Лабораторная работа №5

Васильев А.Р. ИУ5-24М

Цель лабораторной работы: изучение методов предобработки текстов.

Требования к отчету:

Отчет по лабораторной работе должен содержать:

- титульный лист;
- описание задания;
- текст программы;
- экранные формы с примерами выполнения программы. #### Задание - для произвольного предложения или текста решите следующие задачи:

- Токенизация;

- Частеречная разметка;
- Лемматизация;
- Выделение (распознавание) именованных сущностей;
- Разбор предложения.

In [52]:

```
import nltk
import spacy
import numpy as np
from sklearn.datasets import fetch_20newsgroups
nltk.download('punkt')
from nltk import tokenize
import re
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

Будем использовать датасет 20 newsgroups

In [53]:

```
categories = ["rec.autos", "rec.sport.hockey", "sci.crypt", "sci.med", "talk.
newsgroups_train = fetch_20newsgroups(subset='train', categories=categories)
newsgroups_test = fetch_20newsgroups(subset='test', categories=categories)
```

In [54]:

```
unique, frequency = np.unique(newsgroups_train.target,
                              return_counts = True)
```

In [55]:

```
for l, f in zip(unique, frequency):
  print(f'value: {l}, count: {f}')
```

```
value: 0, count: 594
value: 1, count: 600
value: 2, count: 595
```

```
value: 3, count: 594
value: 4, count: 377
```

In [56]:

```
print('Tokenizers NLTK have')
for i in dir(tokenize)[:16]:
  print(i)
```

```
Tokenizers NLTK have
BlanklineTokenizer
LineTokenizer
MWETokenizer
PunktSentenceTokenizer
RegexpTokenizer
ReppTokenizer
SExprTokenizer
SpaceTokenizer
StanfordSegmenter
TabTokenizer
TextTilingTokenizer
ToktokTokenizer
TreebankWordTokenizer
TweetTokenizer
WhitespaceTokenizer
WordPunctTokenizer
```

Токенизация

In [57]:

```
text = newsgroups_train.data[0]
print(text)

cleaned_text = re.sub('[^a-zA-Z0-9 \n\.]', '', text)
print(cleaned_text)
```

```
From: davec@ECE.Concordia.CA (Dave Chu)
Subject: WANTED: OPINIONS ON 75 MG
Nntp-Posting-Host: dreams.ece.concordia.ca
Organization: ECE - Concordia University
Lines: 14

I was wondering if anyone out in net-land have any opinions on MGs
in general.  I know they are not the most reliable cars around but
summer is approaching and they are convertibles `8^).  I'm interested
in a 75 MG but any opinions on MGs would be appreciated.  Thanks.

Dave
                                          |\ |     | |
___________________________/\  /\  /\_____| \|_____| |_____   ___  ___   ___
  Dave Kai-Chui Chu          \/  \/       | /|     | |       |    |     |
  Dept. of Elec. & Comp. Eng.             |/ |     | |       |--  |     |--
  Concordia University                  Voice:(514)848-3115  |___ |___  |___
  1455 de Maisonneuve W. H915           Fax:  (514)848-2802
  Montreal, Quebec, Canada H3G 1M8      Email:davec@ece.concordia.ca
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

From davecECE.Concordia.CA Dave Chu
Subject WANTED OPINIONS ON 75 MG
NntpPostingHost dreams.ece.concordia.ca
Organization ECE  Concordia University
Lines 14

I was wondering if anyone out in netland have any opinions on MGs
in general.  I know they are not the most reliable cars around but
summer is approaching and they are convertibles 8.  Im interested
in a 75 MG but any opinions on MGs would be appreciated.  Thanks.

