- **Цель лабораторной работы**: изучение методов предобработки текстов.

Задание

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

Инициализация текста

In [1]:
```
text1 = 'К началу лета цены на кофе в кофейнях вырастут минимум на 15%. Минимальная
text2 = 'В Совете Федерации прошли консультации по кандидатуре Андрея Белоусова на
text3 = 'Милейшее празднование Дня Победы прямиком из эстонской Нарвы. Несколько ты
```

Ход работы

Токенизация.

In [4]:
```
from spacy.lang.ru import Russian
import spacy
nlp = spacy.load('ru_core_news_sm')
spacy_text1 = nlp(text1)
spacy_text1
```

Out[4]:
```
К началу лета цены на кофе в кофейнях вырастут минимум на 15%. Минимальная стоимос
ть капучино достигнет 200 рублей за объём 200 граммов. Американо объёмом 0,3 подор
ожает до 195 рублей.
```

In [5]:
```
for t in spacy_text1:
    print(t)
```

```
К
началу
лета
цены
на
кофе
в
кофейнях
вырастут
минимум
на
15
%
.
Минимальная
стоимость
капучино
достигнет
200
рублей
за
объём
200
граммов
.
Американо
объёмом
0,3
подорожает
до
195
рублей
.
```

In [6]:
```
spacy_text2 = nlp(text2)
spacy_text2
```

Out[6]:
```
В Совете Федерации прошли консультации по кандидатуре Андрея Белоусова на пост мин
истра обороны.
```

In [7]:
```
spacy_text3 = nlp(text3)
spacy_text3
```

Out[7]:
```
Милейшее празднование Дня Победы прямиком из эстонской Нарвы. Несколько тысяч мест
ных жителей пришли на набережную, чтобы посмотреть на концерт в честь 9 мая у Иван
городской крепости в России. Хотя бы через реку.
```

Частеречная разметка.

In [8]:
```
for token in spacy_text1:
    print('{} - {} - {}'.format(token.text, token.pos_, token.dep_))
```

```
К - ADP - case
началу - NOUN - obl
лета - NOUN - nmod
цены - NOUN - nsubj
на - ADP - case
кофе - NOUN - nmod
в - ADP - case
кофейнях - NOUN - nmod
вырастут - VERB - ROOT
минимум - ADV - advmod
на - ADP - case
15 - NUM - nummod
% - SYM - obl
. - PUNCT - punct
Минимальная - ADJ - amod
стоимость - NOUN - nsubj
капучино - NOUN - nmod
достигнет - VERB - ROOT
200 - NUM - nummod
рублей - NOUN - obj
за - ADP - case
объём - NOUN - obl
200 - NUM - nummod
граммов - NOUN - nmod
. - PUNCT - punct
Американо - VERB - nsubj
объёмом - NOUN - obl
0,3 - NUM - appos
подорожает - VERB - ROOT
до - ADP - case
195 - NUM - nummod
рублей - NOUN - obl
. - PUNCT - punct
```

Лемматизация.

In [9]:
```
for token in spacy_text1:
      print(token, token.lemma, token.lemma_)
```

```
К 1186448856500892395З К
началу 967127417975996310 начало
лета 3178836356482510233 лето
цены 10664777512171039835 цена
на 16191904166009283104 на
кофе 11414041559424983189 кофе
в 15939375860797385675 в
кофейнях 16461198773658873811 кофейня
вырастут 12344508787348253374 вырасти
минимум 7676803677218300408 минимум
на 16191904166009283104 на
15 13771760024209633521 15
% 16590897233515608007 %
. 12646065887601541794 .
Минимальная 3191993238250835527 минимальный
стоимость 4407544826582624941 стоимость
капучино 8693340540434594053 капучино
достигнет 9228071595938805622 достигнуть
200 4266673471014057460 200
рублей 18401144977415468634 рубль
за 8493257168786769949 за
объём 6384255986188200105 объём
200 4266673471014057460 200
граммов 11159605252356613645 грамм
. 12646065887601541794 .
Американо 15679821603825215091 американо
объёмом 6384255986188200105 объём
0,3 16159747554441604937 0,3
подорожает 3119655654249011302 подорожать
до 2372892090537508025 до
195 16204778561300328931 195
рублей 18401144977415468634 рубль
. 12646065887601541794 .
```

Выделение (распознавание) именованных сущностей.

