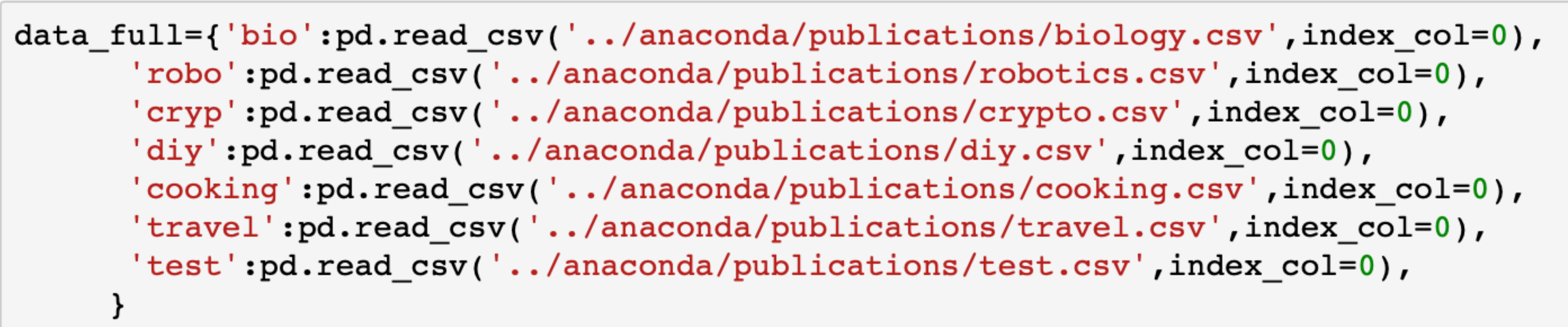
Первичный анализ и поиск похожего текста

Поиск похожих текстов

Дано: набор (корпус) текстов.

Найти: для заданного текстового сообщения (текста) найти ближайший текст из исходного набора.



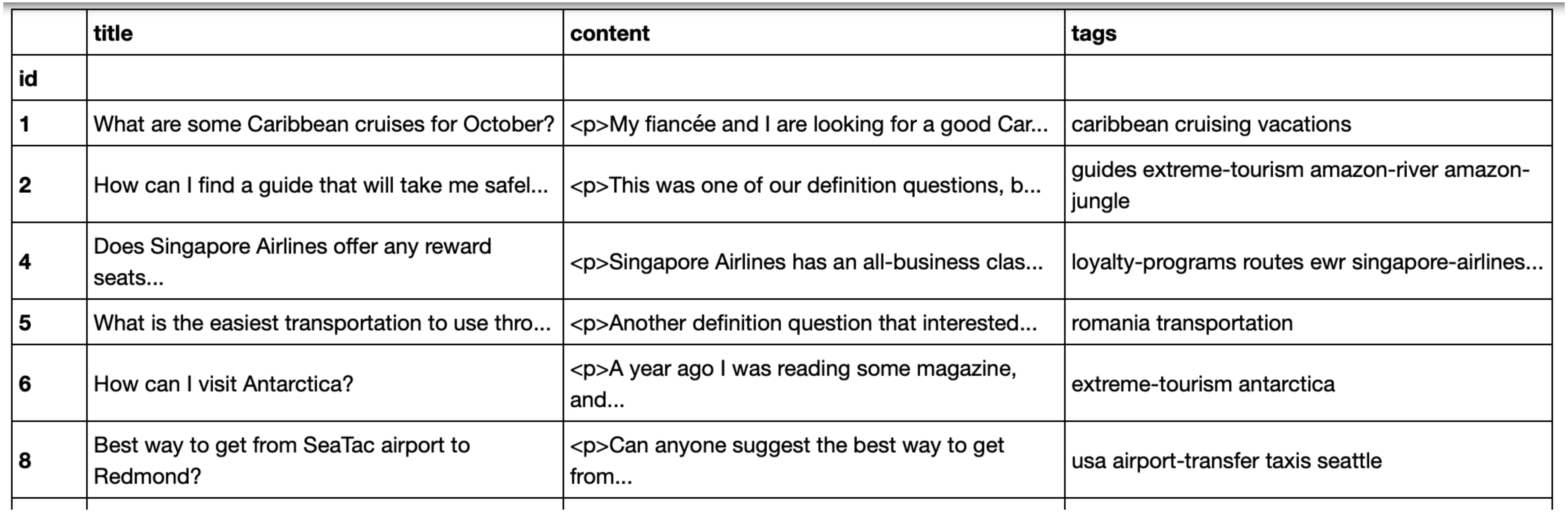
```
data_full={'bio':pd.read_csv('../anaconda/publications/biology.csv',index_col=0),
      'robo':pd.read_csv('../anaconda/publications/robotics.csv',index_col=0),
      'cryp':pd.read_csv('../anaconda/publications/crypto.csv',index_col=0),
      'diy':pd.read_csv('../anaconda/publications/diy.csv',index_col=0),
      'cooking':pd.read_csv('../anaconda/publications/cooking.csv',index_col=0),
      'travel':pd.read_csv('../anaconda/publications/travel.csv',index_col=0),
      'test':pd.read_csv('../anaconda/publications/test.csv',index_col=0),
     }
```

Сообщение: «safely travel tour Amazon jungle»

Поиск похожих текстов

Дано: набор (корпус) текстов.

