Форматы данных (1)

Материалы:

- Макрушин С.В. "Лекция 4: Форматы данных"
- https://docs.python.org/3/library/json.html
- https://docs.python.org/3/library/pickle.html
- https://www.crummy.com/software/BeautifulSoup/bs4/doc.ru/bs4ru.html
- Уэс Маккини. Python и анализ данных

Задачи для совместного разбора

1. Вывести все адреса электронной почты, содержащиеся в адресной книге `addres-book.json`

2. Вывести телефоны, содержащиеся в адресной книге `addres-book.json`

3. По данным из файла `addres-book-q.xml` сформировать список словарей с телефонами каждого из людей.

```
# from bs4 import BeautifulSoup

# with open(r"\E:\Dowloads\TOBD_2021_datasets\food_com\examples\addres-book-q.xml", 'r') as fp:
#   xml = BeautifulSoup(fp, 'xml')

# book = []
# for address in xml.find_all('address'):
#   name = address.find('name').text
#   phones = {phone.attrs['type']: phone.txt for phone in address.find_all('phone')}
#   book.append({"name": name, "phones": phones})


# book
```

Лабораторная работа №4

```
from google.colab import drive
drive.mount('/content/gdrive')
import pandas as pd
```

```
Mounted at /content/gdrive
```

JSON

1.1 Считайте файл `contributors_sample.json`. Воспользовавшись модулем `json`, преобразуйте содержимое файла в соответствующие объекты python. Выведите на экран информацию о первых 3 пользователях.

```
import json
import requests


with open('gdrive/My Drive/Colab Notebooks/files/contributors_sample.json') as f:
   data = json.load(f)

for i in range(3):
  print(data[i])
```

```
{'username': 'uhebert', 'name': 'Lindsey Nguyen', 'sex': 'F', 'address': '01261 Cameron Spring\nTaylorfurt, AK 97791', 'mail': 'jsalazar
{'username': 'vickitaylor', 'name': 'Cheryl Lewis', 'sex': 'F', 'address': '66992 Welch Brooks\nMarshallshire, ID 56004', 'mail': 'bhuds
{'username': 'sheilaadams', 'name': 'Julia Allen', 'sex': 'F', 'address': 'Unit 1632 Box 2971\nDPO AE 23297', 'mail': 'darren44@yahoo.c
```

```
data
```

```
[{'username': 'uhebert',
  'name': 'Lindsey Nguyen',
  'sex': 'F',
  'address': '01261 Cameron Spring\nTaylorfurt, AK 97791',
  'mail': 'jsalazar@gmail.com',
  'jobs': ['Energy engineer',
   'Engineer, site',
   'Environmental health practitioner',
   'Biomedical scientist',
   'Jewellery designer'],
  'id': 35193},
 {'username': 'vickitaylor',
  'name': 'Cheryl Lewis',
  'sex': 'F',
  'address': '66992 Welch Brooks\nMarshallshire, ID 56004',
  'mail': 'bhudson@gmail.com',
  'jobs': ['Music therapist',
   'Volunteer coordinator',
   'Designer, interior/spatial'],
  'id': 91970},
 {'username': 'sheilaadams',
  'name': 'Julia Allen',
  'sex': 'F',
  'address': 'Unit 1632 Box 2971\nDPO AE 23297',
  'mail': 'darren44@yahoo.com',
  'jobs': ['Management consultant',
   'Engineer, structural',
   'Lecturer, higher education',
   'Theatre manager',
   'Designer, textile'],
  'id': 1848091},
 {'username': 'nicole82',
  'name': 'Gina Stevens',
  'sex': 'F',
  'address': '9880 Michelle Bridge\nNew Kimberlybury, WY 02583',
  'mail': 'stevensonsarah@hotmail.com',
  'jobs': ['Mechanical engineer', 'Retail banker', 'Barrister'],
  'id': 50969},
 {'username': 'jean67',
  'name': 'Nicholas Harrington',
  'sex': 'M',
  'address': '9080 Monica Crescent Suite 820\nNorth Deanbury, HI 28977',
  'mail': 'denise42@gmail.com',
  'jobs': ['Network engineer',
   'Youth worker',
   'Primary school teacher',
   'Engineer, broadcasting (operations)'],
  'id': 676820},
 {'username': 'james67',
  'name': 'Allison Gomez',
  'sex': 'F',
  'address': '635 Kenneth Ways Suite 172\nHancockfort, AZ 50544',
  'mail': 'monique02@hotmail.com',
  'jobs': ['Designer, ceramics/pottery',
   'Engineer, energy',
   'Engineer, manufacturing'],
  'id': 64918},
 {'username': 'woodmarissa',
```

