Photo Credit: Pixabay

Building a Content Based Recommender System for Hotels in Seattle

How to use description of a hotel to recommend similar hotels.

Susan Li Follow
Apr 1 · 5 min read

The cold start problem is a well known and well researched problem for recommender systems, where system is not able to recommend items to users. due to three different situation i.e. for new users, for new products and for new websites.

Content-based filtering is the method that solve this problem. Our system first uses the metadata of new products when creating recommendations, while visitor action is secondary for a certain period of time. And our systems recommend a product to a user based upon the category and description of the product.

Content-based recommendation systems may be used in a variety of domains ranging from recommending web pages, news articles, restaurants, television programs, and hotels. The advantage of content-based filtering is that it doesn't have a cold-start problem. If you just start out a new website, or any new products can be recommended right away.

Let's assume we are starting a new online travel agency (OTA), and we have signed up thousands of hotels that are willing to sell on our platform, and we start seeing traffic coming from our website users, but we don't have any users history, therefore, we are going to build a content-based recommendation systems to analyze hotel descriptions to identify hotels that are of particular interest to the user.

We would like to recommend hotels based on the hotels that a user has already booked or viewed using the cosine similarity. We would recommend hotels with the largest similarity to the ones previously booked or viewed or showed interest by the user. Our recommender system is highly dependent on defining an appropriate similarity measure. Eventually, we select a subset of hotels to display to the user or to determine an order in which to display the hotels.

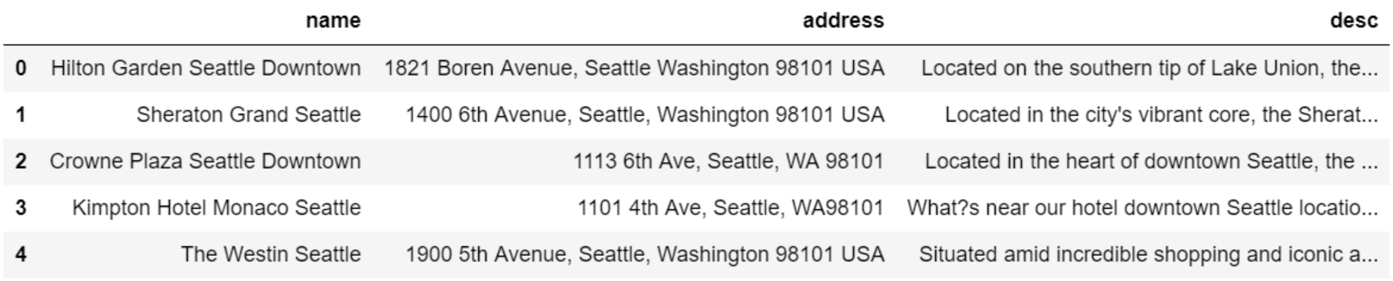
The Data

It's very hard to find public available hotel description data, therefore, I collected them by myself from each hotel's homepage for over 150 hotels in Seattle area, that includes downtown business hotels, boutique hotels and bed and breakfast, airport business hotels, inns near the universities, motels in the middle of nowhere, and so on. The data can be found here.

```
import pandas as pd
import numpy as np
from nltk.corpus import stopwords
from sklearn.metrics.pairwise import linear_kernel
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.decomposition import LatentDirichletAllocation
import re
import random
import plotly.graph_objs as go
import plotly.plotly as py
import cufflinks
```

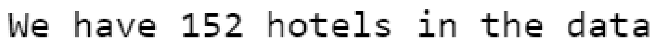
```
pd.options.display.max_columns = 30
from IPython.core.interactiveshell import InteractiveShell
import plotly.figure_factory as ff
InteractiveShell.ast_node_interactivity = 'all'
from plotly.offline import iplot
cufflinks.go_offline()
cufflinks.set_config_file(world_readable=True,
theme='solar')
```

```
df = pd.read_csv('Seattle_Hotels.csv', encoding="latin-1")
df.head()
print('We have ', len(df), 'hotels in the data')
```



	name	address	desc
0	Hilton Garden Seattle Downtown	1821 Boren Avenue, Seattle Washington 98101 USA	Located on the southern tip of Lake Union, the...
1	Sheraton Grand Seattle	1400 6th Avenue, Seattle, Washington 98101 USA	Located in the city's vibrant core, the Sherat...
2	Crowne Plaza Seattle Downtown	1113 6th Ave, Seattle, WA 98101	Located in the heart of downtown Seattle, the ...
3	Kimpton Hotel Monaco Seattle	1101 4th Ave, Seattle, WA98101	What?s near our hotel downtown Seattle locatio...
4	The Westin Seattle	1900 5th Avenue, Seattle, Washington 98101 USA	Situated amid incredible shopping and iconic a...

