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

Отчет
Лабораторная работа № 5
По курсу «Методы машинного обучения»

Предобработка и классификация текстовых данных

ИСПОЛНИТЕЛЬ:
Попов Илья Андреевич
Группа ИУ5-23М

_____

"__"_____2022 г.

ПРЕПОДАВАТЕЛЬ:
Гапанюк Ю.Е.

_____

"__"_____2022 г.

Москва 2022

Popov I.A. IU5-23M lab5

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

```
In [3]:  import nltk
         from nltk import tokenize
```

```
In [38]: text_data = None
         with open('15660.txt', 'r') as f:
             text_data = f.read()
         text_data
```

Out[38]: 'I agree with this assessment. Furthermore, its promotion as\nproviding greater protection than bare voice is quite true, as far\nas it goes. However, the only way for it to fulfill its stated goal\nof letting LE wiretap "terrorists and drug dealers" is to restrict\nstronger techniques. \n\nWiretap targets presently use strong encryption, weak encryption, or\n(the vast majority) no encryption. The latter two classes can be\ntapped. With weak encryption in every phone, the no-encryption\nclass is merged into the weak-encryption class. Will the\nintroduction of Clipper cause targets presently enjoying strong\nprivacy to give up on it? that is, to rely for privacy on a system\nexpressly designed to deny it to people like them? I doubt it. The\nmere introduction of this scheme will give the government *nothing*.\n\nThe stated goal of preventing the degradation of wiretapping \ncapabilities can be fulfilled by restriction of domestic\ncryptography, and only by this restriction. "Clipper" appears to be\nno more than a sop, given to the public to mute any complaints. We\nwould find this a grossly inadequate tradeoff, but I fear the public\nat large will not care. I hate to even mention gun control, but \nmost people seem to think that an `assault weapon\' (as the NYT uses\nthe word) is some sort of automatic weapon, .50 caliber maybe. Who\nwants to have such a thing legal? Well, people know even less about\ncryptology; I suspect that strong cryptography could easily be\nlabeled "too much secrecy for law-abiding citizens to need".\n\n>That\'s not for Clinton (or anyone under him) to say, though. Only the \n>federal and supreme courts can say anything about the constitutionality.\n>Anything the administration or any governmental agency says is opinion at \n>best.\n\nWhat they say is opinion, but what they do is what matters, and will\ncontinue unless overturned. And the courts are reluctant to annul\nlaw or regulation, going to some length to decide cases on other\ngrounds. Furthermore, Congress can get away with quite a bit. They\ncould levy a burdensome tax; this would place enforcement in the\nhands of the BATF, who as we\'ve seen you really don\'t want on your\ncase. They could invoke the Commerce Clause; this seems most\nlikely. This clause will get you anywhere these days. The 18th was\nrequired because the Supreme Court ruled a prohibitory statute\nunconstitutional. In 1970 Congress prohibited many drugs, with a\ntextual nod to the Commerce Clause. The Controlled Substances\nAct of 1970 still stands. I think the government could get away\nwith it.\n\n>Amanda Walker\n\n\t PGP 2 key by finger or e-mail\n Eli   ebrandt@jarthur.claremont.edu\n\n\n'

```
In [9]:  #Токенизация по словам
         nltk_tk_1 = nltk.WordPunctTokenizer()
         nltk_tk_1.tokenize(text_data)
```

Out[9]: ['I',
         'agree',
         'with',
         'this',
         'assessment',
         '.',
         'Furthermore',
         ',',
         'its',
         'promotion',
         'as',
         'providing',
         'greater',
         'protection',
         'than',
         'bare',
         'voice',
         'is',
         'quite',
         'true',
         ',',
         'as',
         'far',
         'as',
         'it',
         'goes',
         '.',
         'However',
         ',',
         'the',
         'only',
         'way',
         'for',