1.2 Выведите уникальные почтовые домены, содержащиеся в почтовых адресах людей

```
mails = []
for user in data:
  mails.append(user['mail'].split('@')[1])
mails = list(set(mails))
mails
```

```
['hotmail.com', 'gmail.com', 'yahoo.com']
```

1.3 Напишите функцию, которая по `username` ищет человека и выводит информацию о нем. Если пользователь с заданным `username` отсутствует, возбудите исключение `ValueError`

```
users = {i:'' for i in range(len(data))}
# users

i = 0
```

```
for user in data:
  users[i] = user['username']
  i += 1


def usersearch(username):
  try:
    print(data[list(users.values()).index('uhebert')])
  except ValueError:
    print('It\'s value error! (:')


usersearch('uhebert')
```

```
{'username': 'uhebert', 'name': 'Lindsey Nguyen', 'sex': 'F', 'address': '01261 Cameron Spring\nTaylorfurt, AK 97791', 'mail': 'jsalazar
```

1.4 Посчитайте, сколько мужчин и женщин присутсвует в этом наборе данных.

```
men_moment = 0
women_moment = 0
for user in data:
  if user['sex'] == 'F':
    women_moment += 1
  else:
    men_moment += 1

print("Female: ", women_moment)
print("Male: ", men_moment)
```

```
Female:  2136
Male:  2064
```

1.5 Создайте `pd.DataFrame contributors`, имеющий столбцы `id`, `username` и `sex`.

```
contributors = pd.read_json('gdrive/My Drive/Colab Notebooks/files/contributors_sample.json')
contributors = contributors[['id', 'username', 'sex']]
contributors
```

	id	username	sex
0	35193	uhebert	F
1	91970	vickitaylor	F
2	1848091	sheilaadams	F
3	50969	nicole82	F
4	676820	jean67	M
...	...	...	...
4195	423555	stevenspencer	F
4196	35251	rwilliams	M
4197	135887	lmartinez	F
4198	212714	brendahill	M
4199	344321	mistyray	F

4200 rows × 3 columns

1.6 Загрузите данные из файла `recipes_sample.csv` (**ЛР2**) в таблицу `recipes`. Объедините `recipes` с таблицей `contributors` с сохранением строк в том случае, если информация о человеке отсутствует в JSON-файле. Для скольких человек информация отсутствует?

```
recipes0 = pd.read_csv('gdrive/My Drive/Colab Notebooks/data/recipes_sample.csv', header = 0)
recipes = pd.DataFrame(recipes0)
recipes.head()
```

	name	id	minutes	contributor_id	submitted	n_steps	description	n_ingredients
0	george s at the cove black bean soup	44123	90	35193	2002-10-25	NaN	an original recipe created by chef scott meska...	18.0
1	healthy for them yogurt popsicles	67664	10	91970	2003-07-26	NaN	my children and their friends ask for my homem...	NaN
2	i can t believe it s spinach	38798	30	1533	2002-08-29	NaN	these were so go, it surprised even me.	8.0

```
df = pd.merge(recipes, contributors, on='id', how='left')
df
```

