Dask Bag

Материалы:

- Макрушин С.В. Лекция 12: Map-Reduce
- https://docs.dask.org/en/latest/bag.html (https://docs.dask.org/en/latest/bag.html)
- JESSE C. DANIEL. Data Science with Python and Dask.

Задачи для совместного разбора

1. Считайте файл `Dostoevskiy Fedor. Igrok - BooksCafe.Net.txt` и разбейте на предложения. Подсчитайте длину (в кол-ве символов) каждого предложения.

2. Считайте файл `Dostoevskiy Fedor. Igrok - BooksCafe.Net.txt` и разбейте на предложения. Выведите предложения, длина которых не более 10 символов.

3. На основе списка предложений из задачи 1-2 создайте `dask.bag` . Рассчитайте среднюю длину предложений в тексте.

4. На основе файла `addres_book.json` создайте `dask.bag` . Посчитайте количество мобильных и рабочих телефонов в наборе данных

Лабораторная работа 12

In [1]:
```
import dask.bag as db
import json
import re
import pandas as pd
```

1. В файлах архиве `reviews_full.zip` находятся файлы, содержащие информацию об отзывах к рецептам в формате JSON Lines. Отзывы разделены на файлы в зависимости от оценки (например, в файле `reviews_1.json` находятся отзывы с оценкой 1). Считайте файлы из этого архива в виде `dask.bag` . Преобразуйте текстовое содержимое файлов в объекты python (с помощью модуля `json`). Выведите на экран первые 5 элементов полученного `bag` .

In [2]:
```
b = db.read_text('reviews_full/*.json').map(json.loads)
b
```

Out[2]:
```
dask.bag<loads, npartitions=6>
```

In [3]:
```
b.take(5)
```

Out[3]:
```
({'user_id': 452355,
  'recipe_id': 292657,
  'date': '2016-05-08',
  'review': 'WOW!!! This is the best. I have never been able to make homemade enchi
ladas that taste like the Mexican restaurants. I made this last night for my family
and they said they will never have enchiladas at the Mexican Restaurants again. Tha
nks for sharing.'},
 {'user_id': 329304,
  'recipe_id': 433404,
  'date': '2006-06-14',
  'review': 'This was good but the dressing needed something and I found it to be a
little too sweet, next time I will experiment with some garlic and herbs and reduce
the sugar slightly, thanks for sharing kcdlong!...Kitten'},
 {'user_id': 227932,
  'recipe_id': 2008187,
  'date': '1985-11-19',
  'review': 'Very good,it was a hit for my family. I used 6 cloves of garlic and ha
d 1 lb beef and  Johnsonville sausage,1/2 lb hot and  1/2 lb honey garlic( which I
wanted to use). That was a perfect combo for us. The sausage gave it nice flavor No
guestion , I will be making this often.'},
 {'user_id': 171468,
  'recipe_id': 270716,
  'date': '2019-05-21',
  'review': 'Made for ZWT-8 Family Picks after I saw these mentioned in the Family
Picks thread. So I got up this morning, made my morning coffee while the oven was h
eating & then made these quick-fix sweet treats B4 I was ready for my 2nd cup of co
ffee. I used 8" tortillas (all I had on-hand) & cut them into 8 wedges. Then I used
2 of my Pamperd Chef gadgets .. my sml basting brush to spread the honey on the war
m tortillas & my flour/sugar shaker to sprinkle the powd sugar/cinnamon mixture ato
p the honeyed tortillas. It might have taken a total of 2 min from oven to tasting!
Yum! Using the same "sweet mix" these would also be good w/butter, cream cheese or
even lime juice. Can\'t wait to try them w/lime juice! Thx for sharing this tasty l
ittle treat w/us.'},
 {'user_id': 91392,
  'recipe_id': 1159916,
  'date': '1972-09-18',
  'review': "Very nice slaw. I especially like that it doesn't have the mayo dressi
ng. I used a prepackaged Cole slaw mix, so i didn't include the carrots nor the nut
s. I also doubled the sauce. :)"})
```

2. Модифицируйте функцию разбора JSON таким образом, чтобы в каждый словарь с информацией об отзыве добавить ключ `rating`. Значение получите на основе названия файла (см. аргумент `include_path`), использовав для этого регулярное выражение.