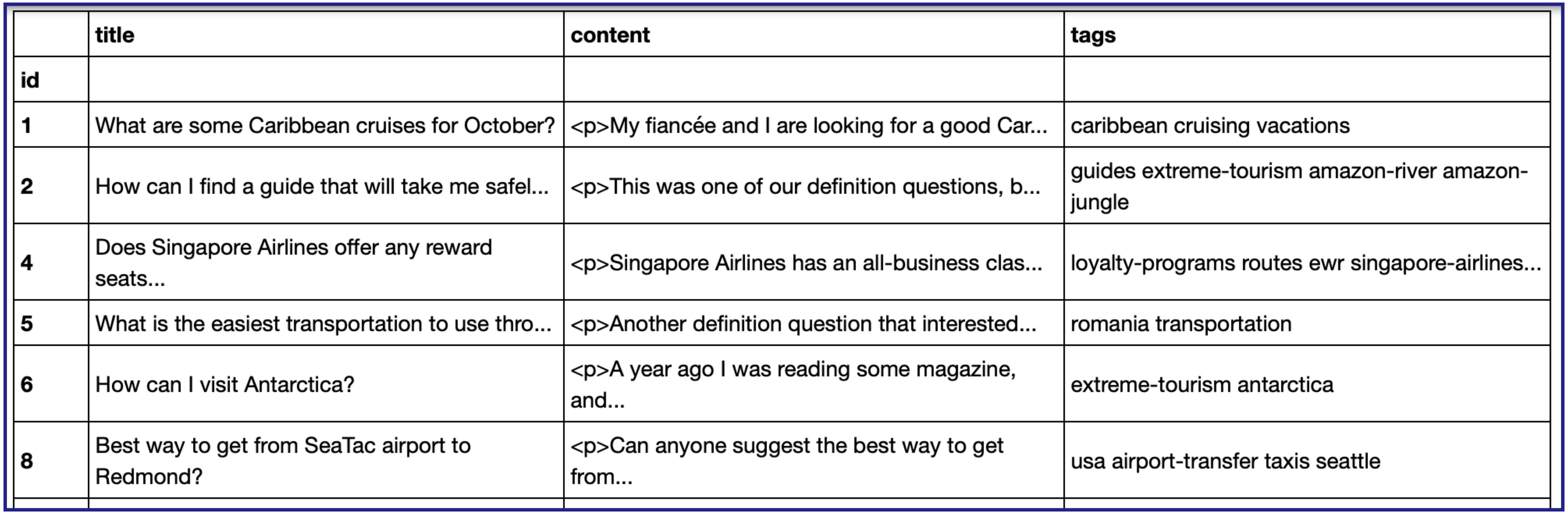
Найти: для заданного текстового сообщения (текста) найти ближайший текст из исходного набора.



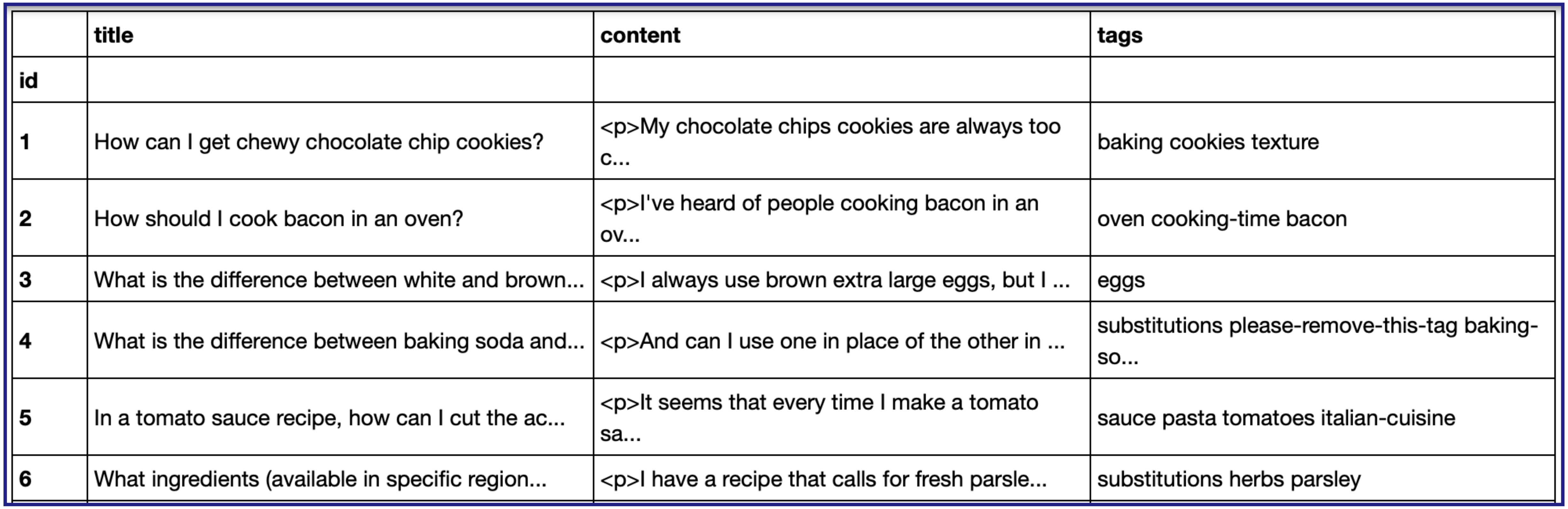
	title	content	tags
id			
1	What are some Caribbean cruises for October?	<p>My fiancée and I are looking for a good Car...	caribbean cruising vacations
2	How can I find a guide that will take me safel...	<p>This was one of our definition questions, b...	guides extreme-tourism amazon-river amazon-jungle
4	Does Singapore Airlines offer any reward seats...	<p>Singapore Airlines has an all-business clas...	loyalty-programs routes ewr singapore-airlines...
5	What is the easiest transportation to use thro...	<p>Another definition question that interested...	romania transportation
6	How can I visit Antarctica?	<p>A year ago I was reading some magazine, and...	extreme-tourism antarctica
8	Best way to get from SeaTac airport to Redmond?	<p>Can anyone suggest the best way to get from...	usa airport-transfer taxis seattle

Сообщение: «safely travel tour Amazon jungle»

