Лабораторная работа №5 по курсу "Методы машинного обучения"

Выполнил: Саврасов П.А. группа ИУ5-24М

Задание

Для произвольного предложения или текста решите следующие задачи:<\p>

1. Токенизация.
2. Частеречная разметка.
3. Лемматизация.
4. Выделение (распознавание) именованных сущностей.
5. Разбор предложения.

In [20]:
```
import spacy
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from spacy import displacy
import warnings
warnings.filterwarnings('ignore')
```

In [10]:
```
with open('text.txt', 'r') as fp:
    text = fp.read()
text
```

Out[10]:
```
'Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data. The result is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them. The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves.Challenges in natural language processing frequently involve speech recognition, natural language understanding, and natural-language generation.'
```

Токенизация

In [14]:
```
nlp = spacy.load('en_core_web_sm')
spacy_text = nlp(text)
for t in spacy_text[:10]:
    print(t)
print('...')
```

```
Natural
language
processing
(
NLP
)
is
a
subfield
of
...
```

Частичная разметка

In [17]:
```
for token in spacy_text[:10]:
    print('{} - {} - {}'.format(token.text, token.pos_, token.dep_))
print('...')
```

```
Natural - ADJ - amod
language - NOUN - compound
processing - NOUN - nsubj
( - PUNCT - punct
NLP - PROPN - appos
) - PUNCT - punct
is - AUX - ROOT
a - DET - det
subfield - NOUN - attr
of - ADP - prep
...
```

Лемматизация

In [18]:
```
for token in spacy_text[:10]:
      print(token, token.lemma, token.lemma_)
```

```
Natural 3743574233330547430 natural
language 8740476009882919263 language
processing 10935198773122488114 processing
( 12638816574900267446 (
NLP 15832915187156881108 NLP
) 3842344029291005339 )
is 10382539506755952630 be
a 11901859001352538922 a
subfield 4202070545653491895 subfield
of 8860501115198325l0 of
```

Выделение (распознавание) именованных сущностей

In [26]:
```
for ent in spacy_text.ents:
    print(ent.text, '->', ent.label_)
```

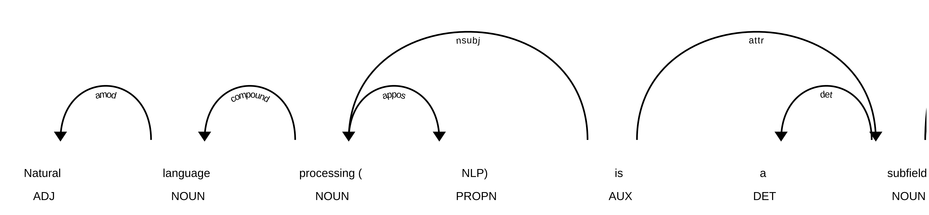
```
NLP -> ORG
```

In [23]:
```
displacy.render(spacy_text, style='ent', jupyter=True)
```

Natural language processing (NLP ORG) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data. The result is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them. The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves.Challenges in natural language processing frequently involve speech recognition, natural language understanding, and natural-language generation.

Разбор предложения

In [24]:
```
displacy.render(spacy_text, style='dep', jupyter=True)
```



In [ ]:
```
```