In [17]:
```
for ent in spacy_text2.ents:
    print(ent.text, ent.label_)
```

```
Совете Федерации ORG
Андрея Белоусова PER
```

In [18]:
```
from spacy import displacy
displacy.render(spacy_text2, style='ent', jupyter=True)
```

В Совете Федерации **ORG** прошли консультации по кандидатуре Андрея Белоусова **PER** на пост министра обороны.

In [20]:
```
print(spacy.explain("ORG"))
```

```
Companies, agencies, institutions, etc.
```

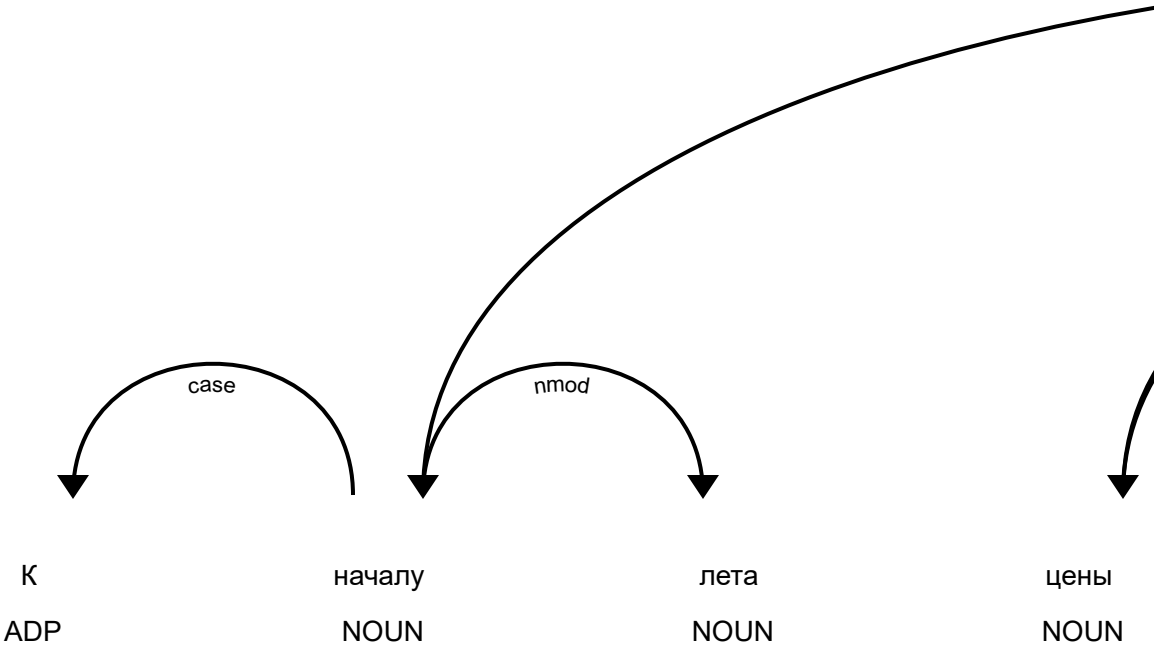
In [21]:
```
print(spacy.explain("PER"))
```

```
Named person or family.
```

