ПРОВЕДЕНИЕ ИССЛЕДОВАНИЙ В СЕТИ ИНТЕРНЕТ С ИСПОЛЬЗОВАНИЕМ РЕГУЛЯРНЫХ ВЫРАЖЕНИЙ

МОДУЛЬ 3
ЛЕКЦИЯ 9

ОРГАНИЗАЦИЯ ПРОВЕДЕНИЯ ИССЛЕДОВАНИЯ В ИНТЕРНЕТ И ПРЕДСТАВЛЕНИЕ ЕГО РЕЗУЛЬТАТОВ

Лектор:
к.ф.-м.н., доцент кафедры программного обеспечения вычислительной техники и автоматизированных систем
Зуев С.В.

ЗАДАЧА ИССЛЕДОВАНИЯ

- **Новые знания** – доля рынка компании по денежному обороту в Ярославской области (в процентах).
- **Предметная область** – услуги и решения в области информационной безопасности.

Уточнение параметров запроса:
Доля рынка измеряется отношением оборота компании в регионе ко всему обороту по предметной области в регионе. Неизвестной является величина оборота в регионе. Ее и надо искать.

ПОИСК ИСТОЧНИКОВ ИНФОРМАЦИИ

- **Краулер** – поисковый робот, собирающий URL'ы страниц, на которых может содержаться то, что вам нужно.

```
def furls(search_str, pages=2):
    urls = []
    content = []
    content.append(requests.get('http://www.google.com/search', params={'q':f'{search_str}'}))
    T = BS(content[0].text)
    Tu = T.findAll('a')
    for t in Tu:
        if re.search('.*q=http.*',t['href']) and not re.search('google|yandex|wiki',t['href']):
            urls.extend([t['href']])
    for i in range(pages):
        soup = BS(content[i].text)
        url = 'http://www.google.com'+soup.find(attrs={'aria-label': 'Следующая страница'})['href']
        content.append(requests.get(url))
        T = BS(content[i+1].text)
        Tu = T.findAll('a')
        for t in Tu:
            if re.search('.*q=http.*',t['href']) and not re.search('google|yandex|wiki',t['href']):
                urls.extend([t['href']])
    return urls
```

ПОИСК ИСТОЧНИКОВ ИНФОРМАЦИИ

- **Предметный парсер** – скрипт, выбирающий на страницах, найденных краулером, целые предложения с нужной информацией.

```
def wurls(word,neg_word,url_list):
    word += '.*'
    rsf = []
    for i,_ in enumerate(url_list):
        u = re.sub('%25','%',re.sub('&.*','',re.sub('/url\?q=','',url_list[i])))
        if u[-4:] == '.pdf':
            rsf.append('файл '+u)
        else:
            try … # обработка ошибки доступа и получение корректного пейлоада ur
          …
            if ur:
                ssoup = str(BS(ur.text))
                sf = re.findall('(<|>)([^<>]*)(<|>)',ssoup)
                for s in sf:
                    ss = s[1]
                    ss = re.sub('\\xa0',' ',ss)
                    prs = '[А-ЯЁ]{1}'+f'[^\.]*{word}.*?\.'
                    rsf.extend(re.findall(prs,ss,re.S))
                for r in rsf:
                    if (re.search(neg_word,r) or (not re.search('20\d{2}',r))) and r[-4:] != '.pdf':
            rsf.remove(r)
    return(rsf)
```

ПРОВЕДЕНИЕ ИССЛЕДОВАНИЯ

Воспользуемся написанными скриптами и сначала соберем все URL, на которых может быть информация о рынке информационной безопасности. Ограничимся 3 страницами выдачи поисковика:

```
Urus = furls('рынок информационной безопасности',3)
```

Выберем далее только те предложения, которые содержат информацию о продажах и не содержат информацию о ценах:

```
R = wurls('прода','цен',Urus)
```

Уберем повторяющиеся предложения и те, которые не содержат денежных единиц или годов:

```
Rset = set(R)
for r in Rset:
    if re.search('долл|руб',r) and re.search('2018|2019|2020',r):
    print(r)
```

Получим то, что увидите на следующем слайде…

ПРОВЕДЕНИЕ ИССЛЕДОВАНИЯ

В результате выдано всего три фразы:

- По итогам 2019 года объем этого рынка увеличился на 14% и достиг 90,6 млрд рублей.
- В 2018 году мировой объем продаж товаров и услуг, связанных с информационной безопасностью, вырастет до 114 млрд долл., что на 12,4% больше прошлогоднего. В 2019 году рынок вырастет еще на 8,7% до 124 млрд долл.
- По итогам 2018 года объем российского рынка систем информационной безопасности увеличился на 10% и составил 79,5 млрд руб.

Отсюда уже можно рассчитать объем рынка в России в указанные годы: он равен

- 72,3 млрд руб. в 2017 году,
- 79,5 млрд руб. в 2018 году,
- 90,6 млрд руб. в 2019 году.

На поиск этой информации вручную ушло бы около часа времени
Информации по Ярославской области, конечно, нет. Поэтому мы рассчитываем ее рынок по пропорциям затрат организаций на информационные и коммуникационные

ЗАВЕРШЕНИЕ ИССЛЕДОВАНИЯ

Доля рынка нашей компании в Ярославской области может быть теперь легко получена:

Наши продажи в 2019 году составили 40 млн рублей

Рынок – 90600*0,255% = 231 млн рублей.

Отсюда доля рынка равна 17%

ИССЛЕДОВАНИЯ НА БОЛЬШИХ ДАННЫХ

Источником обычно являются агрегаторы данных. Пример: данные о землетрясениях на всей планете собираются на разных ресурсах, но ресурс https://earthquake.usgs.gov/ позволяет получать их в формате .csv

Запрос исследования: построить график сейсмической активности в географической области «Камчатка»
Уточнение:

- координаты области (60,1336 с.ш., 155,4445 в.д.) – (50,7503 с.ш., 164,5101 в.д.)
- периодичность – месяц, измеряемая величина – средняя магнитуда

РЕЗУЛЬТАТ ИССЛЕДОВАНИЯ НА БОЛЬШИХ ДАННЫХ

```
import pandas as pd
df = pd.read_csv('https://earthquake.usgs.gov/earthquakes/feed/v1.0/summary/all_month.csv',
            delimiter=',')
dn = df[["latitude", "longitude","mag"]].apply(pd.to_numeric) # чаще всего формат не числовой
dn = dn[(dn['latitude']>50.7503)&(dn['latitude']<60.1336)&
          (dn['longitude']>155.4445)&(dn['longitude']<164.5101)]
dn['mag'].mean()
```

На 29/10/2020:

- Средняя магнитуда за истекший месяц – 4,3 балла
- Дисперсия – 0,15
- Всего было 8 землетрясений, их список приведен
- По данным на каждое, например, 1-е число месяца можно составить график

```
dn = dn[(dn['latitude']>50.7503)&(dn['latitude']<60.1336)&(dn['longitude']>155.4445)&(dn['longitude']<164.5101)]
dn['mag'].mean()
```

```
4.300000000000001
```

```
dn['mag'].std()
```

```
0.1511857892036911
```

```
dn
```

	latitude	longitude	mag
762	55.0602	162.8594	4.4
763	55.3678	162.6313	4.3
4281	53.7484	161.4142	4.4
4391	56.1155	160.1074	4.2
7821	51.0370	156.2908	4.4
9745	52.2107	158.8914	4.5
12894	55.9581	161.5086	4.1
13022	58.8157	158.9730	4.1