

Анализ текста

Семинар 13

Парсинг информации

- Необходимо написать робота, который будет скачивать новости с сайта Лента.Ру и фильтровать их в зависимости от интересов пользователя. От пользователя требуется отмечать интересующие его новости, по которым система будет выделять области его интересов.
- Начнем с загрузки новостей. Для этого нам потребуется метод `requests.get(url)`. Библиотека `requests` предоставляет серьезные возможности для загрузки информации из Интернет. Метод `get` получает URL страницы и возвращает ее содержимое. В нашем случае результат будет получаться в формате `html`.

Как убрать служебную информацию

Количество служебной информации в странице явно превышает объем текста новости. У нас есть два пути: либо использовать библиотеку для извлечения данных HTML и XML BeautifulSoup для получения текста статьи, либо получить текст с использованием регулярных выражений.

Пример с парсингом

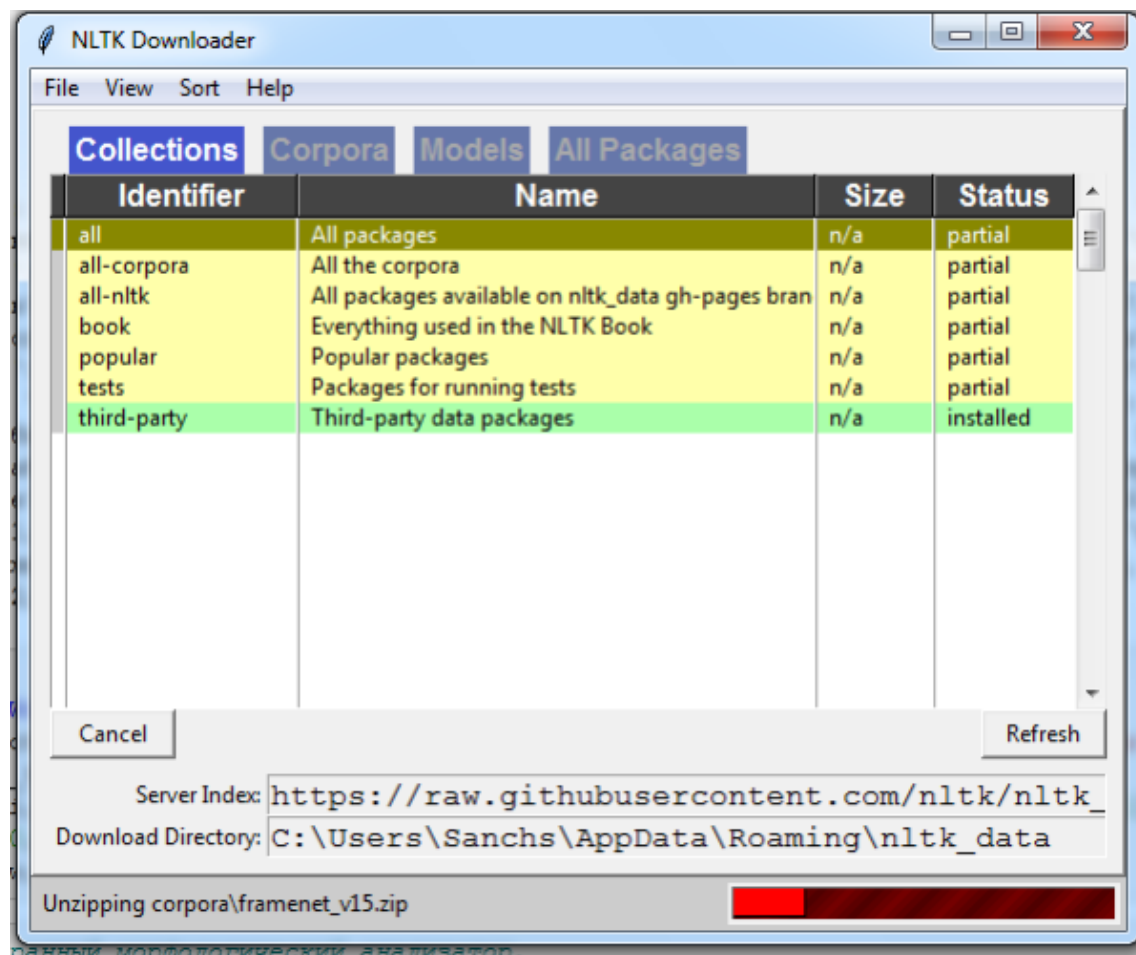
- colab.research.google.com/drive/1sMi5V5DcdYPR6m6_IS0G4c8uJEr70jy0#scrollTo=V30RHrfL7Mre