	name	id	minutes	contributor_id	submitted	n_steps	description	n_ingredients	username	sex
0	george s at the cove black bean soup	44123	90	35193	2002-10-25	NaN	an original recipe created by chef scott meska...	18.0	NaN	NaN
1	healthy for them yogurt popsicles	67664	10	91970	2003-07-26	NaN	my children and their friends ask for my homem...	NaN	NaN	NaN
2	i can t believe it s spinach	38798	30	1533	2002-08-29	NaN	these were so go, it surprised even me.	8.0	NaN	NaN
3	italian gut busters	35173	45	22724	2002-07-27	NaN	my sister-in-law made these for us at a family...	NaN	NaN	NaN
4	love is in the air beef fondue sauces	84797	25	4470	2004-02-23	4.0	i think a fondue is a very romantic casual din...	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...
29995	zurie s holey rustic olive and cheddar bread	267661	80	200862	2007-11-25	16.0	this is based on a french recipe but i changed...	10.0	NaN	NaN
29996	zwetschgenkuchen bavarian plum cake	386977	240	177443	2009-08-24	NaN	this is a traditional fresh plum cake, thought...	11.0	NaN	NaN
29997	zwiebelkuchen southwest german onion cake	103312	75	161745	2004-11-03	NaN	this is a traditional late summer early fall s...	NaN	NaN	NaN
29998	zydeco soup	486161	60	227978	2012-08-29	NaN	this is a delicious soup that i originally fou...	NaN	NaN	NaN
29999	cookies by design cookies on a stick	298512	29	506822	2008-04-15	9.0	i've heard of the 'cookies by design' company,...	10.0	NaN	NaN

30000 rows × 10 columns

```
df.isnull().sum()
```

```
name                 0
id                   0
minutes              0
contributor_id       0
submitted            0
n_steps          11190
description        623
n_ingredients     8880
username         29830
sex              29830
dtype: int64
```

pickle

2.1 На основе файла `contributors_sample.json` создайте словарь следующего вида:

```
 {
    должность: [список username людей, занимавших эту должность]
 }
```

```
jobs = []

for user in data:
  for job in user['jobs']:
```

```
    jobs.append(job)

jobs = list(set(jobs))

new = [[[job],[]] for job in jobs]

# print(new)

for user in data:
  for job in user['jobs']:
    new[jobs.index(job)][1].append(user['username'])

jobs_dict = {new[i][0][0]:new[i][1] for i in range(len(new))}

jobs_dict
```

```
       'tpearson',
       'carriewhite',
       'stephaniewade',
       'vcollins',
       'upage',
       'angeladavis',
       'shirley03',
       'theresaeaton',
       'sandra23',
       'christophermartin',
       'riversrachel',
       'adriana33',
       'emilywilliams',
       'danielbarton',
       'michelle53',
       'sandra18',
       'wmcdaniel'],
      'Engineer, energy': ['james67',
       'davidreynolds',
       'andrew58',
       'devinmoore',
       'christopher06',
       'vasquezkevin',
       'elizabethwaters',
       'christopherdaniel',
       'rachelproctor',
       'adamlee',
       'fowlerbrian',
       'thompsonkristina',
       'shannon14',
       'brittany65',
       'dustinwhite',
       'avaughan',
       'ashley59',
       'xbutler',
       'lewisalexandra',
       'igreen',
       'ryanbarajas',
       'avilachelsea'],
      'Animal technologist': ['huntterri',
       'qwoodard',
       'lsmith',
       'david26',
       'elizabethschwartz',
       'brooksstephen',
       'jeanne11',
       'lindseyanna',
       'elainerodriguez',
       'gstrickland',
       'owilson',
       'gordonlaura',
       'christopher05',
       'ryan37',
       'melaniecamacho',
       'victor08',
       'hunterchristopher',
       'karensilva',
       'cynthiaadams']}
```

2.2 Сохраните результаты в файл `job_people.pickle` и в файл `job_people.json` с использованием форматов pickle и JSON соответственно. Сравните объемы получившихся файлов. При сохранении в JSON укажите аргумент `indent` .

```
import pickle
```

```
with open('gdrive/My Drive/Colab Notebooks/files/job_people.pickle', 'wb') as file:
  pickle.dump(jobs_dict, file, protocol=pickle.HIGHEST_PROTOCOL)
```

```
with open('gdrive/My Drive/Colab Notebooks/files/job_people.json', 'w') as file:
  json.dump(jobs_dict, file, indent=2)
```