In [3]:
```
def flatten(rec):
    record = json.loads(rec[0])
    return {
        'user_id': record['user_id'],
        'recipe_id': record['recipe_id'],
        'date': record['date'],
        'review': record['review'],
        'rating': int(rec[1].split('/')[-1][8:9])
    }
```

In [4]:
```
b_path = db.read_text('reviews_full/*.json', include_path=True)
b_new = b_path.map(flatten)
b_new.take(5)
```

Out[4]:
```
({'user_id': 452355,
  'recipe_id': 292657,
  'date': '2016-05-08',
  'review': 'WOW!!! This is the best. I have never been able to make homemade enchi
ladas that taste like the Mexican restaurants. I made this last night for my family
and they said they will never have enchiladas at the Mexican Restaurants again. Tha
nks for sharing.',
  'rating': 0},
 {'user_id': 329304,
  'recipe_id': 433404,
  'date': '2006-06-14',
  'review': 'This was good but the dressing needed something and I found it to be a
little too sweet, next time I will experiment with some garlic and herbs and reduce
the sugar slightly, thanks for sharing kcdlong!...Kitten',
  'rating': 0},
 {'user_id': 227932,
  'recipe_id': 2008187,
  'date': '1985-11-19',
  'review': 'Very good,it was a hit for my family. I used 6 cloves of garlic and ha
d 1 lb beef and  Johnsonville sausage,1/2 lb hot and  1/2 lb honey garlic( which I
wanted to use). That was a perfect combo for us. The sausage gave it nice flavor No
guestion , I will be making this often.',
  'rating': 0},
 {'user_id': 171468,
  'recipe_id': 270716,
  'date': '2019-05-21',
  'review': 'Made for ZWT-8 Family Picks after I saw these mentioned in the Family
Picks thread. So I got up this morning, made my morning coffee while the oven was h
eating & then made these quick-fix sweet treats B4 I was ready for my 2nd cup of co
ffee. I used 8" tortillas (all I had on-hand) & cut them into 8 wedges. Then I used
2 of my Pamperd Chef gadgets .. my sml basting brush to spread the honey on the war
m tortillas & my flour/sugar shaker to sprinkle the powd sugar/cinnamon mixture ato
p the honeyed tortillas. It might have taken a total of 2 min from oven to tasting!
Yum! Using the same "sweet mix" these would also be good w/butter, cream cheese or
even lime juice. Can\'t wait to try them w/lime juice! Thx for sharing this tasty l
ittle treat w/us.',
  'rating': 0},
 {'user_id': 91392,
  'recipe_id': 1159916,
  'date': '1972-09-18',
  'review': "Very nice slaw. I especially like that it doesn't have the mayo dressi
ng. I used a prepackaged Cole slaw mix, so i didn't include the carrots nor the nut
s. I also doubled the sauce. :)",
  'rating': 0})
```

3. Посчитайте количество отзывов в исходном датасете.

In [8]:
```
b.filter(lambda record: record['review']).count().compute()
```

Out[8]:
```
9057540
```

In [7]:
```
b.count().compute()
```

Out[7]:
```
9057540
```

4. Отфильтруйте `bag`, сохранив только отзывы, оставленные в 2014 и 2015 годах.

In [5]:
```
reviews = b_new.filter(lambda record: record['date'].split('-')[0] == '2014' or reco
```

5. Выполните препроцессинг отзывов:
   - привести строки к нижнему регистру
   - обрезать пробельные символы в начале и конце строки
   - удалите все символы, кроме английских букв и пробелов

Примените препроцессинг ко всем записям из `bag`, полученного в задании 4.

In [32]:
```
import re

def preprocessing(record):
    r = record['review']
    rec = r.lower().strip()
    out = " ".join(re.findall(r"[a-zA-Z0-9]+", rec))
    return {
        'user_id': record['user_id'],
        'recipe_id': record['recipe_id'],
        'date': record['date'],
        'review': out,
        'rating': record['rating']
    }
```

In [29]:
```
reviews.take(1)
```

Out[29]:
```
({'user_id': 229850,
  'recipe_id': 1300038,
  'date': '2014-10-03',
  'review': 'Took this to a New Year&#039;s Eve Party. Everyone loved it! It&#039;s
absolutely perfect, the flavor, the crunch, just delicious!',
  'rating': 0},)
```