Анализ текста

- Токенизация – выделение отдельных слов в предложении
- Морфологический анализ – разбор по частям речи
- Синтаксический анализ
- Семантический анализ

Установка библиотеки nltk для морфологического анализа

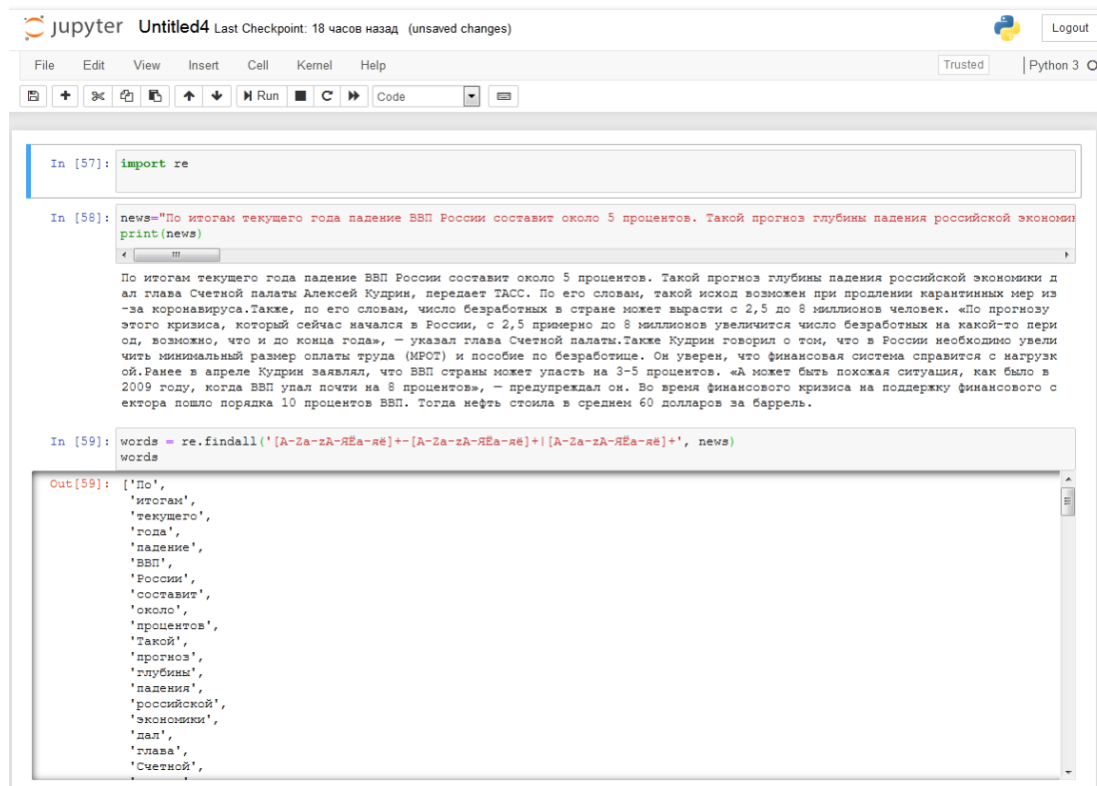
`import nltk` # Иностраный морфологический анализатор



анный морфологический анализатор.