'it',
'to',
'fulfill',
'its',
'stated',
'goal',
'of',
'letting',
'LE',
'wiretap',
'"',
'terrorists',
'and',
'drug',
'dealers',
'"',
'is',
'to',
'restrict',
'stronger',
'techniques',
'.',
'Wiretap',
'targets',
'presently',
'use',
'strong',
'encryption',
',',
'weak',
'encryption',
',',
'or',
'(',
'the',
'vast',
'majority',
')',
'no',
'encryption',
'.',
'The',
'latter',
'two',
'classes',
'can',
'be',
'tapped',
'.',
'With',
'weak',
'encryption',
'in',
'every',
'phone',
',',
'the',
'no',
'-',
'encryption',
'class',
'is',
'merged',
'into',
'the',
'weak',
'-',
'encryption',
'class',
'.',
'Will',
'the',
'introduction',
'of',
'Clipper',
'cause',
'targets',
'presently',
'enjoying',
'strong',
'privacy',
'to',
'give',

up on it? that is, to rely for privacy on a system expressly designed to deny it to people like them? I doubt it. The mere introduction of this scheme will give the government *nothing*. The stated goal of preventing the degradation of wiretapping capabilities can be fulfilled by restriction of domestic cryptography, and only by this restriction. "Clipper" appears to be no more than a sop, given to the public to

```
'mute',
'any',
'complaints',
'.',
'We',
'would',
'find',
'this',
'a',
'grossly',
'inadequate',
'tradeoff',
',',
'but',
'I',
'fear',
'the',
'public',
'at',
'large',
'will',
'not',
'care',
'.',
'I',
'hate',
'to',
'even',
'mention',
'gun',
'control',
',',
'but',
'most',
'people',
'seem',
'to',
'think',
'that',
'an',
'`',
'assault',
'weapon',
'"',
'(',
'as',
'the',
'NYT',
'uses',
'the',
'word',
')',
'is',
'some',
'sort',
'of',
'automatic',
'weapon',
',',
'.',
'50',
'caliber',
'maybe',
'.',
'Who',
'wants',
'to',
'have',
'such',
'a',
'thing',
'legal',
'?',
'Well',
',',
'people',
'know',
'even',
'less',
'about',
'cryptology',
';',
'I',
```

```
'suspect',
'that',
'strong',
'cryptography',
'could',
'easily',
'be',
'labeled',
'"',
'too',
'much',
'secrecy',
'for',
'law',
'-',
'abiding',
'citizens',
'to',
'need',
'".',
'>',
'That',
"'",
's',
'not',
'for',
'Clinton',
'(',
'or',
'anyone',
'under',
'him',
')',
'to',
'say',
',',
'though',
'.',
'Only',
'the',
'>',
'federal',
'and',
'supreme',
'courts',
'can',
'say',
'anything',
'about',
'the',
'constitutionality',
'.',
'>',
'Anything',
'the',
'administration',
'or',
'any',
'governmental',
'agency',
'says',
'is',
'opinion',
'at',
'>',
'best',
'.',
'What',
'they',
'say',
'is',
'opinion',
',',
'but',
'what',
'they',
'do',
'is',
'what',
'matters',
',',
'and',
'will',
```

continue unless overturned. And the courts are reluctant to annul law or regulation, going to some length to decide cases on other grounds. Furthermore, Congress can get away with quite a bit. They could levy a burdensome tax; this would place enforcement in the hands of the BATF, who as we've seen you really don't want on your case. They could invoke the Commerce Clause; this seems most likely.

```
    'This',
    'clause',
    'will',
    'get',
    'you',
    'anywhere',
    'these',
    'days',
    '.',
    'The',
    '18th',
    'was',
    'required',
    'because',
    'the',
    'Supreme',
    'Court',
    'ruled',
    'a',
    'prohibitory',
    'statute',
    'unconstitutional',
    '.',
    'In',
    '1970',
    'Congress',
    'prohibited',
    'many',
    'drugs',
    ',',
    'with',
    'a',
    'textual',
    'nod',
    'to',
    'the',
    'Commerce',
    'Clause',
    '.',
    'The',
    'Controlled',
    'Substances',
    'Act',
    'of',
    '1970',
    'still',
    'stands',
    '.',
    'I',
    'think',
    'the',
    'government',
    'could',
    'get',
    'away',
    'with',
    'it',
    '.',
    '>',
    'Amanda',
    'Walker',
    'PGP',
    '2',
    'key',
    'by',
    'finger',
    'or',
    'e',
    '-',
    'mail',
    'Eli',
    'ebrandt',
    '@',
    'jarthur',
    '.',
    'claremont',
    '.',
    'edu']
```

In [11]:
```python
#Токенизация по предложениям
nltk.download('punkt')
```

```python
nltk_tk_sents = nltk.tokenize.sent_tokenize(text_data)
print(len(nltk_tk_sents))
nltk_tk_sents
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\Ilya\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping tokenizers\punkt.zip.
30
```