In [33]:
```
reviews_prep = reviews.map(preprocessing)
reviews_prep.take(5)
```

Out[33]:
```
({'user_id': 229850,
  'recipe_id': 1300038,
  'date': '2014-10-03',
  'review': 'took this to a new year 039 s eve party everyone loved it it 039 s abs
olutely perfect the flavor the crunch just delicious',
  'rating': 0},
 {'user_id': 2706705,
  'recipe_id': 133747,
  'date': '2015-05-08',
  'review': 'simple and easy way to enjoy a slice of pizza any time well toasted br
ead is the key really toast it i put a bit of pizza sauce underneath my cheese for
a more pizza like flavor i used sourdough bread medium cheddar cheese fast fun grea
t idea made for 1 2 3 hits tag game',
  'rating': 0},
 {'user_id': 945545,
  'recipe_id': 898468,
  'date': '2015-06-30',
  'review': 'delish i wanted to make this spicy so i used hot enchilada sauce and j
alapeno refried beans i forgot to buy the onions so i doctored up the beans with on
ion powder and granulated garlic added the olives under the cheese and baked uncove
red for the 25 minutes served with pico de gallo sour cream and avocado chunks fant
astic thanks for sharing lazyme',
  'rating': 0},
 {'user_id': 262915,
  'recipe_id': 1657686,
  'date': '2015-11-29',
  'review': 'this is seriously delicious i mixed it up the night before and baked i
t in the morning before work made the sauce the night before and refrigerated overn
ight and heated in microwave such an unusual combination but it kind of reminded me
of the pancake on a stick that my kids used to get at the convenience store sausage
wrapped in a pancake kind of like a corn dog making it again this week and will try
to get a pic this time',
  'rating': 0},
 {'user_id': 2000408662,
  'recipe_id': 128848,
  'date': '2015-08-17',
  'review': 'many thanks found the recipe truly marvellous i can now make full use
of the lemons growing in my garden the marmalade was delicious and very easy to mak
e since it was the first time i didn 039 t want to use a large quantity and so i ha
lved the measures after converting them into grams please could we have measurement
s of ingredients in recipes in the metric system i e grams and litres and not in po
unds now i am going to search for a recipe to make lemon squash',
  'rating': 0})
```

6. Посчитайте количество отзывов в датасете, полученном в результате решения задачи 5. В случае ошибок прокомментируйте результат и исправьте функцию препроцессинга.

In [31]:
```
reviews_prep.filter(lambda record: record['review']).count().compute()
```

Out[31]:
```
735227
```

7. Посчитайте, как часто в наборе, полученном в задании 5, встречается та или иная оценка

In [32]:
```
dict(reviews_prep.map(lambda record: record['rating']).frequencies())
```

Out[32]:
```
{0: 42472, 1: 9246, 2: 9380, 3: 26532, 4: 119413, 5: 528231}
```

8. Найдите среднее значение `rating` в выборке

In [9]:
```
rating_values = dict(reviews_prep.map(lambda record: record['rating']).frequencies()
```

In [15]:
```
s = 0
c = sum(rating_values.values())
# print(c)
for rate, count in rating_values.items():
    s += rate*count
print(s/c)
```

```
4.388036296673077
```

9. Используя метод `foldby` , подсчитать максимальную длину отзывов в зависимости от оценки `rating` в наборе, полученном в задании 5.

In [9]:
```
def binop(t, x):
    return max((t, x), key=lambda x: len(x['review']))
```

In [34]:
```
max_rate_reviews = reviews_prep.foldby(lambda x: x['rating'], binop).compute()
```

In [35]:
```
max_rate_reviews
```

Out[35]:
```
[(0,
  {'user_id': 386585,
   'recipe_id': 1039476,
   'date': '2014-05-10',
   'review': 'i don 039 t need to make this recipe myself to know that it 039 s
good because it 039 s very similar to 1 that i first learned to make decades ago
i only gave it a 4 instead of a 5 because i know that i would prefer the saucier
veggie filled version that i have been making for years lt br gt lt br gt this c
asserole is very similar to one that my family has been making for well over 20
years my sister found the original recipe for a quot chicken num num quot casser
ole in a magazine or from one of those subscription recipe collections the chick
en num num recipe she found uses 1 can each of cream of chicken and cream of mus
hroom soups but no milk or broth to thin out the condensed soups the 1 2 cup may
o as well as 2 cans of french cut green beens well drained all of the shredded c
heese is blended into the creamy chicken amp green been fillng amp the biscuits
are topped with finely crushed crakers the chicken num num recipe called for put
ting the biscuits onto the filling right away topping them with crushed crackers
and putting the completely assembled casserole into the oven for a longer cook t
ime in order to get the filling hot enough to cook the bottoms of the biscuits w
```

In [36]:
```
for rate in max_rate_reviews:
    print(rate[0], ':', len(rate[1]['review']))
```

```
0 : 6741
1 : 3202
2 : 2906
3 : 3214
4 : 6741
5 : 5505
```

In [31]:
```
import re
check_str = 'first off i have to admit that i have written this review a couple time
out = ''
final = " ".join(re.findall(r"[a-zA-Z0-9]+", check_str))
final
```

