Лабораторная работа №5: "Предобработка текста."

ИУ5-23 Зорин Арсений

Задание:

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

In [16]:
```
from spacy.lang.ru import Russian
import spacy
from spacy import displacy
```

In [17]:
```
sentence = 'Язык разрабатывался как язык программирования для создания высокоэффективных программ, работающих н

nlp = spacy.load('ru_core_news_sm')
text = nlp(sentence)
text
```

Out[17]:
```
Язык разрабатывался как язык программирования для создания высокоэффективных программ, работающих на современны
х распределённых системах и многоядерных процессорах, Иван.
```

In [18]:
```
for token in text:
    print('{} - {} - {}'.format(token.text, token.pos_, token.dep_))
```

```
Язык - NOUN - nsubj:pass
разрабатывался - VERB - ROOT
как - SCONJ - case
язык - NOUN - obl
программирования - NOUN - nmod
для - ADP - case
создания - NOUN - nmod
высокоэффективных - ADJ - amod
программ - NOUN - nmod
, - PUNCT - punct
работающих - VERB - acl
на - ADP - case
современных - ADJ - amod
распределённых - ADJ - amod
системах - NOUN - obl
и - CCONJ - cc
многоядерных - ADJ - amod
процессорах - NOUN - conj
, - PUNCT - punct
Иван - PROPN - appos
. - PUNCT - punct
```

In [19]:
```
# лемматизация
for token in text:
      print(token, token.lemma, token.lemma_)
```

```
Язык 14510553211863083651 язык
разрабатывался 15564489334748586254 разрабатываться
как 13039644133688645009 как
язык 14510553211863083651 язык
программирования 8074303456646587670 программирование
для 10075485332184864679 для
создания 18173957039368943813 создание
высокоэффективных 16834134106748613661 высокоэффективный
программ 12303722236974168530 программа
, 2593208677638477497 ,
работающих 18093316098418404276 работать
на 16191904166009283104 на
современных 5694862869184012350 современный
распределённых 931496266546309172 распределённый
системах 10806820580119676155 система
и 15015917632809974589 и
многоядерных 5217752035983917432 многоядерный
процессорах 4505449870847683111 процессор
, 2593208677638477497 ,
Иван 1346038828973845847 иван
. 12646065887601541794 .
```

In [20]:
```
# выделение (распознавание) именованных сущностей
for ent in text.ents:
    print(ent.text, ent.label_)
```

```
Иван PER
```

In [21]:
```
displacy.render(text, style='ent', jupyter=True)
```

Язык разрабатывался как язык программирования для создания высокоэффективных программ, работающих на современных распределённых системах и многоядерных процессорах, Иван PER .

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