Dave
```

```
Dave KaiChui Chu
Dept. of Elec.  Comp. Eng.
Concordia University                  Voice5148483115
1455 de Maisonneuve W. H915           Fax  5148482802
Montreal Quebec Canada H3G 1M8      Emaildavecece.concordia.ca
```

In [60]:

```
tokenizer_wp = nltk.WordPunctTokenizer()
tokens = tokenizer_wp.tokenize(cleaned_text)
```

In [63]:

```
from spacy.lang.en import English
import spacy
nlp = spacy.load("en_core_web_sm")
spacy_text1 = nlp(cleaned_text)
```

Частеречная разметка (Part-Of-Speech tagging, POS-tagging)

In [64]:

```
for token in spacy_text1:
    print('{} - {} - {}'.format(token.text, token.pos_, token.dep_))
```

```
From - ADP - ROOT
davecECE.Concordia - PROPN - pobj
. - PUNCT - punct
CA - PROPN - compound
Dave - PROPN - compound
Chu - PROPN - ROOT

 - SPACE -
Subject - PROPN - compound
WANTED - PROPN - compound
OPINIONS - PROPN - ROOT
ON - ADP - prep
75 - NUM - nummod
MG - PROPN - pobj

 - SPACE -
NntpPostingHost - PROPN - ROOT
dreams.ece.concordia.ca - PROPN - nmod

 - SPACE -
Organization - PROPN - compound
ECE - PROPN - nmod
  - SPACE -
Concordia - PROPN - compound
University - PROPN - ROOT

 - SPACE -
Lines - NOUN - ROOT
14 - NUM - nummod


 - SPACE -
I - PRON - nsubj
was - AUX - aux
wondering - VERB - ROOT
if - SCONJ - mark
anyone - PRON - nsubj
out - ADP - prep
in - ADP - prep
```

```
netland - NOUN - pobj
have - AUX - ccomp
any - DET - det
opinions - NOUN - dobj
on - ADP - prep
MGs - PROPN - pobj

 - SPACE -
in - ADP - prep
general - ADJ - amod
. - PUNCT - punct
  - SPACE -
I - PRON - nsubj
know - VERB - ROOT
they - PRON - nsubj
are - AUX - ccomp
not - PART - neg
the - DET - det
most - ADV - advmod
reliable - ADJ - amod
cars - NOUN - attr
around - ADV - prep
but - CCONJ - cc

 - SPACE -
summer - NOUN - nsubj
is - AUX - aux
approaching - VERB - ROOT
and - CCONJ - cc
they - PRON - nsubj
are - AUX - ROOT
convertibles - NOUN - attr
8 - NUM - nummod
. - PUNCT - punct
  - SPACE -
I - PRON - nsubj
m - VERB - npadvmod
interested - ADJ - ROOT

 - SPACE -
in - ADP - prep
a - DET - det
75 - NUM - nummod
MG - PROPN - pobj
but - CCONJ - cc
any - DET - det
opinions - NOUN - nsubjpass
on - ADP - prep
MGs - NOUN - pobj
would - VERB - aux
be - AUX - auxpass
appreciated - VERB - conj
. - PUNCT - punct
  - SPACE -
Thanks - NOUN - ROOT
. - PUNCT - punct


 - SPACE -
Dave - PROPN - ROOT



   - SPACE -
Dave - PROPN - compound
KaiChui - PROPN - compound
Chu - PROPN - ROOT

   - SPACE -
```

```
Dept - PROPN - dep
. - PROPN - appos
of - ADP - prep
Elec - PROPN - pobj
. - PUNCT - punct
  - SPACE - 
Comp - PROPN - appos
. - PUNCT - punct
Eng - NOUN - ROOT
. - PUNCT - punct

   - SPACE - 
Concordia - PROPN - compound
University - PROPN - compound
                  - SPACE - 
Voice5148483115 - PROPN - ROOT

   - SPACE - 
1455 - NUM - appos
de - PROPN - punct
Maisonneuve - PROPN - compound
W. - PROPN - compound
H915 - PROPN - intj
           - SPACE - 
Fax - PROPN - ROOT
  - SPACE - 
5148482802 - NUM - nummod

   - SPACE - 
Montreal - PROPN - compound
Quebec - PROPN - compound
Canada - PROPN - ROOT
H3 - PROPN - ROOT
G - PROPN - ROOT
1M8 - PROPN - nummod
      - SPACE - 
Emaildavecece.concordia.ca - PROPN - ROOT


 - SPACE - 
```

Лемматизация

In [65]:

```python
for token in spacy_text1:
      print(token, token.lemma, token.lemma_)
```