Out[11]: ['\n\nI agree with this assessment.',
 'Furthermore, its promotion as\nproviding greater protection than bare voice is quite true, as far\nas it goes
.',
 'However, the only way for it to fulfill its stated goal\nof letting LE wiretap "terrorists and drug dealers"
is to restrict\nstronger techniques.',
 'Wiretap targets presently use strong encryption, weak encryption, or\n(the vast majority) no encryption.',
 'The latter two classes can be\ntapped.',
 'With weak encryption in every phone, the no-encryption\nclass is merged into the weak-encryption class.',
 'Will the\nintroduction of Clipper cause targets presently enjoying strong\nprivacy to give up on it?',
 'that is, to rely for privacy on a system\nexpressly designed to deny it to people like them?',
 'I doubt it.',
 'The\nmere introduction of this scheme will give the government *nothing*.',
 'The stated goal of preventing the degradation of wiretapping\ncapabilities can be fulfilled by restriction of
domestic\ncryptography, and only by this restriction.',
 '"Clipper" appears to be\nno more than a sop, given to the public to mute any complaints.',
 'We\nwould find this a grossly inadequate tradeoff, but I fear the public\nat large will not care.',
 "I hate to even mention gun control, but\nmost people seem to think that an `assault weapon' (as the NYT uses\
nthe word) is some sort of automatic weapon, .50 caliber maybe.",
 'Who\nwants to have such a thing legal?',
 'Well, people know even less about\ncryptology; I suspect that strong cryptography could easily be\nlabeled "t
oo much secrecy for law-abiding citizens to need".',
 ">That's not for Clinton (or anyone under him) to say, though.",
 'Only the \n>federal and supreme courts can say anything about the constitutionality.',
 '>Anything the administration or any governmental agency says is opinion at \n>best.',
 'What they say is opinion, but what they do is what matters, and will\ncontinue unless overturned.',
 'And the courts are reluctant to annul\nlaw or regulation, going to some length to decide cases on other\ngrou
nds.',
 'Furthermore, Congress can get away with quite a bit.',
 "They\ncould levy a burdensome tax; this would place enforcement in the\nhands of the BATF, who as we've seen
you really don't want on your\ncase.",
 'They could invoke the Commerce Clause; this seems most\nlikely.',
 'This clause will get you anywhere these days.',
 'The 18th was\nrequired because the Supreme Court ruled a prohibitory statute\nunconstitutional.',
 'In 1970 Congress prohibited many drugs, with a\ntextual nod to the Commerce Clause.',
 'The Controlled Substances\nAct of 1970 still stands.',
 'I think the government could get away\nwith it.',
 '>Amanda Walker\n\n\t PGP 2 key by finger or e-mail\n   Eli   ebrandt@jarthur.claremont.edu']

```python
#используем другую библиотеку для частиречной разметки
import spacy
#from spacy.lang.en import English
nlp = spacy.load('en_core_web_sm')
sp_tokens = nlp(text_data)
sp_tokens
```

Out[39]: I agree with this assessment. Furthermore, its promotion as
providing greater protection than bare voice is quite true, as far
as it goes. However, the only way for it to fulfill its stated goal
of letting LE wiretap "terrorists and drug dealers" is to restrict
stronger techniques.

Wiretap targets presently use strong encryption, weak encryption, or
(the vast majority) no encryption. The latter two classes can be
tapped. With weak encryption in every phone, the no-encryption
class is merged into the weak-encryption class. Will the
introduction of Clipper cause targets presently enjoying strong
privacy to give up on it? that is, to rely for privacy on a system
expressly designed to deny it to people like them? I doubt it. The
mere introduction of this scheme will give the government *nothing*.

The stated goal of preventing the degradation of wiretapping
capabilities can be fulfilled by restriction of domestic
cryptography, and only by this restriction. "Clipper" appears to be
no more than a sop, given to the public to mute any complaints. We
would find this a grossly inadequate tradeoff, but I fear the public
at large will not care. I hate to even mention gun control, but
most people seem to think that an `assault weapon' (as the NYT uses
the word) is some sort of automatic weapon, .50 caliber maybe. Who
wants to have such a thing legal? Well, people know even less about
cryptology; I suspect that strong cryptography could easily be

labeled "too much secrecy for law-abiding citizens to need".