2.3 Считайте файл `job_people.pickle` и продемонстрируйте, что данные считались корректно.

```
with open('gdrive/My Drive/Colab Notebooks/files/job_people.pickle', 'rb') as file:
  new_jobs_dict = pickle.load(file)
```

```
new_jobs_dict
```


```
       'tpearson',
       'carriewhite',
       'stephaniewade',
       'vcollins',
       'upage',
       'angeladavis',
       'shirley03',
       'theresaeaton',
       'sandra23',
       'christophermartin',
       'riversrachel',
       'adriana33',
       'emilywilliams',
       'danielbarton',
       'michelle53',
       'sandra18',
       'wmcdaniel'],
      'Engineer, energy': ['james67',
       'davidreynolds',
       'andrew58',
       'devinmoore',
       'christopher06',
       'vasquezkevin',
       'elizabethwaters',
       'christopherdaniel',
       'rachelproctor',
       'adamlee',
       'fowlerbrian',
       'thompsonkristina',
       'shannon14',
       'brittany65',
       'dustinwhite',
       'avaughan',
       'ashley59',
       'xbutler',
       'lewisalexandra',
       'igreen',
       'ryanbarajas',
       'avilachelsea'],
      'Animal technologist': ['huntterri',
       'qwoodard',
       'lsmith',
       'david26',
       'elizabethschwartz',
       'brooksstephen',
       'jeanne11',
       'lindseyanna',
       'elainerodriguez',
       'gstrickland',
       'owilson',
       'gordonlaura',
       'christopher05',
       'ryan37',
       'melaniecamacho',
       'victor08',
       'hunterchristopher',
       'karensilva',
       'cynthiaadams']}
```


XML

3.1 По данным файла `steps_sample.xml` сформируйте словарь с шагами по каждому рецепту вида `{id_рецепта: ["шаг1", "шаг2"]}`. Сохраните этот словарь в файл `steps_sample.json`

```
from bs4 import BeautifulSoup

with open(r"gdrive/My Drive/Colab Notebooks/files/steps_sample.xml", 'r') as fp:
  recipe_steps = BeautifulSoup(fp, 'xml')


recipes_dict = {}

for recipe in recipe_steps.find_all('recipe'):
  id = int(recipe.find('id').text)
  steps = [step.text for step in recipe.find_all('step')]
  recipes_dict[id] = steps

recipes_dict
```

```
       'spread 1 / 4 cup of the mixture on each slice of toasted bread',
       'sprinkle evenly with walnuts',
       'top with a slice of cheese',
       'place tuna melts on a baking sheet',
       'broil 5 inches away from the heat until the cheese is melted'],
      31748: ['place brisket',
       'onion , garlic and cloves in large dutch oven',
       'add water to cover',
       'bring to a boil',
       'reduce heat , cover and simmer 2 1 / 2 to 3 hours or until tender',
       'drain brisket , cover and refrigerate up to 24 hours',
       'to prepare glaze , combine apple jelly , wine , mustard , green onions , salt , peppercorns and curry powder in small saucepan
     and heat until jelly melts , stirring occasionally',
       'place brisket in shallow roasting pan',
       'brush with glaze and roast in preheated 325-degree oven 45 minutes , basting frequently with glaze',
       'place brisket on heated serving platter and garnish with parsley and tomato roses',
       'carve brisket into thin slices and serve with remaining glaze',
       'brisket may also be cooked on charcoal grill for 30 minutes , basting often with glaze'],
      20143: ['in medium saucepan , combine brown sugar , cornstarch , 1 / 4 teaspoon caraway seed , and salt',
       'stir in 11 / 2 cups apple juice and lemon juice',
       'cook , stirring , over medium heat until thickened',
       'set aside',
       'place roast in oven-proof skillet',
       'insert meat thermometer',
       'roast at 325 degrees for 1 hour',
       'brush with 1 / 2 cup apple juice mixture',
       'roast 2 hours longer',
       'spoon combined sauerkraut , apples , carrot , and 1 / 2 cup remaining apple juice and 1 / 2 teaspoon caraway seed around pork
     roast',
       'spoon remaining apple juice mixture over pork roast and sauerkraut',
       'cover',
       'roast 30 to 60 minutes longer , or until meat thermometer registers 155 to 160 degrees',
       'let roast stand 5 to 10 minutes',
       'slice to serve'],
      524289: ['for stuffing , in a large skillet , cook bacon over medium high heat until crisp',
       'remove with a slotted spoon and rain on paper towels',
       'reserve 2 teaspoons drippings in skillet',
       'cook leeks in reserved drippings for 3 to 4 minutes or until almost tender',
       'add the wine , thyme , salt , and pepper',
       'cook and stir until liquid is evaporated',
       'remove from heat',
       'stir in cooked bacon',
       'for glaze , stir together the apple jelly and mustard',
       'set aside',
       'trim fat from chops',
       'make a pocket in each chop by cutting from fat side almost to , but not through , the opposite side',
       'spoon stuffing into pockets with chops',
       'if necessary , secure with wooden toothpicks',
       'sprinkle with additional salt and pepper',
       'place chops on rack of an uncovered grill directly over medium coals or medium low flame on your gas grill',
       'grill about 20 minutes or until juices run clear , turning once , and brushing with glaze the last five minutes of grilling',
       'remove toothpicks before serving'],
      203930: ['place bread slices on an ungreased baking sheet',
       'broil 3-4 inches from the heat for 1-2 minutes or until golden brown',
       'combine the apple , goat cheese , thyme , oregano and pepper',
       'sprinkle over bread',
       'broil 1 minute or until cheese is softened'],
      ...}
```