Поиск похожих текстов

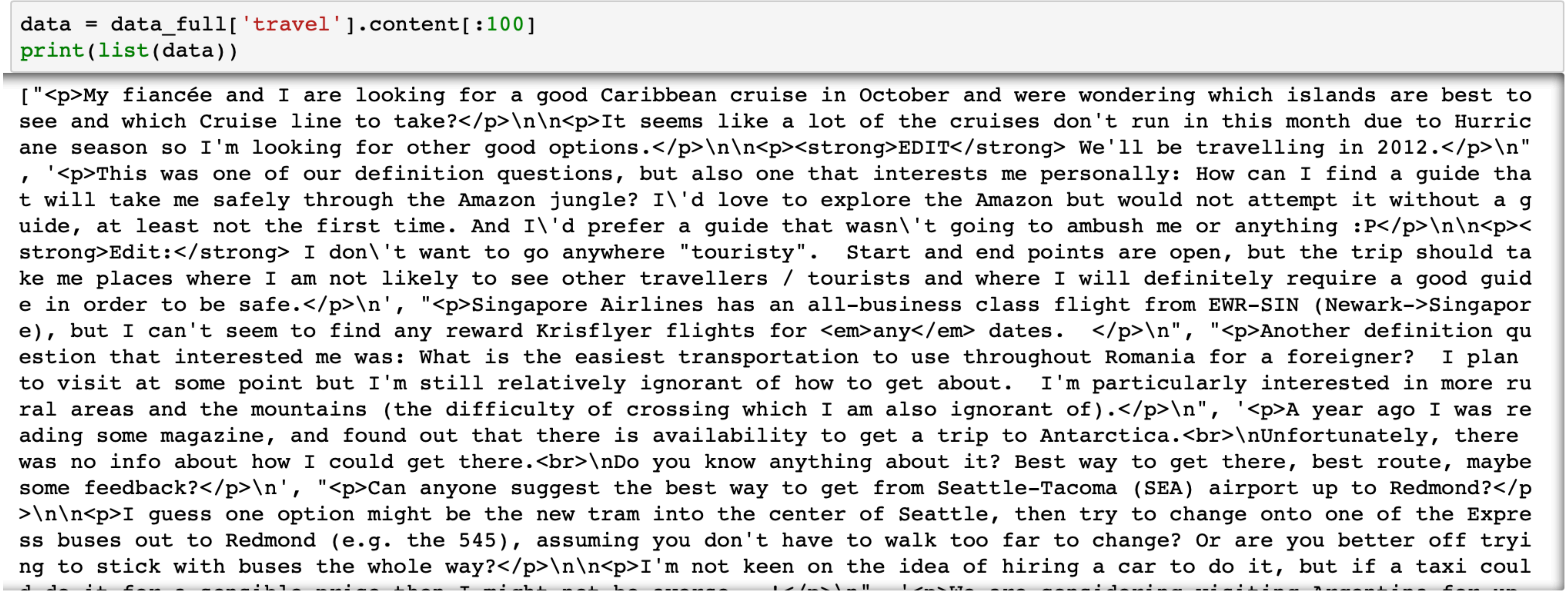


	title	content	tags
id			
1	What are some Caribbean cruises for October?	<p>My fiancée and I are looking for a good Car...	caribbean cruising vacations
2	How can I find a guide that will take me safel...	<p>This was one of our definition questions, b...	guides extreme-tourism amazon-river amazon-jungle
4	Does Singapore Airlines offer any reward seats...	<p>Singapore Airlines has an all-business clas...	loyalty-programs routes ewr singapore-airlines...
5	What is the easiest transportation to use thro...	<p>Another definition question that interested...	romania transportation
6	How can I visit Antarctica?	<p>A year ago I was reading some magazine, and...	extreme-tourism antarctica
8	Best way to get from SeaTac airport to Redmond?	<p>Can anyone suggest the best way to get from...	usa airport-transfer taxis seattle



	title	content	tags
id			
1	How can I get chewy chocolate chip cookies?	<p>My chocolate chips cookies are always too c...	baking cookies texture
2	How should I cook bacon in an oven?	<p>I've heard of people cooking bacon in an ov...	oven cooking-time bacon
3	What is the difference between white and brown...	<p>I always use brown extra large eggs, but I ...	eggs
4	What is the difference between baking soda and...	<p>And can I use one in place of the other in ...	substitutions please-remove-this-tag baking-so...
5	In a tomato sauce recipe, how can I cut the ac...	<p>It seems that every time I make a tomato sa...	sauce pasta tomatoes italian-cuisine
6	What ingredients (available in specific region...	<p>I have a recipe that calls for fresh parsle...	substitutions herbs parsley

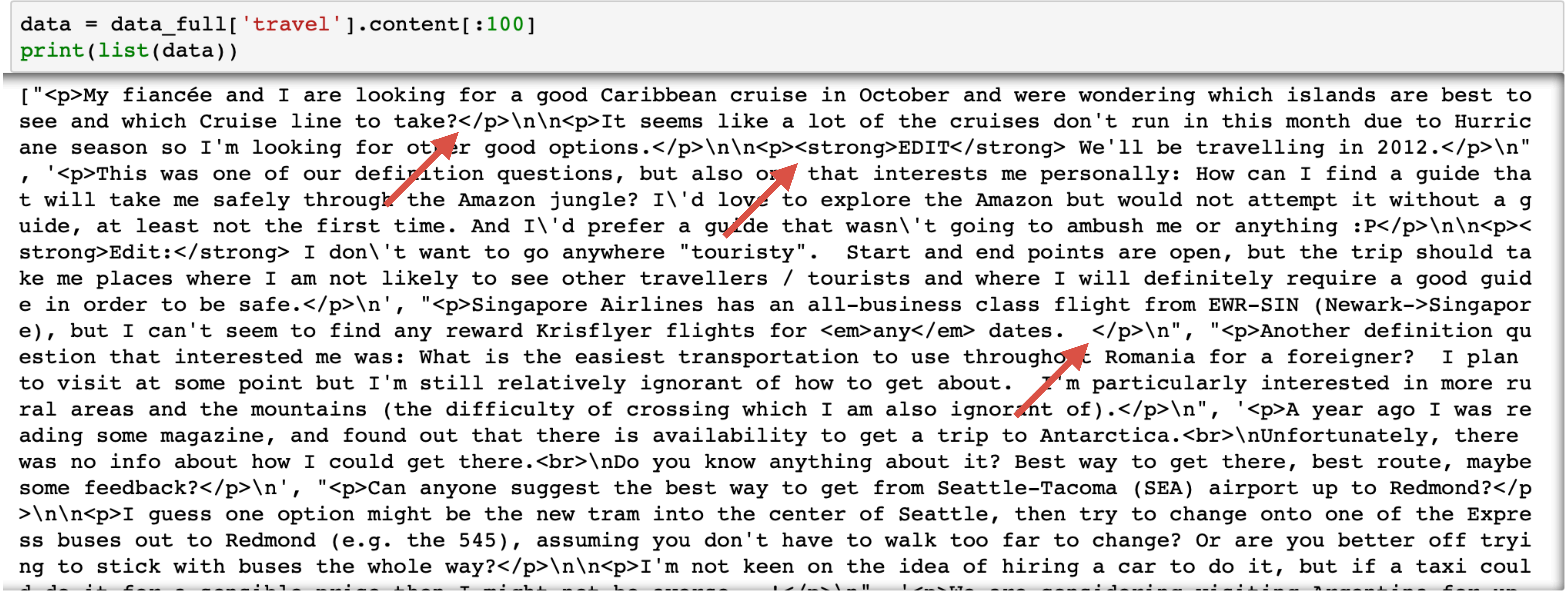
Поиск похожих текстов



```python
data = data_full['travel'].content[:100]
print(list(data))
```

```
["<p>My fiancée and I are looking for a good Caribbean cruise in October and were wondering which islands are best to
see and which Cruise line to take?</p>\n\n<p>It seems like a lot of the cruises don't run in this month due to Hurric
ane season so I'm looking for other good options.</p>\n\n<p><strong>EDIT</strong> We'll be travelling in 2012.</p>\n"
, '<p>This was one of our definition questions, but also one that interests me personally: How can I find a guide tha
t will take me safely through the Amazon jungle? I\'d love to explore the Amazon but would not attempt it without a g
uide, at least not the first time. And I\'d prefer a guide that wasn\'t going to ambush me or anything :P</p>\n\n<p><
strong>Edit:</strong> I don\'t want to go anywhere "touristy".  Start and end points are open, but the trip should ta
ke me places where I am not likely to see other travellers / tourists and where I will definitely require a good guid
e in order to be safe.</p>\n', "<p>Singapore Airlines has an all-business class flight from EWR-SIN (Newark->Singapor
e), but I can't seem to find any reward Krisflyer flights for <em>any</em> dates.  </p>\n", "<p>Another definition qu
estion that interested me was: What is the easiest transportation to use throughout Romania for a foreigner?  I plan
to visit at some point but I'm still relatively ignorant of how to get about.  I'm particularly interested in more ru
ral areas and the mountains (the difficulty of crossing which I am also ignorant of).</p>\n", '<p>A year ago I was re
ading some magazine, and found out that there is availability to get a trip to Antarctica.<br>\nUnfortunately, there
was no info about how I could get there.<br>\nDo you know anything about it? Best way to get there, best route, maybe
some feedback?</p>\n', "<p>Can anyone suggest the best way to get from Seattle-Tacoma (SEA) airport up to Redmond?</p
>\n\n<p>I guess one option might be the new tram into the center of Seattle, then try to change onto one of the Expre
ss buses out to Redmond (e.g. the 545), assuming you don't have to walk too far to change? Or are you better off tryi
ng to stick with buses the whole way?</p>\n\n<p>I'm not keen on the idea of hiring a car to do it, but if a taxi coul
d do it for a sensible price then I might not be averse.  '</p>\n", '<p>We are considering visiting Argentina for up
```