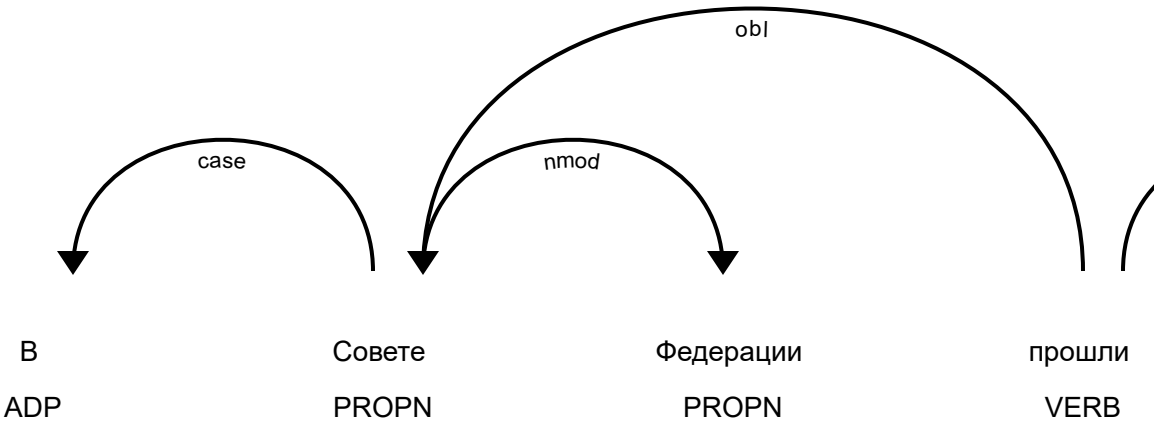
Разбор предложения.

In [22]:
```
from spacy import displacy
```

In [23]:
```
displacy.render(spacy_text1, style='dep', jupyter=True)
```



In [24]:
```
displacy.render(spacy_text2, style='dep', jupyter=True)
```




```
In [27]: print(spacy.explain("case"))
```

```
case marking
```

```
In [29]: print(spacy.explain("NOUN"))
```

```
noun
```

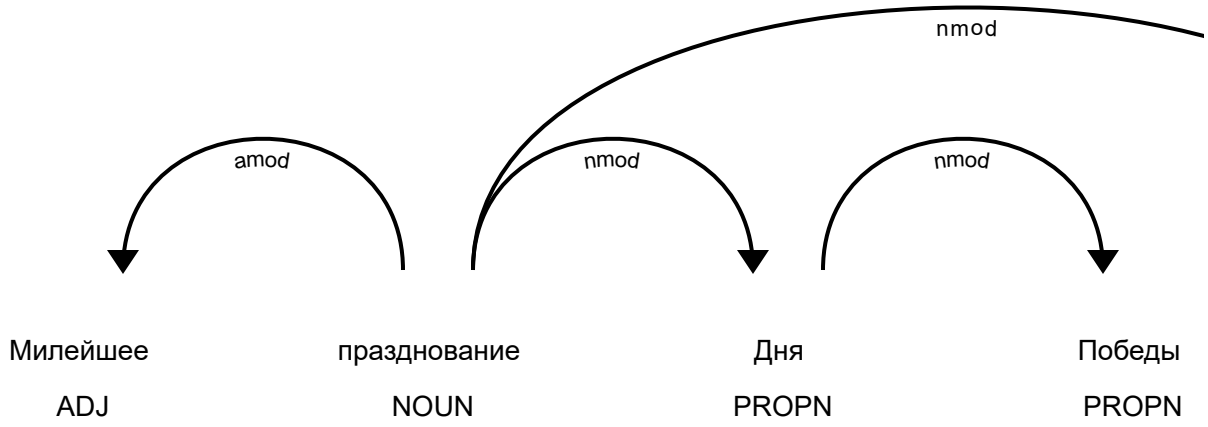
```
In [30]: print(spacy.explain("obl"))
```

```
oblique nominal
```

```
In [31]: print(spacy.explain("nsubj"))
```

```
nominal subject
```

```
In [32]: displacy.render(spacy_text3, style='dep', jupyter=True)
```


nmod
amod
nmod
nmod
Милейшее
ADJ
празднование
NOUN
Дня
PROPN
Победы
PROPN