```
From 7831658034963690409 from
davecECE.Concordia 14523868423411891367 davecECE.Concordia
. 12646065887601541794 .
CA 8902657483871908647 CA
Dave 15237984737769454380 Dave
Chu 11811199503084153866 Chu

 962983613142996970 

Subject 10828730662571105685 Subject
WANTED 7965972062204850113 WANTED
OPINIONS 3594955551980145801 OPINIONS
ON 5640369432778651323 on
75 6571748649314214940 75
MG 7932641007965793912 MG

 962983613142996970 

NntpPostingHost 11894535311416990441 NntpPostingHost
dreams.ece.concordia.ca 10305115850927697332 dreams.ece.concordia.ca
```

```
 962983613142996970

Organization 685973656875656808 Organization
ECE 366147149345845489 ECE
  8532415787641010193
Concordia 9848136328145832432 Concordia
University 986123062041987797 University

 962983613142996970

Lines 9545763306533606446 line
14 9798277639574861054 14


 908432558851201422


I 561228191312463089 -PRON-
was 10382539506755952630 be
wondering 17230765341337091640 wonder
if 12446819118446800910 if
anyone 444920330522528470 anyone
out 1696981056005371314 out
in 3002984154512732771 in
netland 9247405803685480328 netland
have 14692702688101715474 have
any 13148361048351484388 any
opinions 14536103007527724270 opinion
on 5640369432778651323 on
MGs 12856644686425764130 MGs

 962983613142996970

in 3002984154512732771 in
general 4476931165537661438 general
. 12646065887601541794 .
  8532415787641010193
I 561228191312463089 -PRON-
know 7743033266031195906 know
they 561228191312463089 -PRON-
are 10382539506755952630 be
not 447765159362469301 not
the 7425985699627899538 the
most 11104729984170784471 most
reliable 10751479260089776908 reliable
cars 17545852598994811774 car
around 3194226484742107227 around
but 14560795576765492085 but

 962983613142996970

summer 14937584329648122761 summer
is 10382539506755952630 be
approaching 13730216158190965 approach
and 2283656566040971221 and
they 561228191312463089 -PRON-
are 10382539506755952630 be
convertibles 3964087356174508259 convertible
8 5117079446564601502 8
. 12646065887601541794 .
  8532415787641010193
I 561228191312463089 -PRON-
m 10382539506755952630 be
interested 2072552851738343 61 interested

 962983613142996970

in 3002984154512732771 in
a 11901859001352538922 a
```

```
75 6571748649314214940 75
MG 7932641007965793912 MG
but 14560795576765492085 but
any 13148361048351484388 any
opinions 14536103007527724270 opinion
on 5640369432778651323 on
MGs 722646293792979593 mg
would 6992604926141104606 would
be 10382539506755952630 be
appreciated 11915230412772266274 appreciate
. 12646065887601541794 .
  8532415787641010193
Thanks 8960195134057108264 thank
. 12646065887601541794 .


 9084325588512014 22


Dave 15237984737769454380 Dave



   1783162232174459540



Dave 15237984737769454380 Dave
KaiChui 8748783432222834126 KaiChui
Chu 11811199503084153866 Chu

   8029423334175508679

Dept 7320582945065383184 Dept
. 12646065887601541794 .
of 886050111519832510 of
Elec 1958910702233705206 Elec
. 12646065887601541794 .
  8532415787641010193
Comp 6958077383709244653 Comp
. 12646065887601541794 .
Eng 528727853987738021 eng
. 12646065887601541794 .

   16621976668123663401

Concordia 9848136328145832432 Concordia
University 986123062041987797 University
                  9687856286444532230
Voice5148483115 17156059720282402116 Voice5148483115

   9100049831318385606

1455 6397100970172086592 1455
de 11144093025662894627 de
Maisonneuve 16399122424279130147 Maisonneuve
W. 7347484796042509418 W.
H915 9135924853700306158 H915
           11427293743432821961
Fax 17030259339538587598 Fax
  8532415787641010193
5148482802 9829312210527632909 5148482802

   11295366195010100045

Montreal 17406125053521963953 Montreal
Quebec 13048886353597724116 Quebec
Canada 12493166723054806753 Canada
H3 12131622538089761714 H3
```