Table 1



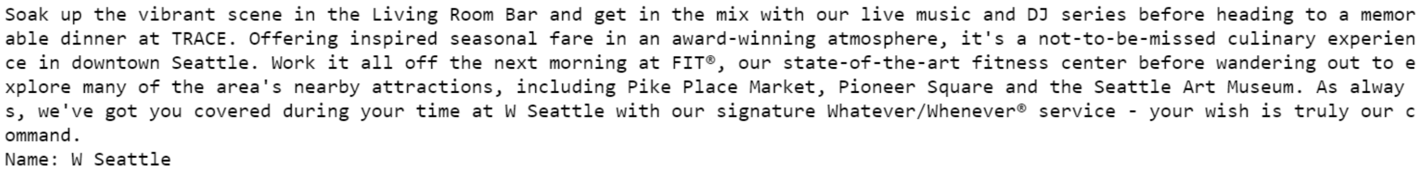
```
We have 152 hotels in the data
```

Have a look few hotel name and description pairs.

```
def print_description(index):
    example = df[df.index == index][['desc', 'name']].v
    if len(example) > 0:
        print(example[0])
```

print_description.py

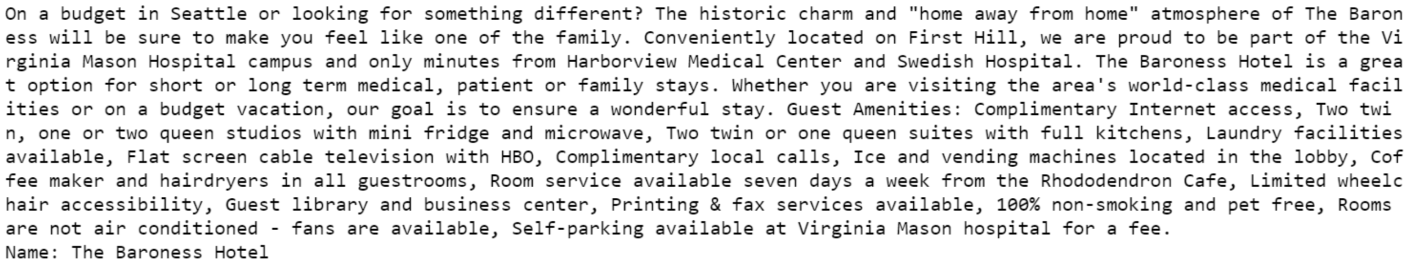
```
print_description(10)
```



```
Soak up the vibrant scene in the Living Room Bar and get in the mix with our live music and DJ series before heading to a memor
able dinner at TRACE. Offering inspired seasonal fare in an award-winning atmosphere, it's a not-to-be-missed culinary experien
ce in downtown Seattle. Work it all off the next morning at FIT®, our state-of-the-art fitness center before wandering out to e
xplore many of the area's nearby attractions, including Pike Place Market, Pioneer Square and the Seattle Art Museum. As alway
s, we've got you covered during your time at W Seattle with our signature Whatever/Whenever® service - your wish is truly our c
ommand.
Name: W Seattle
```

Figure 1

```
print_description(100)
```



```
On a budget in Seattle or looking for something different? The historic charm and "home away from home" atmosphere of The Baron
ess will be sure to make you feel like one of the family. Conveniently located on First Hill, we are proud to be part of the Vi
rginia Mason Hospital campus and only minutes from Harborview Medical Center and Swedish Hospital. The Baroness Hotel is a grea
t option for short or long term medical, patient or family stays. Whether you are visiting the area's world-class medical facil
ities or on a budget vacation, our goal is to ensure a wonderful stay. Guest Amenities: Complimentary Internet access, Two twi
n, one or two queen studios with mini fridge and microwave, Two twin or one queen suites with full kitchens, Laundry facilities
available, Flat screen cable television with HBO, Complimentary local calls, Ice and vending machines located in the lobby, Cof
fee maker and hairdryers in all guestrooms, Room service available seven days a week from the Rhododendron Cafe, Limited wheelc
hair accessibility, Guest library and business center, Printing & fax services available, 100% non-smoking and pet free, Rooms
are not air conditioned - fans are available, Self-parking available at Virginia Mason hospital for a fee.
Name: The Baroness Hotel
```