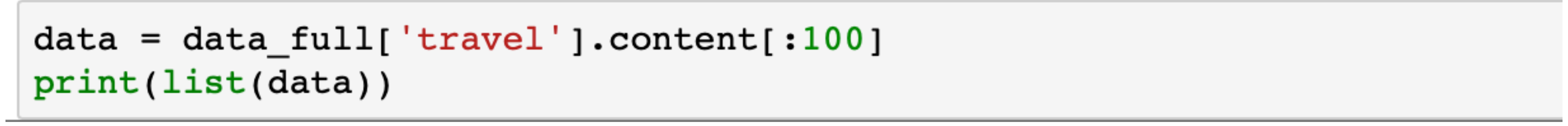
Поиск похожих текстов



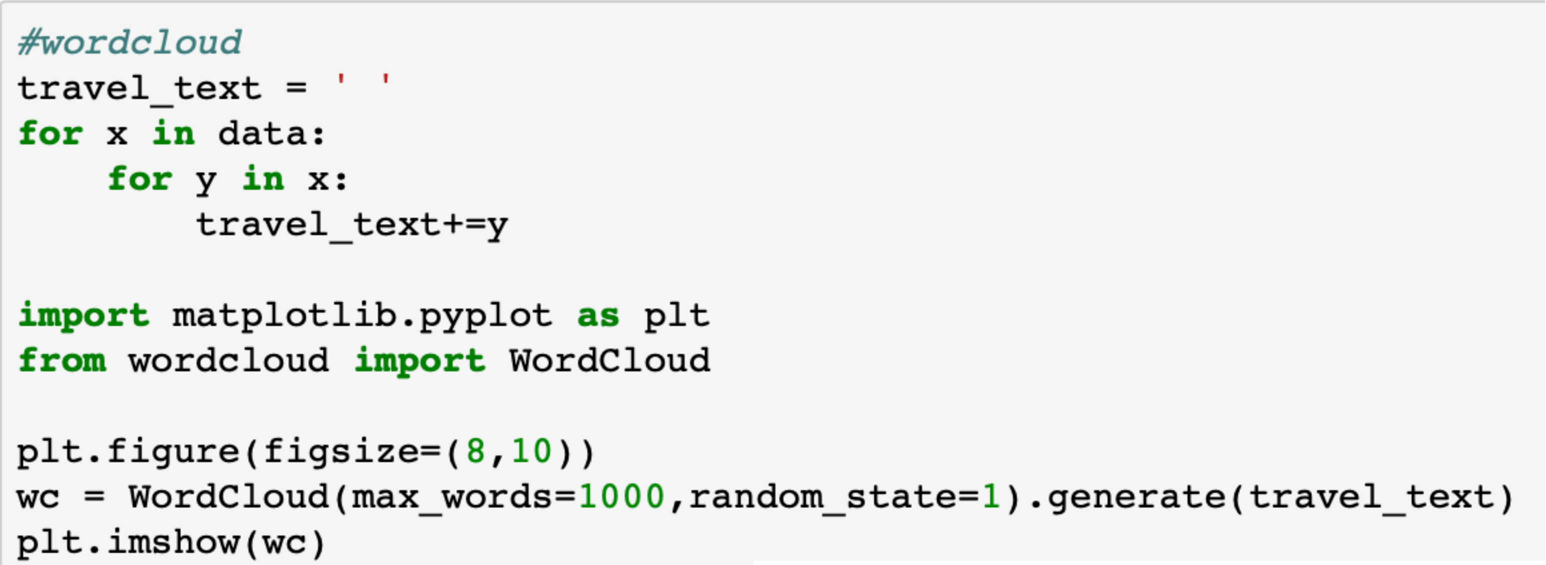
```
data = data_full['travel'].content[:100]
print(list(data))
```

```
["<p>My fiancée and I are looking for a good Caribbean cruise in October and were wondering which islands are best to
see and which Cruise line to take?</p>\n\n<p>It seems like a lot of the cruises don't run in this month due to Hurric
ane season so I'm looking for other good options.</p>\n\n<p><strong>EDIT</strong> We'll be travelling in 2012.</p>\n"
, '<p>This was one of our definition questions, but also one that interests me personally: How can I find a guide tha
t will take me safely through the Amazon jungle? I\'d love to explore the Amazon but would not attempt it without a g
uide, at least not the first time. And I\'d prefer a guide that wasn\'t going to ambush me or anything :P</p>\n\n<p><
strong>Edit:</strong> I don\'t want to go anywhere "touristy".  Start and end points are open, but the trip should ta
ke me places where I am not likely to see other travellers / tourists and where I will definitely require a good guid
e in order to be safe.</p>\n', "<p>Singapore Airlines has an all-business class flight from EWR-SIN (Newark->Singapor
e), but I can't seem to find any reward Krisflyer flights for <em>any</em> dates.  </p>\n", "<p>Another definition qu
estion that interested me was: What is the easiest transportation to use throughout Romania for a foreigner?  I plan
to visit at some point but I'm still relatively ignorant of how to get about.  I'm particularly interested in more ru
ral areas and the mountains (the difficulty of crossing which I am also ignorant of).</p>\n", '<p>A year ago I was re
ading some magazine, and found out that there is availability to get a trip to Antarctica.<br>\nUnfortunately, there
was no info about how I could get there.<br>\nDo you know anything about it? Best way to get there, best route, maybe
some feedback?</p>\n', "<p>Can anyone suggest the best way to get from Seattle-Tacoma (SEA) airport up to Redmond?</p
>\n\n<p>I guess one option might be the new tram into the center of Seattle, then try to change onto one of the Expre
ss buses out to Redmond (e.g. the 545), assuming you don't have to walk too far to change? Or are you better off tryi
ng to stick with buses the whole way?</p>\n\n<p>I'm not keen on the idea of hiring a car to do it, but if a taxi coul
d do it for a sensible price then I might not be averse.  '</p>\n", '<p>We are considering visiting Argentina for up
```