```
G 16829822105563051112 G
1M8 9540698899994290761 1M8
      17579141535385064505
Emaildavecece.concordia.ca 13465759656340681344 Emaildavecece.concordia.ca


 908432558851201422
```

Выделение (распознавание) именованных сущностей

In [67]:

```
for ent in spacy_text1.ents:
  print(ent.text, ent.label_)
```

```
davecECE.Concordia ORG
Dave Chu PERSON
75 CARDINAL
ECE ORG
Concordia University ORG
summer DATE
8 CARDINAL
75 CARDINAL
Dave


   PERSON
Dave KaiChui Chu PERSON
Dept. GPE
Elec GPE
Comp ORG
Concordia University ORG
1455 DATE
de Maisonneuve W. H915 PERSON
5148482802 CARDINAL
Montreal Quebec ORG
1M8 CARDINAL
```

In [78]:

```
text = """
The Eiffel Tower is a wrought-iron lattice tower on the Champ de Mars in Pari
It is named after the engineer Gustave Eiffel, whose company designed and bui
Locally nicknamed "La dame de fer" (French for "Iron Lady"),
 it was constructed from 1887 to 1889 as the entrance to the 1889 World's Fai
  initially criticised by some of France's leading artists and intellectuals
   but it has become a global cultural icon of France and one of the most rec
    The Eiffel Tower is the most-visited paid monument in the world; 6.91 mil

"""

text_without_spaces = ' '.join([i.strip() for i in text.split(' ')])
```

In [79]:

```
etower_text = nlp(text_without_spaces)
for ent in etower_text.ents:
  print(ent.text, ent.label_)
```

```
The Eiffel Tower FAC
the Champ de Mars FAC
Paris GPE
France GPE
Gustave Eiffel PERSON
La dame de fer WORK_OF_ART
French NORP
```

```
Iron Lady WORK_OF_ART
1887 to 1889 DATE
1889 DATE
World's Fair EVENT
France GPE
France GPE
one CARDINAL
The Eiffel Tower FAC
6.91 million CARDINAL
2015 DATE
```

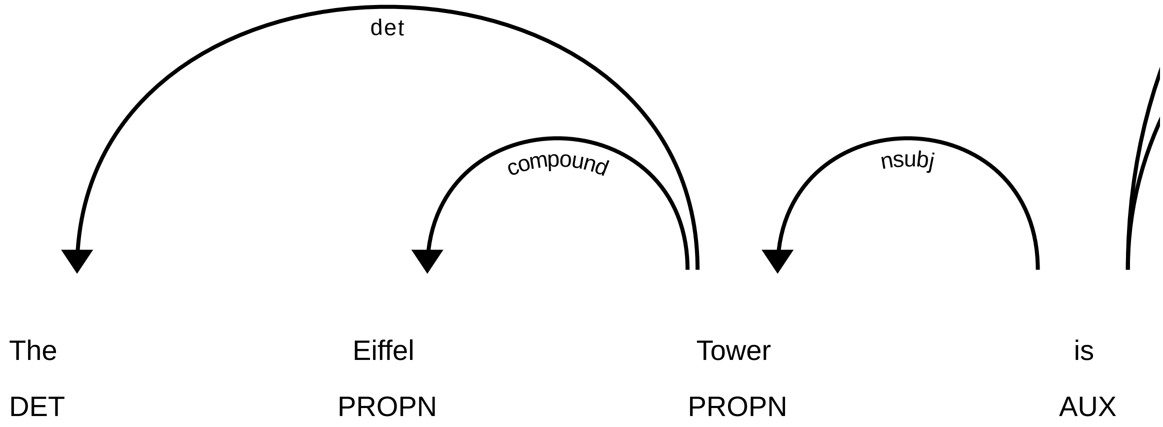
Разбор предложения

In [80]:

```
from spacy import displacy
```

In [82]:

```
for sentence in etower_text.sents:
  displacy.render(sentence, style='dep', jupyter=True)
  break
```


det
compound
nsubj
The
DET
Eiffel
PROPN
Tower
PROPN
is
AUX