Токенизация

- Задача выделения слов в заданном тексте с помощью регулярных выражений



The screenshot shows a Jupyter Notebook interface with the following content:

```
In [57]: import re
```

```
In [58]: news="По итогам текущего года падение ВВП России составит около 5 процентов. Такой прогноз глубины падения российской экономики дал глава Счетной палаты Алексей Кудрин, передает ТАСС. По его словам, такой исход возможен при продлении карантинных мер из-за коронавируса. Также, по его словам, число безработных в стране может вырасти с 2,5 до 8 миллионов человек. «По прогнозу этого кризиса, который сейчас начался в России, с 2,5 примерно до 8 миллионов увеличится число безработных на какой-то период, возможно, что и до конца года», — указал глава Счетной палаты. Также Кудрин говорил о том, что в России необходимо увеличить минимальный размер оплаты труда (МРОТ) и пособие по безработице. Он уверен, что финансовая система справится с нагрузкой. Ранее в апреле Кудрин заявлял, что ВВП страны может упасть на 3-5 процентов. «А может быть похожая ситуация, как было в 2009 году, когда ВВП упал почти на 8 процентов», — предупреждал он. Во время финансового кризиса на поддержку финансового сектора пошло порядка 10 процентов ВВП. Тогда нефть стоила в среднем 60 долларов за баррель."
```

```
In [59]: words = re.findall('[А-З а-з А-Я я а-я]+|[А-З а-з А-Я я]+|[А-З а-з А-Я я]+', news)
words
```

```
Out[59]: ['По',
          'итогами',
          'текущего',
          'года',
          'падение',
          'ВВП',
          'России',
          'составит',
          'около',
          'процентов',
          'Такой',
          'прогноз',
          'глубины',
          'падения',
          'российской',
          'экономики',
          'дал',
          'глава',
          'счетной',
          'палаты',
          'Кудрин',
          'передаёт',
          'ТАСС',
          'По',
          'его',
          'словами',
          'такой',
          'исход',
          'возможен',
          'при',
          'продлении',
          'карантинных',
          'мер',
          'из-за',
          'коронавируса',
          'Также',
          'по',
          'его',
          'словами',
          'число',
          'безработных',
          'в',
          'стране',
          'может',
          'вырасти',
          'с',
          '2,5',
          'до',
          '8',
          'миллионов',
          'человек',
          '«По',
          'прогнозу',
          'этого',
          'кризиса',
          'который',
          'сейчас',
          'начался',
          'в',
          'России',
          'с',
          '2,5',
          'примерно',
          'до',
          '8',
          'миллионов',
          'увеличится',
          'число',
          'безработных',
          'на',
          'какой-то',
          'период',
          'возможно',
          'что',
          'и',
          'до',
          'конца',
          'года',
          '»,',
          'указал',
          'глава',
          'Счетной',
          'палаты',
          'Также',
          'Кудрин',
          'говорил',
          'о',
          'том',
          'что',
          'в',
          'России',
          'необходимо',
          'увеличить',
          'минимальный',
          'размер',
          'оплаты',
          'труда',
          'МРОТ',
          'и',
          'пособие',
          'по',
          'безработице',
          'Он',
          'уверен',
          'что',
          'финансовая',
          'система',
          'справится',
          'с',
          'нагрузкой',
          'Ранее',
          'в',
          'апреле',
          'Кудрин',
          'заявлял',
          'что',
          'ВВП',
          'страны',
          'может',
          'упасть',
          'на',
          '3-5',
          'процентов',
          '«А',
          'может',
          'быть',
          'похожая',
          'ситуация',
          'как',
          'было',
          'в',
          '2009',
          'году',
          'когда',
          'ВВП',
          'упал',
          'почти',
          'на',
          '8',
          'процентов',
          '»,',
          'предупреждал',
          'он',
          'Во',
          'время',
          'финансового',
          'кризиса',
          'на',
          'поддержку',
          'финансового',
          'сектора',
          'пошло',
          'порядка',
          '10',
          'процентов',
          'ВВП',
          'Тогда',
          'нефть',
          'стоила',
          'в',
          'среднем',
          '60',
          'долларов',
          'за',
          'баррель'"]
```