3.2 По данным файла `steps_sample.xml` сформируйте словарь следующего вида: `кол-во_шагов_в_рецепте: [список_id_рецептов]`

```
counter = {id:len(recipes_dict[id]) for id in recipes_dict.keys()}

all_steps = list(set(counter.values()))

counter_t = [[[step],[]] for step in all_steps]
```

```
for id in counter.keys():
  counter_t[all_steps.index(counter[id])][1].append(id)

steps_count = {counter_t[i][0][0]:counter_t[i][1] for i in range(len(counter_t))}

steps_count
       65382,
       30748,
       332745,
       288096,
       441302,
       86410,
       338453,
       370574],
      38: [310570,
       392181,
       51546,
       56109,
       336218,
       321405,
       195558,
       321190,
       36961,
       279328,
       234964,
       414937,
       90995],
      39: [254615, 423704, 267289, 418474, 226102, 486865, 306303, 46098],
      40: [424845, 42845, 228948, 155194, 224729, 297282, 125499, 154787],
      41: [111596, 264210, 202753, 95756, 179670, 153130],
      42: [159935, 84074, 69628, 239666, 234912],
      43: [518120, 37639, 48954, 261193, 192436, 137853, 418779],
      44: [233539,
       354868,
       246387,
       47467,
       343087,
       13950,
       358726,
       235584,
       366252,
       314409,
       56107],
      45: [325571, 338907, 187751, 465115, 39387],
      46: [288125],
      47: [366685, 200780, 398107, 181411],
      48: [99823, 107529],
      49: [254347, 498384],
      51: [536747, 435430, 524744],
      52: [185860],
      53: [197341, 196010, 169943],
      55: [66524, 263072],
      56: [509506],
      60: [8409],
      62: [400101],
      63: [470251],
      65: [510490],
      67: [448591],
      68: [389774],
      69: [373670],
      70: [692],
      73: [536360],
      87: [337926],
      88: [284729]}
```

3.3 Получите список рецептов, в этапах выполнения которых есть информация о времени (часы или минуты). Для отбора подходящих рецептов обратите внимание на атрибуты соответствующих тэгов.

```
time_recipes = []

for recipe in recipe_steps.find_all('recipe'):
  id = int(recipe.find('id').text)
  flag = False
  for step in recipe.find_all('step', has_minutes='1'):
    flag = True
  if flag:
    time_recipes.append(id)

time_recipes
```

```
      369891,
      371041,
      313162,
      176060,
      484256,
      474905,
      370325,
      369937,
      142253,
      424727,
      227557,
      14400,
      362411,
      330133,
      431305,
      311508,
      313788,
      384067,
      235297,
      37639,
      380657,
      277862,
      320795,
      408652,
      491444,
      349392,
      174435,
      512041,
      328233,
      133726,
      54968,
      255200,
      394303,
      57857,
      6776,
      330897,
      214541,
      90193,
      33255,
      344202,
      467357,
      93910,
      354868,
      41183,
      22025,
      108081,
      142454,
      263038,
      19582,
      159472,
      211407,
      364119,
      486822,
      439979,
      86386,
      443041,
      315671,
      ...]
```