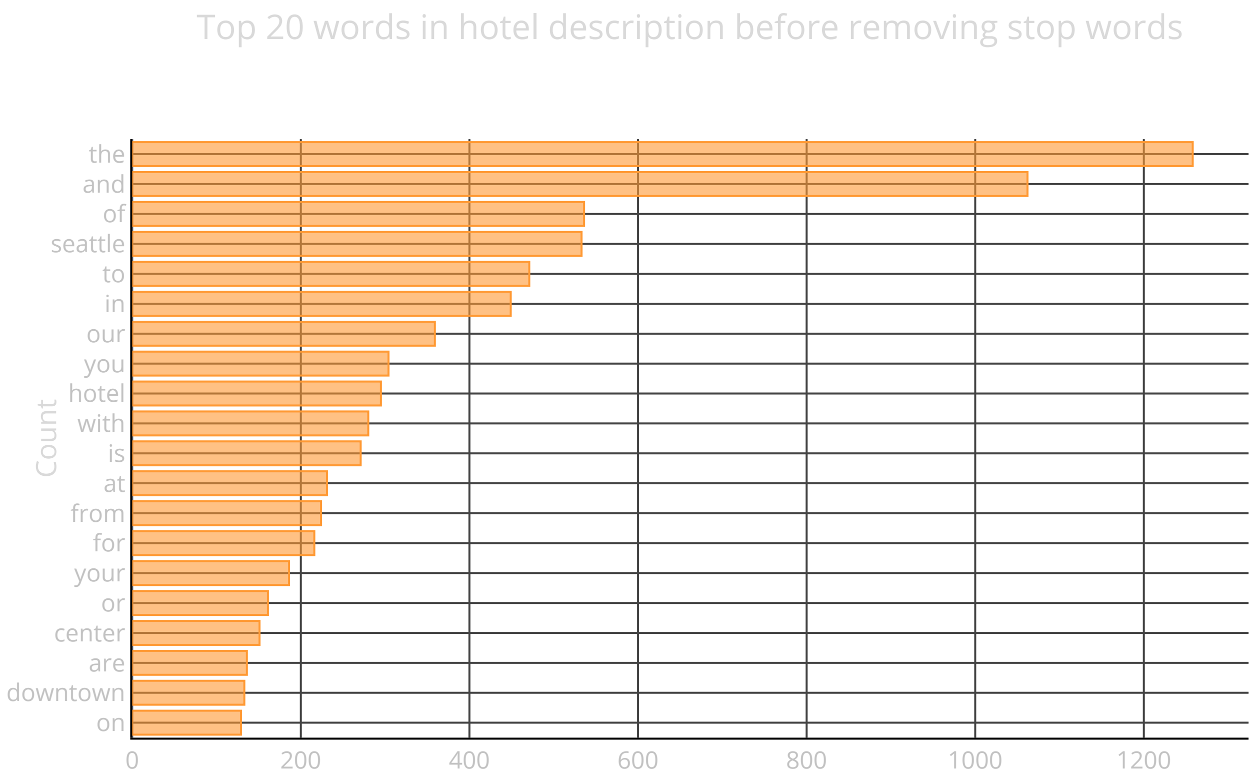
Figure 2

EDA

Token (vocabulary) Frequency Distribution Before Removing Stop Words

```
def get_top_n_words(corpus, n=None):
    vec = CountVectorizer().fit(corpus)
    bag_of_words = vec.transform(corpus)
    sum_words = bag_of_words.sum(axis=0)
    words_freq = [(word, sum_words[0, idx]) for word,
    words_freq =sorted(words_freq, key = lambda x: x[1
    return words_freq[:n]
common_words = get_top_n_words(df['desc'], 20)
```