Токенизация

- Задача подсчета слов в заданном тексте

```
In [59]: words = re.findall('[A-Za-zА-ЯЁа-яё]+-[A-Za-zА-ЯЁа-яё]+|[A-Za-zА-ЯЁа-яё]+', news)
words
```

```
'по',
'безработице',
'Он',
'уверен',
'что',
'финансовая',
'система',
'справится',
'с',
'нагрузкой',
'Ранее',
'в',
'апреле',
'Кудрин',
'заявлял',
'что',
'ВВП',
'страны',
'может',
```

```
In [60]: from collections import Counter # Подсчет частоты
```

```
wdict = Counter(words) # Объект сразу посчитает частоты элементов списка.
print(wdict)
print({w:n for w,n in wdict.items() if n>1}) # Вывод слов, которые встречаются больше 1 раза
```

```
Counter({'в': 6, 'ВВП': 4, 'процентов': 4, 'на': 4, 'что': 4, 'По': 3, 'России': 3, 'Кудрин': 3, 'может': 3, 'с': 3, 'до': 3, 'года': 2, 'глава': 2, 'Счетной': 2, 'палаты': 2, 'его': 2, 'словам': 2, 'Также': 2, 'по': 2, 'число': 2, 'безработных': 2, 'миллионов': 2, 'кризиса': 2, 'и': 2, 'финансового': 2, 'итогах': 1, 'текущего': 1, 'падение': 1, 'составит': 1, 'около': 1, 'Такой': 1, 'прогноз': 1, 'глубины': 1, 'падения': 1, 'российской': 1, 'экономики': 1, 'дал': 1, 'Алексей': 1, 'передает': 1, 'ТАСС': 1, 'такой': 1, 'исход': 1, 'возможен': 1, 'при': 1, 'продлении': 1, 'карантинных': 1, 'мер': 1, 'из-за': 1, 'коронавируса': 1, 'стране': 1, 'вырасти': 1, 'человек': 1, 'прогнозу': 1, 'этого': 1, 'который': 1, 'сейчас': 1, 'начался': 1, 'примерно': 1, 'увеличится': 1, 'какой-то': 1, 'период': 1, 'возможно': 1, 'конца': 1, 'указал': 1, 'говорил': 1, 'о': 1, 'том': 1, 'необходимо': 1, 'увеличить': 1, 'минимальный': 1, 'размер': 1, 'оплаты': 1, 'труда': 1, 'МРОТ': 1, 'пособие': 1, 'безработице': 1, 'Он': 1, 'уверен': 1, 'финансовая': 1, 'система': 1, 'справится': 1, 'нагрузкой': 1, 'Ранее': 1, 'апреле': 1, 'заявлял': 1, 'страны': 1, 'упасть': 1, 'А': 1, 'быть': 1, 'покожая': 1, 'ситуация': 1, 'как': 1, 'было': 1, 'году': 1, 'когда': 1, 'упал': 1, 'почти': 1, 'предупреждал': 1, 'он': 1, 'Во': 1, 'время': 1, 'поддержку': 1, 'сектора': 1, 'пошло': 1, 'порядка': 1, 'Тогда': 1, 'нефть': 1, 'стоила': 1, 'среднем': 1, 'долларов': 1, 'за': 1, 'баррель': 1})
{'По': 3, 'года': 2, 'ВВП': 4, 'России': 3, 'процентов': 4, 'глава': 2, 'Счетной': 2, 'палаты': 2, 'Кудрин': 3, 'его': 2, 'словам': 2, 'Также': 2, 'по': 2, 'число': 2, 'безработных': 2, 'в': 6, 'может': 3, 'с': 3, 'до': 3, 'миллионов': 2, 'кризиса': 2, 'на': 4, 'что': 4, 'и': 2, 'финансового': 2}
```