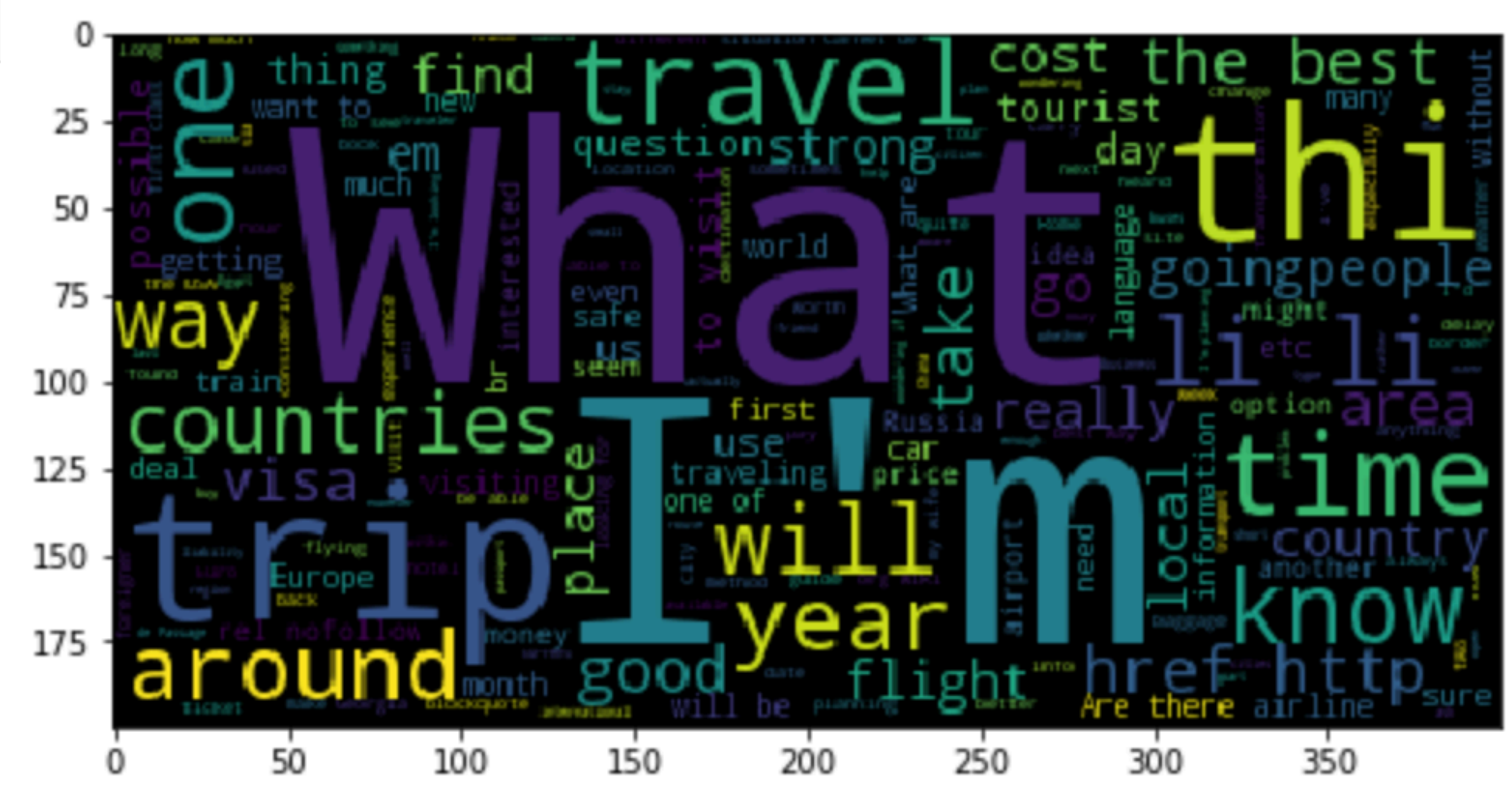
Облако слов



```
data = data_full['travel'].content[:100]
print(list(data))
```



```
#wordcloud
travel_text = ' '
for x in data:
    for y in x:
        travel_text+=y

import matplotlib.pyplot as plt
from wordcloud import WordCloud

plt.figure(figsize=(8,10))
wc = WordCloud(max_words=1000,random_state=1).generate(travel_text)
plt.imshow(wc)
plt.show()
```



Мера сходства

Необходимо определить меру сходства текстов (сообщений)

Одна из возможных мер - **редакционное расстояние** - минимальное число операций редактирования, обеспечивающее преобразование одного слова в другое.

Операции редактирования: вставка, удаление, замена символа.

Такой алгоритм является затратным.

Мера сходства

Необходимо определить меру сходства текстов (сообщений)

Обобщение

Расстояние между сообщениями - количество слов, которые необходимо добавить или удалить для преобразования одного текста в другой.

Такой подход неустойчив относительно порядка слов в предложении и аналогично трудозатратен.