>That's not for Clinton (or anyone under him) to say, though. Only the
>federal and supreme courts can say anything about the constitutionality.
>Anything the administration or any governmental agency says is opinion at
>best.

What they say is opinion, but what they do is what matters, and will
continue unless overturned. And the courts are reluctant to annul
law or regulation, going to some length to decide cases on other
grounds. Furthermore, Congress can get away with quite a bit. They
could levy a burdensome tax; this would place enforcement in the
hands of the BATF, who as we've seen you really don't want on your
case. They could invoke the Commerce Clause; this seems most
likely. This clause will get you anywhere these days. The 18th was
required because the Supreme Court ruled a prohibitory statute
unconstitutional. In 1970 Congress prohibited many drugs, with a
textual nod to the Commerce Clause. The Controlled Substances
Act of 1970 still stands. I think the government could get away
with it.

>Amanda Walker

        PGP 2 key by finger or e-mail
    Eli   ebrandt@jarthur.claremont.edu

```
In [40]:  for i in range(10):
              print(sp_tokens[i])
```

```
I
agree
with
this
assessment
.

Furthermore
,
its
```

```
In [41]:  #Частеречная разметка
          for i in range(30):
              print('{} - {} - {}'.format(sp_tokens[i].text, sp_tokens[i].pos_, sp_tokens[i].dep_))
```

```
I - PRON - nsubj
agree - VERB - ROOT
with - ADP - prep
this - DET - det
assessment - NOUN - pobj
. - PUNCT - punct
  - SPACE - dep
Furthermore - ADV - advmod
, - PUNCT - punct
its - PRON - poss
promotion - NOUN - nsubj
as - ADP - prep

  - SPACE - dep
providing - VERB - csubj
greater - ADJ - amod
protection - NOUN - dobj
than - ADP - prep
bare - ADJ - amod
voice - NOUN - pobj
is - AUX - ROOT
quite - ADV - advmod
true - ADJ - acomp
, - PUNCT - punct
as - ADV - advmod
far - ADV - advmod

  - SPACE - dep
as - SCONJ - mark
it - PRON - nsubj
goes - VERB - advcl
. - PUNCT - punct
```

```python
#Лемматизация
for i in range(30):
    print(sp_tokens[i], sp_tokens[i].lemma, sp_tokens[i].lemma_)
```

```
I 4690420944186131903 I
agree 3895058773709925002 agree
with 12510949447758279278 with
this 1995909169258310477 this
assessment 16732015447806033820 assessment
. 12646065887601541794 .
  8532415787641010193
Furthermore 13267237348481798544 furthermore
, 2593208677638477497 ,
its 12513610393978129441 its
promotion 15623665570904952346 promotion
as 7437575085468336610 as

 962983613142996970

providing 3805418597433244348 provide
greater 8881679497796027013 great
protection 9357963436562240989 protection
than 10794458019344880855 than
bare 4665181751273064962 bare
voice 257396040179354343 voice
is 10382539506755952630 be
quite 11176473678897442000 quite
true 7434368892455186804 true
, 2593208677638477497 ,
as 7437575085468336610 as
far 6785115870927179060 far

 962983613142996970

as 7437575085468336610 as
it 10239237003504588839 it
goes 8004577259940138793 go
. 12646065887601541794 .
```

```python
#Выделение (распознавание) именованных сущностей
from spacy import displacy
displacy.render(sp_tokens, style='ent', jupyter=True)
```

I agree with this assessment. Furthermore, its promotion as

providing greater protection than bare voice is quite true, as far

as it goes. However, the only way for it to fulfill its stated goal

of letting LE wiretap "terrorists and drug dealers" is to restrict

stronger techniques.

Wiretap targets presently use strong encryption, weak encryption, or

(the vast majority) no encryption. The latter two **CARDINAL** classes can be

tapped. With weak encryption in every phone, the no-encryption

class is merged into the weak-encryption class. Will the

introduction of Clipper cause targets presently enjoying strong

privacy to give up on it? that is, to rely for privacy on a system

expressly designed to deny it to people like them? I doubt it. The

mere introduction of this scheme will give the government *nothing*.