Токенизация

Задание

- **Объект исследования – lenta.ru.**
- **Найти 7 слов с наибольшей частотой**
- 11.10-12.30 – анализ с lenta.ru новостей от 20 марта 2024 г.
- 13.00-14.20 – анализ с lenta.ru новостей от 25 марта 2024 г.
- 14.40-16.00 – анализ с lenta.ru новостей от 30 марта 2024 г.
- 16.20-17.40 – анализ с lenta.ru новостей от 5 апреля 2024 г.

Библиотеки. Морфологический анализ

Часть речи	Pymorphy	Mystem	NLTK
Существительное	NOUN	S	NN
Прилагательное	ADJF, ADJS	A	JJ
Глагол	VERB	V	VB
Причастие	PRTF, PRTS	V	JJ
Деепричастие	GRND	V	JJ?

Токенизация и морфологический анализ текста

```
In [2]: import nltk
```

```
In [ ]:
```

```
In [3]: sentence = """At eight o'clock on Thursday morning  
... Arthur didn't feel very good."""  
tokens = nltk.word_tokenize(sentence)  
tokens
```

```
Out[3]: ['At',  
         'eight',  
         "o'clock",  
         'on',  
         'Thursday',  
         'morning',  
         'Arthur',  
         'did',  
         "n't",  
         'feel',  
         'very',  
         'good',  
         '.']
```

```
In [4]: tagged = nltk.pos_tag(tokens)  
tagged[0:6]
```

```
Out[4]: [('At', 'IN'),  
         ('eight', 'CD'),  
         ("o'clock", 'NN'),  
         ('on', 'IN'),  
         ('Thursday', 'NNP'),  
         ('morning', 'NN')]
```

Морфологический анализ текста

```
In [38]: import nltk
```

```
In [45]: news="By the end of this year, Russia's GDP will fall by about 5 percent. This forecast of the depth of the fall of the Russia"
def normalizeNLTK(news):
    tokens = nltk.pos_tag(nltk.word_tokenize(news))
    words = []
    for t in tokens:
        if t[0] != t[1]:
            words.append(t[0]+'_'+t[1])
    return words

normalizeNLTK(news)
```

```
Out[45]: ['By_IN',
           'the_DT',
           'end_NN',
           'of_IN',
           'this_DT',
           'year_NN',
           'Russia_NNP',
           "'s_POS",
           'GDP_NNP',
           'will_MD',
           'fall_VB',
           'by_IN',
           'about_IN',
           '5_CD',
           'percent_NN',
           'This_DT',
           'forecast_NN',
           'of_IN',
           'the_DT',
           'the_DT']
```

Задание

- **С помощью регулярных выражений убрать все цифры и определить части речи оставшихся слов в тексте .**
- «By the end of this year, Russia's GDP will fall by about 5 percent. This forecast of the depth of the fall of the Russian economy was given by the head of the accounting chamber Alexey Kudrin, reports TASS. According to him, such an outcome is possible if the quarantine measures are extended due to the coronavirus. Also, according to him, the number of unemployed in the country may grow from 2.5 to 8 million people. According to the forecast of this crisis, which has now begun in Russia, the number of unemployed will increase from 2.5 to about 8 million for some period, possibly until the end of the year, the head of the accounting chamber said. Kudrin also said that Russia needs to increase the minimum wage (minimum wage) and unemployment benefits. He is confident that the financial system will cope with the load. Earlier in April, Kudrin said that the country's GDP could fall by 3-5 percent. There may be a similar situation as in 2009, when GDP fell by almost 8 percent, he warned. During the financial crisis, about 10 percent of GDP went to support the financial sector. Back then, oil cost an average of \$ 60 per barrel.»