Мера сходства

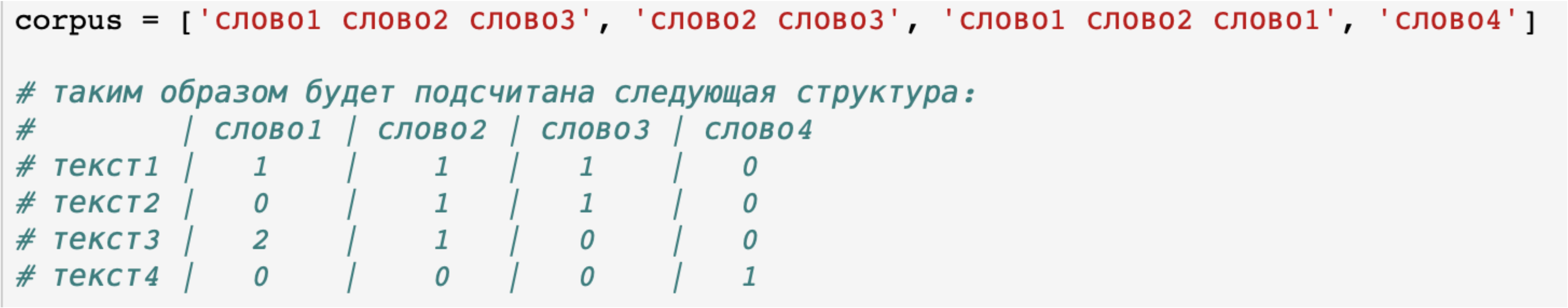
Необходимо определить меру сходства текстов (сообщений)

Если не учитывать порядок слов, то можем получить меру на основе **векторизации.**

Сообщению ставится в соответствие вектор пар «слово - количество вхождений».

Сообщения сравниваются на основе сравнения (поиска расстояния) векторов.

Векторизация

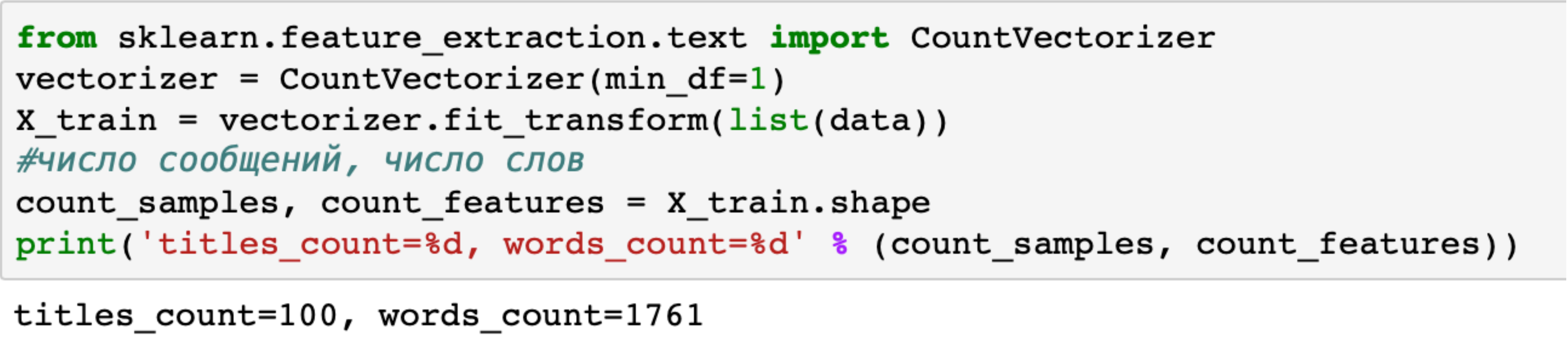


```
corpus = ['слово1 слово2 слово3', 'слово2 слово3', 'слово1 слово2 слово1', 'слово4']

# таким образом будет подсчитана следующая структура:
#        | слово1 | слово2 | слово3 | слово4
# текст1 |   1    |    1   |   1    |   0
# текст2 |   0    |    1   |   1    |   0
# текст3 |   2    |    1   |   0    |   0
# текст4 |   0    |    0   |   0    |   1
```

Один из минусов - необходимо указывать сразу все тексты, новые слова добавить в корпус не получится.

Векторизация

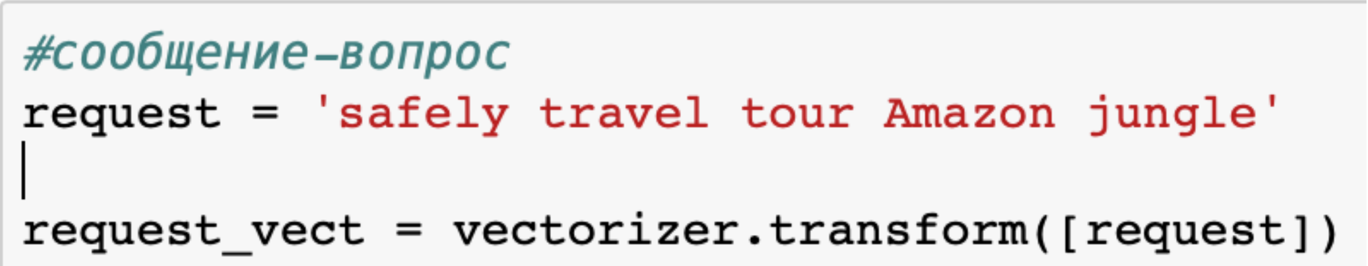
Создадим объект векторизатор, где *min_df* - минимальное число вхождений слова.



```
from sklearn.feature_extraction.text import CountVectorizer
vectorizer = CountVectorizer(min_df=1)
X_train = vectorizer.fit_transform(list(data))
#число сообщений, число слов
count_samples, count_features = X_train.shape
print('titles_count=%d, words_count=%d' % (count_samples, count_features))
```

```
titles_count=100, words_count=1761
```

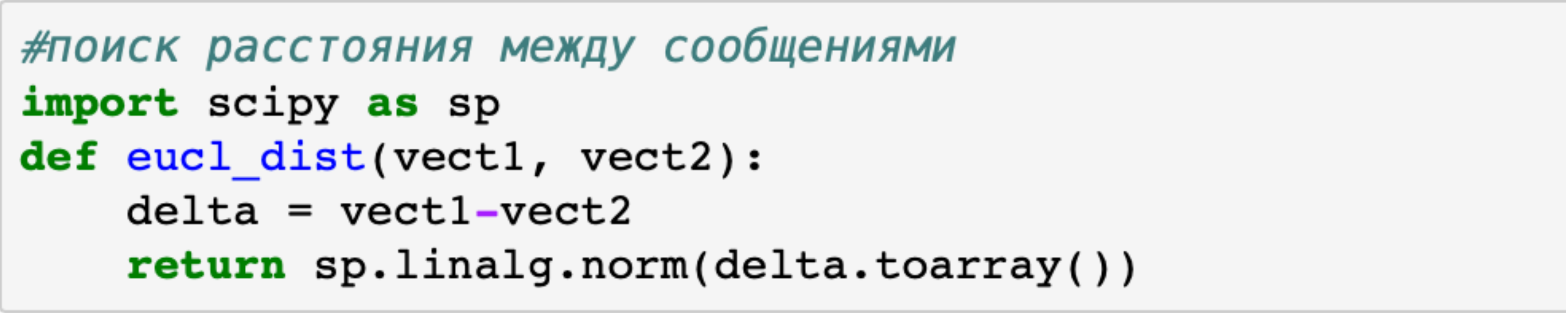
Создадим сообщение-вопрос и построим для него вектор.



```
#сообщение-вопрос
request = 'safely travel tour Amazon jungle'

request_vect = vectorizer.transform([request])
```