The stated goal of preventing the degradation of wiretapping

capabilities can be fulfilled by restriction of domestic

cryptography, and only by this restriction. "Clipper" appears to be

no more than a sop, given to the public to mute any complaints. We

would find this a grossly inadequate tradeoff, but I fear the public

at large will not care. I hate to even mention gun control, but

most people seem to think that an `assault weapon' (as the NYT **ORG** uses

the word) is some sort of automatic weapon, .50 caliber maybe. Who

wants to have such a thing legal? Well, people know even less about

cryptology; I suspect that strong cryptography could easily be

labeled "too much secrecy for law-abiding citizens to need".

>That's not for Clinton **PERSON** (or anyone under him) to say, though. Only the

>federal and supreme courts can say anything about the constitutionality.

>Anything the administration or any governmental agency says is opinion at

>best.

What they say is opinion, but what they do is what matters, and will

continue unless overturned. And the courts are reluctant to annul

law or regulation, going to some length to decide cases on other

grounds. Furthermore, Congress **ORG** can get away with quite a bit. They

could levy a burdensome tax; this would place enforcement in the

hands of the BATF, who as we've seen you really don't want on your

case. They could invoke the Commerce Clause **ORG** ; this seems most

likely. This clause will get you anywhere these days **DATE** . The 18th **ORDINAL** was

required because the Supreme Court **ORG** ruled a prohibitory statute

unconstitutional. In 1970 **DATE** Congress **ORG** prohibited many drugs, with a

textual nod to the Commerce Clause **ORG** . The Controlled Substances

Act of 1970 **DATE** still stands. I think the government could get away

with it.
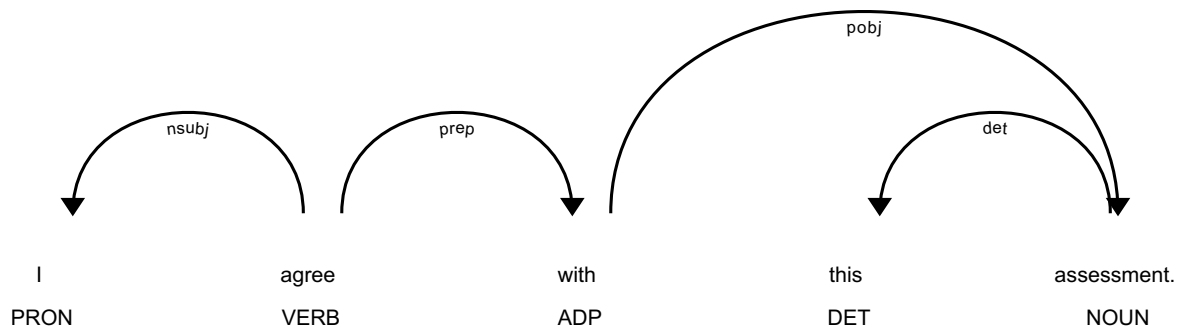

> Amanda Walker **ORG**

PGP 2 **CARDINAL** key by finger or e-mail

Eli **ORG** ebrandt@jarthur.claremont.edu


```
In [43]:    #Разбор предложения
            sp_sentence1 = sp_tokens[:6]
            sp_sentence1
```

```
Out[43]:    I agree with this assessment.
```

```python
In [28]:  displacy.render(sp_sentence1, style='dep', jupyter=True)
```


