Лабораторная работа по теме "Предобработка текста"

Выполнил: Пакало Александр Сергеевич ИУ5-22М

Задание

Для произвольного предложения или текста решите следующие задачи:

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

In [1]:
```
import spacy
from spacy.lang.ru import Russian

# Changes display style for spacy.displacy.
IS_JUPYTER = True
```

In [2]:
```
text1 = "Natural Language Toolkit (NLTK) - одна из наиболее старых и из
# Хорошо подходит для тестирования частей речи (2 пункт) и разбора пред
# (пункт 5).
text2 = "На косой косе Косой косил траву косой косой."
# Хорошо подходит для тестирования Named Entity Recognition (4 пункт).
text3 = (
    "Москва - столица России, по преданию ее основал князь Юрий Долгору
)
#
my_text = "Nyandex - отличная компания, ведь в ней (не) работают такие

CHOSEN_TEXT = my_text
```

1. Токенизация

In [3]:
```
nlp = spacy.load("ru_core_news_sm")
spacy_text = nlp(CHOSEN_TEXT)

for token in spacy_text:
    print(token)
```

```
Nyandex
–
отличная
компания
,
ведь
в
ней
(
не
)
работают
такие
люди
как
Пвинкович
```

2. Частеречная разметка

In [4]:
```
import pandas as pd

# You may need reduce or other method for larger texts because it can g
# for each column we iterate over all dataset.
token_text_column = [token.text for token in spacy_text]
token_position_column = [token.pos_ for token in spacy_text]
token_dep_column = [token.dep_ for token in spacy_text]
token_explanation_column = [f"{spacy.explain(position)}, {spacy.explain

pd.DataFrame(
    data={
        "text": token_text_column,
        "position": token_position_column,
        "dep": token_dep_column,
        "explanation": token_explanation_column,
    }
)
```

Out[4]:

	text	position	dep	explanation
0	Nyandex	PROPN	nsubj	proper noun, nominal subject
1	-	PROPN	amod	proper noun, adjectival modifier
2	отличная	ADJ	amod	adjective, adjectival modifier
3	компания	NOUN	ROOT	noun, root
4	,	PUNCT	punct	punctuation, punctuation
5	ведь	SCONJ	mark	subordinating conjunction, marker
6	в	ADP	case	adposition, case marking
7	ней	PRON	obl	pronoun, oblique nominal
8	(	PUNCT	punct	punctuation, punctuation
9	не	PART	advmod	particle, adverbial modifier
10	)	PUNCT	punct	punctuation, punctuation
11	работают	VERB	acl	verb, clausal modifier of noun (adjectival cla...
12	такие	DET	det	determiner, determiner
13	люди	NOUN	nsubj	noun, nominal subject
14	как	SCONJ	case	subordinating conjunction, case marking
15	Пвинкович	PROPN	obl	proper noun, oblique nominal

3. Лемматизация

In [5]:
```
# You may need reduce or other method for larger texts because it can g
# for each column we iterate over all dataset.
token_lemma_column = [token.lemma_ for token in spacy_text]

pd.DataFrame(
    data={
        "text": token_text_column,
        "lemma": token_lemma_column,
    }
)
```

Out[5]:

	text	lemma
0	Nyandex	nyandex
1	-	-
2	отличная	отличный
3	компания	компания
4	,	,
5	ведь	ведь
6	в	в
7	ней	ней
8	(	(
9	не	не
10	)	)
11	работают	работать
12	такие	такой
13	люди	человек
14	как	как
15	Пвинкович	пвинкович

4. Выделение (распознавание) именованных сущностей

In [6]:
```
from spacy import displacy

displacy.render(spacy_text, style="ent", jupyter=IS_JUPYTER)
```

Nyandex - отличная компания **ORG** , ведь в ней (не) работают такие люди как Пвинкович **PER**

In [7]:
```
entity_text_column = [entity.text for entity in spacy_text.ents]
entity_label_column = [entity.label_ for entity in spacy_text.ents]
entity_explanation_column = [spacy.explain(entity) for entity in entity

pd.DataFrame(
    data={
        "entity": entity_text_column,
        "label": entity_label_column,
        "explanation": entity_explanation_column,
    }
)
```

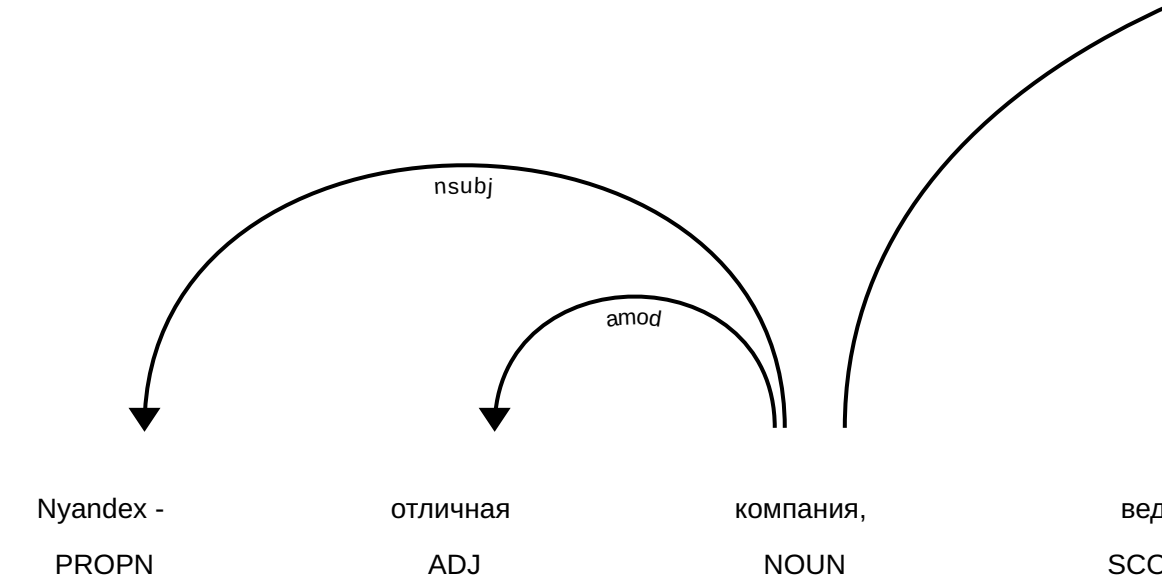
Out[7]:

	entity	label	explanation
0	Nyandex - отличная компания	ORG	Companies, agencies, institutions, etc.
1	Пвинкович	PER	Named person or family.

5. Разбор предложения

In [8]:
```
from spacy import displacy

displacy.render(spacy_text, style="dep", jupyter=IS_JUPYTER)
```



Вывод

Библиотека `spacy` отлично справляется с задачами преобработки и анализа текстов. С помощью неё мы можем токенизировать текст, определить леммы, выделить именованные сущности и даже разобрать предложение.