3.4 Загрузите данные из файла `recipes_sample.csv` (**ЛР2**) в таблицу `recipes`. Для строк, которые содержат пропуски в столбце `n_steps`, заполните этот столбец на основе файла `steps_sample.xml`. Строки, в которых столбец `n_steps` заполнен, оставьте без изменений.

```
for i in range(len(recipes)):
    if str(recipes.loc[i]['n_steps']) == "nan":
      recipes.loc[[i], 'n_steps'] = counter[recipes.loc[i,'id']]
```

```
recipes
```

	name	id	minutes	contributor_id	submitted	n_steps	description	n_ingredients
0	george s at the cove black bean soup	44123	90	35193	2002-10-25	11.0	an original recipe created by chef scott meska...	18.0
1	healthy for them yogurt popsicles	67664	10	91970	2003-07-26	3.0	my children and their friends ask for my homem...	NaN
2	i can t believe it s spinach	38798	30	1533	2002-08-29	5.0	these were so go, it surprised even me.	8.0
3	italian gut busters	35173	45	22724	2002-07-27	7.0	my sister-in-law made these for us at a family...	NaN
4	love is in the air beef fondue sauces	84797	25	4470	2004-02-23	4.0	i think a fondue is a very romantic casual din...	NaN
...	...	...	...	...	...	...	...	...
29995	zurie s holey rustic olive and cheddar bread	267661	80	200862	2007-11-25	16.0	this is based on a french recipe but i changed...	10.0
29996	zwetschgenkuchen bavarian plum cake	386977	240	177443	2009-08-24	22.0	this is a traditional fresh plum cake, thought...	11.0
29997	zwiebelkuchen southwest german onion cake	103312	75	161745	2004-11-03	10.0	this is a traditional late summer early fall s...	NaN
29998	zydeco soup	486161	60	227978	2012-08-29	7.0	this is a delicious soup that i originally fou...	NaN
29999	cookies by design cookies on a stick	298512	29	506822	2008-04-15	9.0	i've heard of the 'cookies by design' company,...	10.0

30000 rows × 8 columns

3.5 Проверьте, содержит ли столбец `n_steps` пропуски. Если нет, то преобразуйте его к целочисленному типу и сохраните результаты в файл `recipes_sample_with_filled_nsteps.csv`

```
recipes.n_steps.isnull().sum()
```

```
0
```

```
recipes['n_steps'] = recipes['n_steps'].astype(int)
```

```
recipes
```

	name	id	minutes	contributor_id	submitted	n_steps	description	n_ingredients
0	george s at the cove black bean soup	44123	90	35193	2002-10-25	11	an original recipe created by chef scott meska...	18.0
1	healthy for them yogurt popsicles	67664	10	91970	2003-07-26	3	my children and their friends ask for mv homem...	NaN

```
recipes.to_csv('gdrive/My Drive/Colab Notebooks/files/recipes_sample_with_filled_nsteps.csv')
```

						29	me.	
3	italian gut busters	35173	45	22724	2002-07-27	7	my sister-in-law made these for us at a family...	NaN
4	love is in the air beef fondue sauces	84797	25	4470	2004-02-23	4	i think a fondue is a very romantic casual din...	NaN
...	...	...	...	...	...	...	...	...
29995	zurie s holey rustic olive and cheddar bread	267661	80	200862	2007-11-25	16	this is based on a french recipe but i changed...	10.0
	zwetschgenkuchen bavarian plum				2009-08-		this is a traditional fresh plum cake,	

✓ 0s completed at 3:03 PM