р', 1], ['в ы п у с к а л а',
1], ['г р а н н и н г з', 1], ['д и р е к т о р о м', 1], ['к о т
о р а я', 1], ['р а б о т а л', 1], ['с в е р л а', 1], ['ф и р
м ы', 1]]
[['б о л ь ш и м и', 1], ['к о р о т к о й', 1], ['к р е п к и
й', 1], ['к р у п н ы й', 1], ['м у ж ч и н а', 1], ['о ч е н
ь', 2], ['у с а м и', 1], ['ш е е й', 1]]
[['д а р с л и', 1], ['м и с с и с', 1], ['д в о и х', 1], ['з а
т о', 1], ['л и х в о й', 1], ['с в е т л о в о л о с о й', 1],
['х в а т и л о', 1], ['х у д о й', 1], ['ш е и', 1]]
[['б о л ь ш у ю', 1], ['в р е м е н и', 1], ['и з г о р о д и',
1], ['п о д а р к о м', 1], ['п о л ь з о в а л а с ь', 1], ['п р
и р о д ы', 1], ['с а д о в ы е', 1], ['с о с е д я м и', 1],
['у с е р д н о', 1], ['ч а с т ь', 1], ['ш п и о н я', 1], ['э
т и м', 1]]
[['д а р с л и', 1], ['в с е м', 1], ['д а д л и', 1], ['л у ч ш
```

```
len(dictionary.keys())
```

12848

```
from gensim.models import TfidfModel
tfidf = TfidfModel(corpus)

for doc in tfidf[corpus[:10]]:
    print([[dictionary[id], round(freq, 2)] for id, freq in doc])
```

```
[['б и р ю ч и н н о м у', 0.25], ['б о г у', 0.23], ['в ы ж и л', 0.21], ['г а р р и', 0.05], ['г л а в а', 0.17], ['г о р д о с т ь ю', 0.23], ['д а р с л и', 0.12], ['д ж о а н', 0.25], ['д о м а', 0.16], ['к а м е н ь', 0.14], ['к о т о р ы й', 0.12], ['к э т л и н', 0.25], ['л ю д и', 0.15], ['м а л ь ч и к', 0.15], ['м и с с и с', 0.14], ['м и с т е р', 0.13], ['м о г л и', 0.15], ['н о м е р', 0.18], ['н о р м а л ь н ы е', 0.25], ['п е р в а я', 0.22], ['п о т т е р', 0.13], ['п р о е з д у', 0.25], ['р о у л и н г', 0.25], ['с к а з а т ь', 0.14], ['с л а в а', 0.22], ['с о в е р ш е н н о', 0.18], ['ф и л о с о ф с к и й', 0.2], ['ч е т ы р е', 0.16]]
[['м о г л и', 0.19], ['с о в е р ш е н н о', 0.23], ['к о г о', 0.23], ['о д о б р я л и', 0.32], ['о ж и д а т ь', 0.26], ['п о д о б н о й', 0.32], ['п о с л е д н и м и', 0.32], ['с т р а н н о м', 0.32], ['т а и н с т в е н н о м', 0.32], ['у ч а с т и я', 0.32], ['ч е м н и б у д ь', 0.29], ['ч е п у х и', 0.32]]
[['д а р с л и', 0.19], ['м и с т е р', 0.21], ['в ы п у с к а л а', 0.39], ['г р а н н и н г з', 0.39], ['д и р е к т о р о м', 0.39], ['к о т о р а я', 0.24], ['р а б о т а л', 0.39], ['с в е р л а', 0.34], ['ф и р м ы', 0.39]]
[['б о л ь ш и м и', 0.32], ['к о р о т к о й', 0.36], ['к р е п к и й', 0.39], ['к р у п н ы й', 0.36], ['м у ж ч и н а', 0.32], ['о ч е н ь', 0.34], ['у с а м и', 0.36], ['ш е е й', 0.39]]
[['д а р с л и', 0.19], ['м и с с и с', 0.22], ['д в о и х', 0.39], ['з а т о', 0.33], ['л и х в о й', 0.39], ['с в е т л о в о л
```

Word2Vec(word to vec) ¶

```
with open('../input/preprocessed/with_pos_tag (3).txt', 'r') as f:
    data = f.read().split('\n')
```

```
with_pos_tags = [sent.split(',') for sent in data]
```

```
with_pos_tags[:5]
```

```
[['д ж о а н_NOUN',
  'к э т л и н а_NOUN',
  'р о у л и н г а_NOUN',
  'г а р р и_NOUN',
  'п о т т е р_NOUN',
  'ф и л о с о ф с к и я_NOUN',
  'к а м е н ь_NOUN',
  'г л а в а_NOUN',
  'п е р в ы й_ADJF',
  'м а л ь ч и к_NOUN',
  'к о т о р ы й_ADJF',
  'в ы ж и т ь_INFN',
  'м и с т е р_NOUN',
  'м и с с и с_NOUN',
  'д а р с л и т ь_INFN',
  'и з_PREP',
  'д о м_NOUN',
  'н о м е р_NOUN',
  'ч е т ы р е_NUMR',
  'п о_PREP',
  'б и р ю ч и н н ы й_ADJF',
  'п р о е з д_NOUN',
  'м о ч ь_NOUN',
  'б ы_PRCL',
  'г о р д о с т ь_NOUN',
  'с к а з а т ь_INFN',
  'ч т о_CONJ',
  'о н и_NPRO'
```

```
import gensim.downloader as api
model = api.load("word2vec-ruscorpora-300")  # download the mod
```

```
model.most_similar(positive=['ж е н щ и н а_NOUN', 'к о р о л
```

```
[('к о р о л е в а_NOUN', 0.7313905358314514),
 ('г е р ц о г_NOUN', 0.6502389311790466),
 ('п р и н ц е с с а_NOUN', 0.6266286373138428),
 ('г е р ц о г и н я_NOUN', 0.6240381598472595),
 ('к о р о л е в с т в о_NOUN', 0.6094207763671875),
 ('з ю д е р м а н л а н д с к и й_ADJ', 0.6084389686584473),
 ('д у р л а х с к и й_ADJ', 0.608166515827179),
 ('у л ь р и к : : э л е о н о р а_NOUN', 0.6073106527328491),
 ('м а к с и м и л и а н о в_NOUN', 0.6057005524635315),
 ('п р и н ц_NOUN', 0.5984029173851013)]
```

Список литературы

[1] Гапанюк Ю. Е. Лабораторная работа «Обработка признаков (часть2)» [Электронный ресурс] https://github.com/ugapanyuk/ml_course_2021/wiki/LAB_MMO__FEATURES