Поиск похожего текста

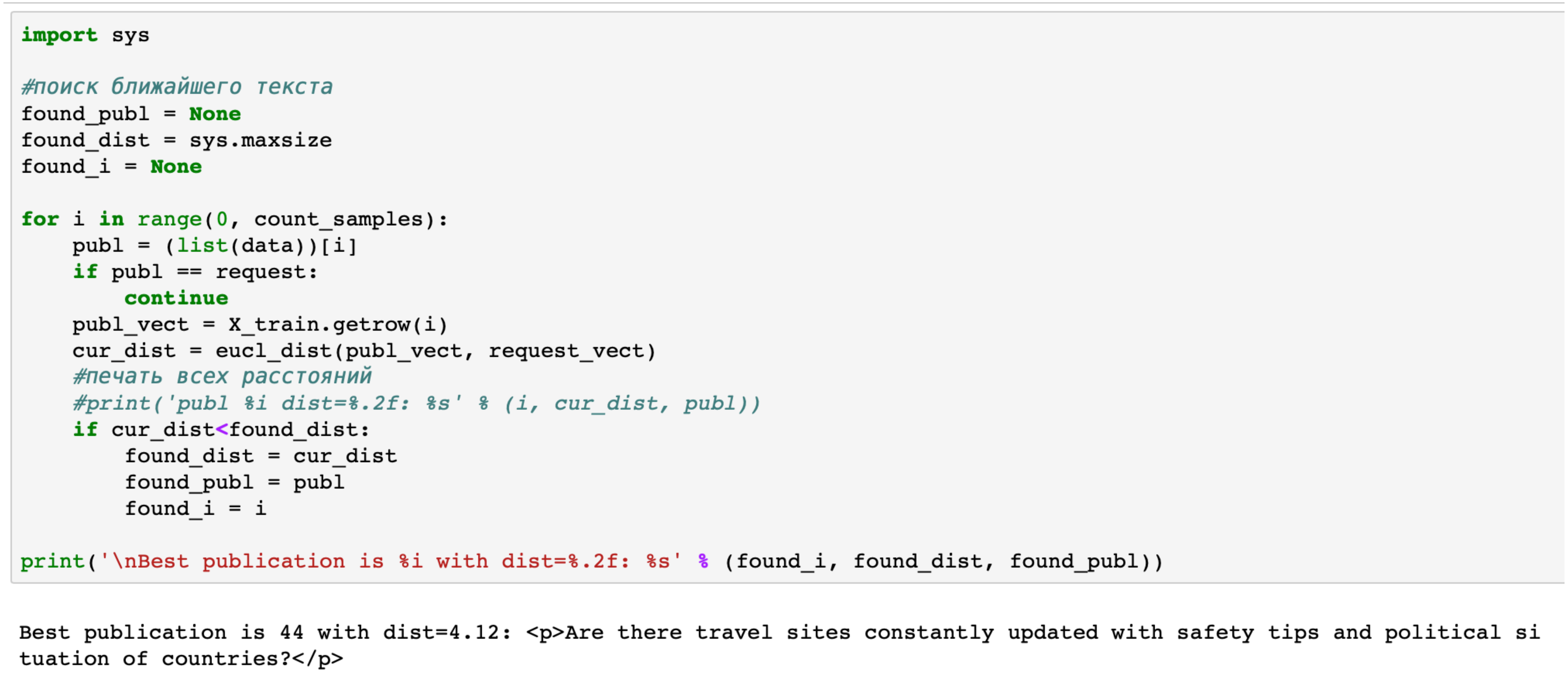
Создадим метод вычисления евклидова расстояния между векторами.



```python
#поиск расстояния между сообщениями
import scipy as sp
def eucl_dist(vect1, vect2):
    delta = vect1-vect2
    return sp.linalg.norm(delta.toarray())
```

Поиск похожего текста

Выполним поиск вектора с минимальным расстоянием до вектора исходного сообщения.

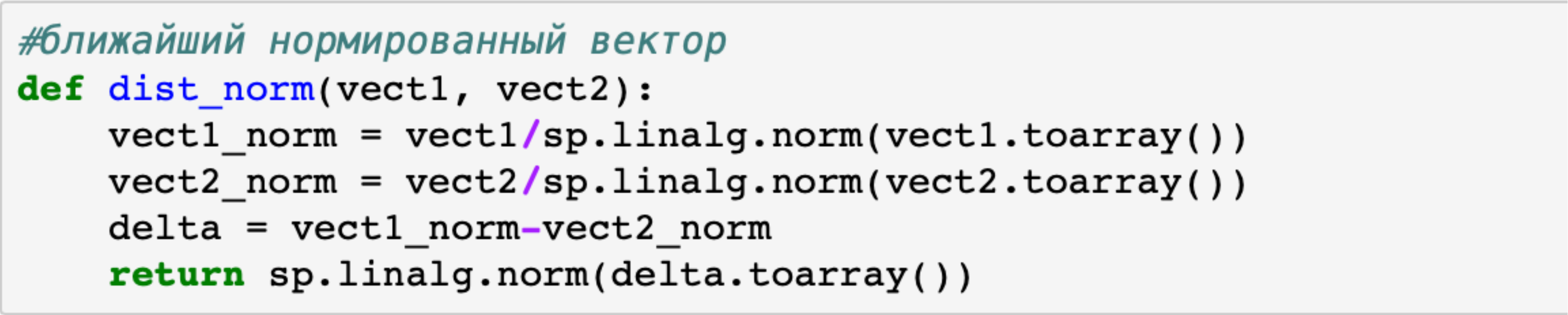


```
import sys

#поиск ближайшего текста
found_publ = None
found_dist = sys.maxsize
found_i = None

for i in range(0, count_samples):
    publ = (list(data))[i]
    if publ == request:
        continue
    publ_vect = X_train.getrow(i)
    cur_dist = eucl_dist(publ_vect, request_vect)
    #печать всех расстояний
    #print('publ %i dist=%.2f: %s' % (i, cur_dist, publ))
    if cur_dist<found_dist:
        found_dist = cur_dist
        found_publ = publ
        found_i = i

print('\nBest publication is %i with dist=%.2f: %s' % (found_i, found_dist, found_publ))
```

```
Best publication is 44 with dist=4.12: <p>Are there travel sites constantly updated with safety tips and political si
tuation of countries?</p>
```