Классификация текста

```python
In [30]:  #Способ 1: TfidfVectorizer
          import numpy as np
          import pandas as pd
          import sklearn
          from sklearn.feature_extraction.text import TfidfVectorizer
```

```python
In [44]:  text_data_list = None
          with open('15660.txt', 'r') as f:
              text_data_list = f.readlines()
          text_data_list
```

```
Out[44]: ['I agree with this assessment.  Furthermore, its promotion as\n',
          'providing greater protection than bare voice is quite true, as far\n',
          'as it goes.  However, the only way for it to fulfill its stated goal\n',
          'of letting LE wiretap "terrorists and drug dealers" is to restrict\n',
          'stronger techniques.  \n',
          '\n',
          'Wiretap targets presently use strong encryption, weak encryption, or\n',
          '(the vast majority) no encryption.  The latter two classes can be\n',
          'tapped.  With weak encryption in every phone, the no-encryption\n',
          'class is merged into the weak-encryption class.  Will the\n',
          'introduction of Clipper cause targets presently enjoying strong\n',
          'privacy to give up on it?  that is, to rely for privacy on a system\n',
          'expressly designed to deny it to people like them?  I doubt it.  The\n',
          'mere introduction of this scheme will give the government *nothing*.\n',
          '\n',
          'The stated goal of preventing the degradation of wiretapping\n',
          'capabilities can be fulfilled by restriction of domestic\n',
          'cryptography, and only by this restriction.  "Clipper" appears to be\n',
          'no more than a sop, given to the public to mute any complaints.  We\n',
          'would find this a grossly inadequate tradeoff, but I fear the public\n',
          'at large will not care.  I hate to even mention gun control, but\n',
          "most people seem to think that an `assault weapon' (as the NYT uses\n",
          'the word) is some sort of automatic weapon, .50 caliber maybe.  Who\n',
          'wants to have such a thing legal?  Well, people know even less about\n',
          'cryptology; I suspect that strong cryptography could easily be\n',
          'labeled "too much secrecy for law-abiding citizens to need".\n',
          '\n',
          ">That's not for Clinton (or anyone under him) to say, though.  Only the \n",
          '>federal and supreme courts can say anything about the constitutionality.\n',
          '>Anything the administration or any governmental agency says is opinion at \n',
          '>best.\n',
          '\n',
          'What they say is opinion, but what they do is what matters, and will\n',
          'continue unless overturned.  And the courts are reluctant to annul\n',
          'law or regulation, going to some length to decide cases on other\n',
          'grounds.  Furthermore, Congress can get away with quite a bit.  They\n',
          'could levy a burdensome tax; this would place enforcement in the\n',
          "hands of the BATF, who as we've seen you really don't want on your\n",
          'case.  They could invoke the Commerce Clause; this seems most\n',
          'likely.  This clause will get you anywhere these days.  The 18th was\n',
          'required because the Supreme Court ruled a prohibitory statute\n',
          'unconstitutional.  In 1970 Congress prohibited many drugs, with a\n',
          'textual nod to the Commerce Clause.  The Controlled Substances\n',
          'Act of 1970 still stands.  I think the government could get away\n',
          'with it.\n',
```

```
'\n',
'>Amanda Walker\n',
'\n',
'\t PGP 2 key by finger or e-mail\n',
'  Eli   ebrandt@jarthur.claremont.edu\n',
'\n',
'\n']
```

In [45]:
```python
tfidfv = TfidfVectorizer(ngram_range=(1,3))
tfidf_ngram_features = tfidfv.fit_transform(text_data_list)
tfidf_ngram_features
```

Out[45]:
```
<52x951 sparse matrix of type '<class 'numpy.float64'>'
        with 1126 stored elements in Compressed Sparse Row format>
```

In [46]:
```python
# Размер нулевой строки
len(tfidf_ngram_features.todense()[0].getA1())
```

Out[46]: 951

In [47]:
```python
# Непустые значения нулевой строки
[i for i in tfidf_ngram_features.todense()[0].getA1() if i>0]
```

Out[47]:
```
[0.22821699663041792,
 0.22821699663041792,
 0.22821699663041792,
 0.16959798496893802,
 0.22821699663041792,
 0.22821699663041792,
 0.22821699663041792,
 0.20658246355111792,
 0.22821699663041792,
 0.22821699663041792,
 0.20658246355111792,
 0.22821699663041792,
 0.22821699663041792,
 0.22821699663041792,
 0.22821699663041792,
 0.1542480394660581,
 0.22821699663041792,
 0.22821699663041792,
 0.16959798496893802,
 0.22821699663041792,
 0.22821699663041792]
```

In [1]:
```python
#Способ 2: Fasttext
import fasttext
```

In [2]:
```python
ft_model = fasttext.load_model('C:/Users/Ilya/Documents/!University/MMO_2022/cc.en.300.bin')
```

Warning : `load_model` does not return WordVectorModel or SupervisedModel any more, but a `FastText` object which is very similar.

In [5]:
```python
ft_model.get_nearest_neighbors('target')
```

Out[5]:
```
[(0.819719672203064, 'targets'),
 (0.6835477948188782, 'target.The'),
 (0.6587862372398376, 'targeting'),
 (0.6565704941749573, 'target.'),
 (0.6332108974456787, 'targer'),
 (0.6261875629425049, 'target.A'),
 (0.6221246719360352, 'traget'),
 (0.621788421348572, 'targeted'),
 (0.6163643002510071, 'targetting'),
 (0.6080957055091858, 'target.If')]
```