Out[31]: 'first off i have to admit that i have written this review a couple times already a
nd admittedly i was so angry i couldnt write it properly so after trying to settle
down a bit here goesltbrgtltbrgta the recipe is very poorly written i have provided
many recipes over the years to family friends and as contributions to charityfundra
iser cookbooks i write recipes in painstaking detailnothing unnecessaryso that the
cook who is using my recipe doesnt have to guess about what they need to do everyth
ing must be spelled out preciselyltbrgtltbrgtb when i am following a recipe i prepa
re all the ingredients as listed and described completely so when its time to use t
hem im not scurrying around for somethingltbrgt i direction number two is quotfry t
he onion till golden in a couple of tablespoons of lightly browned butter the recip
e calls for tbsp of butter direction number is quotfry them slowly in plenty of but
terquot what is plenty of butter this needs to be more specific when cooks have pre
pared something many times they progress beyond needing specific amounts often beca
use they have personalized the recipe and its their own but this and every publishe
d recipe must be done as though the chef is making the recipe for the first time be
cause this was my first time as i said i followed the directions exactlyltbrgt ii b
ased on the list of ingredients i made sure i had soy sauce on hand admittedly the
recipe called for teaspoon of soy sauce really in a pound of meat teaspoon of soy s
auce but wait after making sure you have the soy sauce it is never mentioned again
where are we to put the soy sauce in the meat in the cream sauceltbrgt iii the ingr
edients call for quot cup cream half and half or cup milk im sorry but this is unin
telligible should we use cream or half and half or milk why is quothalf and halfquo
t in brackets cream is cream half and half is half and half and milk is milk decide
and then tell us what to do if we want to change it to accommodate a particular lif
estyle we will change it but please be clear about what the cook is supposed to use
ie tell us what the standard of quotexcellencequot is for this recipe what in your
mind makes it bestltbrgtltbrgtc after assembling the called for ingredients and abs
olutely omitting the water because by the time all the other ingredients were assem
bled it was raw soup not able to be formed into meatballs i could not begin to move
forward with shaping them into ballsltbrgtltbrgtd i suggest the following which i d
id in an effort to save the i spent on meat and ingredients i didnt have on hand ag
ain if you want to make changes to make it your own ok but i can only tell you what
worked for meltbrgt i use cream not half and half and not milk use cupltbrgt ii com
pletely omit the waterltbrgt iiibased on this recipe use at least c dried breadcrum
bs and more if necessaryltbrgt iv add some flavor to the meat there is nothing but
allspice and that according to the ltbrgtquotsourcequot is optional i dont know abo
ut you but i think meatballs are basically miniround meatloaves i do not make a mea
tloaf devoid of seasoning why would i make a meatball devoid of seasoning yes i kno
w that people rave about the sauce but i have learned this in my years of cooking i
f the meat itself is not seasoned and only depends on the sauce its in for flavor w
hat you get is a mouthful of unflavored meat with some sauce on it i learned this t
he hard way with quotthe best chili you will ever tastequot when i thought the chil
i was magnificent until i tasted the meat and it was like having a mouthful of flav
orlesssomething meat needs to be able to stand on its own and then be enhanced by a
saucethe sauce cannot make up for flavorless meat i dont know what ikea meatballs a
re likeperhaps they are supposed to be bland and let the sauce do the workpersonall
y i added two teaspoons of cumin two teaspoons of ground coriander two teaspoons of
ground fennel and cloves of garlic to go along with the onionltbrgt v chill the mea
tball mixture for hours more is ok less is not a good idealtbrgt vi dont just cook
the meatballs slowly in plenty of butter use two tablespoons of butter clarified if
possible and two tablespoons of olive oil because you are cooking slowly so as to m
ake sure the meatball is cooked through without the outside being ruined you can us
e olive oil despite it relatively low smoking point it will be okbut cook on medium
heat at best im not being specific here because there are so many permutations of r
angesi use a viking stovetop and have complete controlso anything wrong is my fault
to lower and raise temp immediately ltbrgtltbrgti have now based on the original re
cipe and only adding extra breadcrumbs and omitting the watertried frying these lit
tle clumps they are flavorless by themselves and beg for the seasonings i mentioned
above the texture is wonderful but the meatballs remain flavorlessltbrgtheres a gre

```
at chance for you to season to your hearts contentor notits up to you how you want
your meatballs to tasteltbrgtltbrgtanyway im sorry that this didnt turn out the way
it should thank you dee for giving us a great starting point to make some incredo m
eatballs i suggest making changes which will make these little morsels quite tasty
i cant comment on the cream sauce yet'
```