Поиск похожего текста

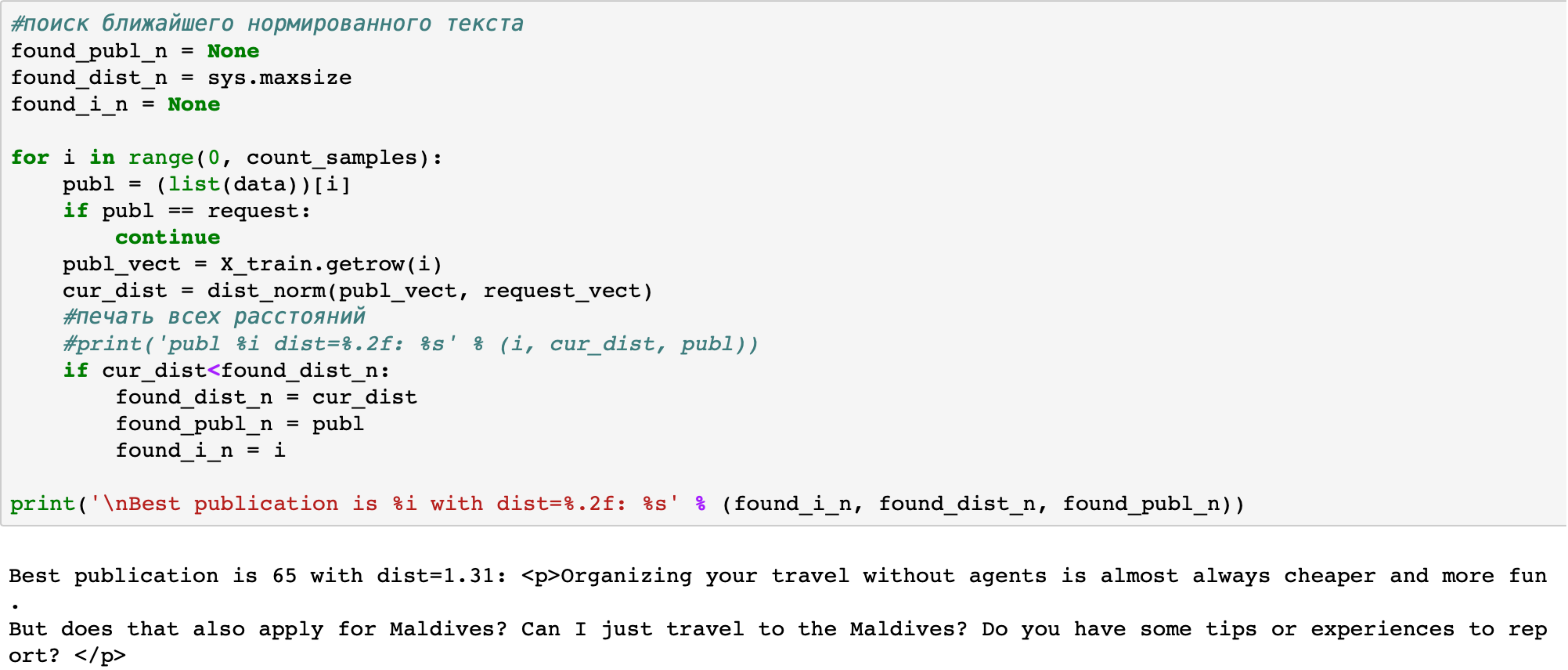
Нормируем векторы.



```
#ближайший нормированный вектор
def dist_norm(vect1, vect2):
    vect1_norm = vect1/sp.linalg.norm(vect1.toarray())
    vect2_norm = vect2/sp.linalg.norm(vect2.toarray())
    delta = vect1_norm-vect2_norm
    return sp.linalg.norm(delta.toarray())
```

Поиск похожего текста

Выполним поиск вектора с минимальным расстоянием до вектора исходного сообщения.



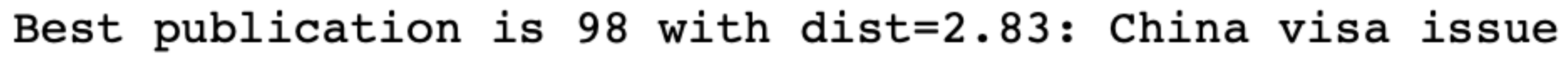
```python
#поиск ближайшего нормированного текста
found_publ_n = None
found_dist_n = sys.maxsize
found_i_n = None

for i in range(0, count_samples):
    publ = (list(data))[i]
    if publ == request:
        continue
    publ_vect = X_train.getrow(i)
    cur_dist = dist_norm(publ_vect, request_vect)
    #печать всех расстояний
    #print('publ %i dist=%.2f: %s' % (i, cur_dist, publ))
    if cur_dist<found_dist_n:
        found_dist_n = cur_dist
        found_publ_n = publ
        found_i_n = i

print('\nBest publication is %i with dist=%.2f: %s' % (found_i_n, found_dist_n, found_publ_n))
```

```
Best publication is 65 with dist=1.31: <p>Organizing your travel without agents is almost always cheaper and more fun
.
But does that also apply for Maldives? Can I just travel to the Maldives? Do you have some tips or experiences to rep
ort? </p>
```

Поиск похожего текста

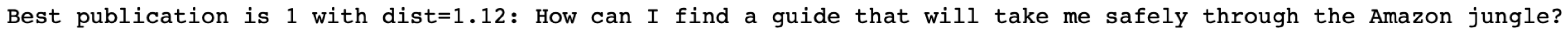
Возьмем для наглядности не содержимое статей, а их заголовки.

Наиболее похожие текст ДО нормирования:



```
Best publication is 98 with dist=2.83: China visa issue
```

Наиболее похожие текст ПОСЛЕ нормирования:



```
Best publication is 1 with dist=1.12: How can I find a guide that will take me safely through the Amazon jungle?
```

Исходное сообщение: «safely travel tour Amazon jungle»

Задание 2

1. Построить облако слов по корпусу текстов;
2. Создать векторизатор по корпусу текстов;
3. Определить меру сходства текстов, которую вы будете использовать, и выполнить поиск ближайших текстов.