unigram_distribution.py



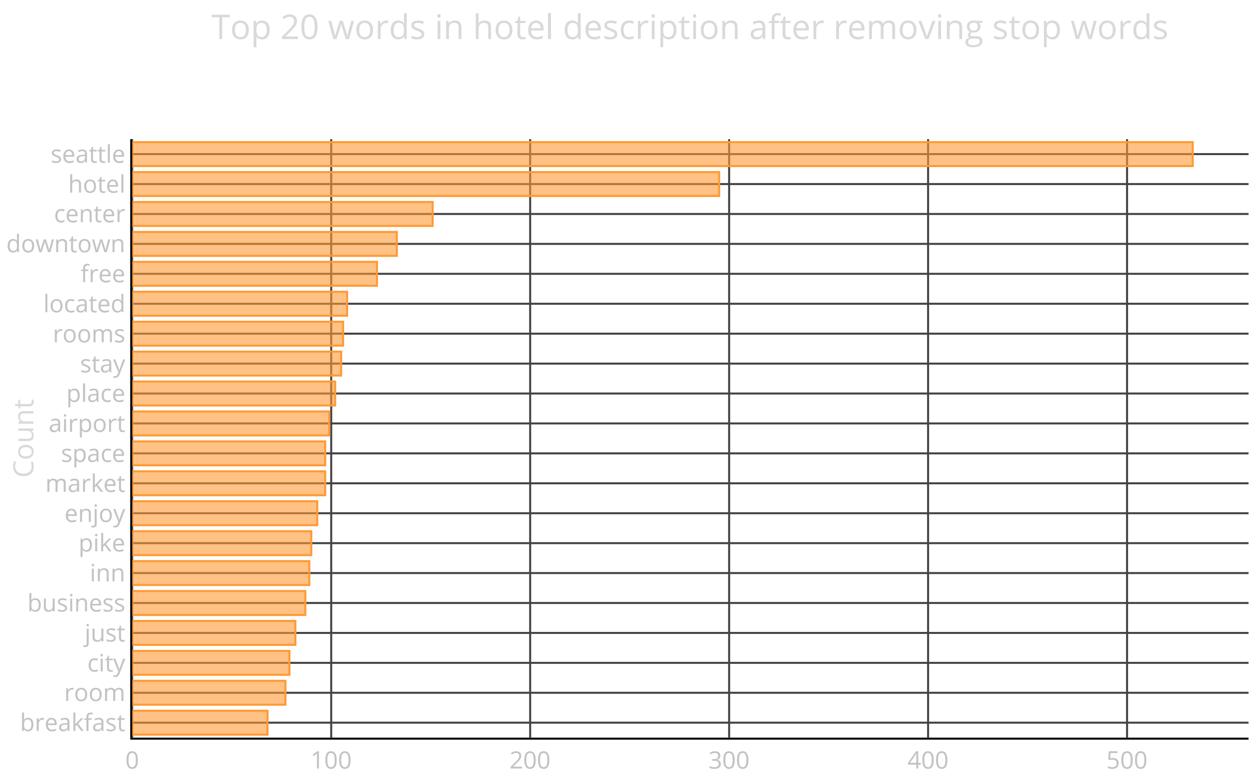
EDIT CHART

Figure 3

Token (vocabulary) Frequency Distribution After Removing Stop Words

```
def get_top_n_words(corpus, n=None):
    vec = CountVectorizer(stop_words='english').fit(cc
    bag_of_words = vec.transform(corpus)
    sum_words = bag_of_words.sum(axis=0)
    words_freq = [(word, sum_words[0, idx]) for word,
    words_freq =sorted(words_freq, key = lambda x: x[1
    return words_freq[:n]
common_words = get_top_n_words(df['desc'], 20)
```

unigram_distribution_stopwords_removed.py



EDIT CHART

Figure 4

Bigrams Frequency Distribution Before Removing Stop Words

```
def get_top_n_bigram(corpus, n=None):
    vec = CountVectorizer(ngram_range=(2, 2)).fit(corp
    bag_of_words = vec.transform(corpus)
    sum_words = bag_of_words.sum(axis=0)
    words_freq = [(word, sum_words[0, idx]) for word,
    words_freq =sorted(words_freq, key = lambda x: x[1
    return words_freq[:n]
common_words = get_top_n_bigram(df['desc'], 20)
```

bigrams_distribution.py

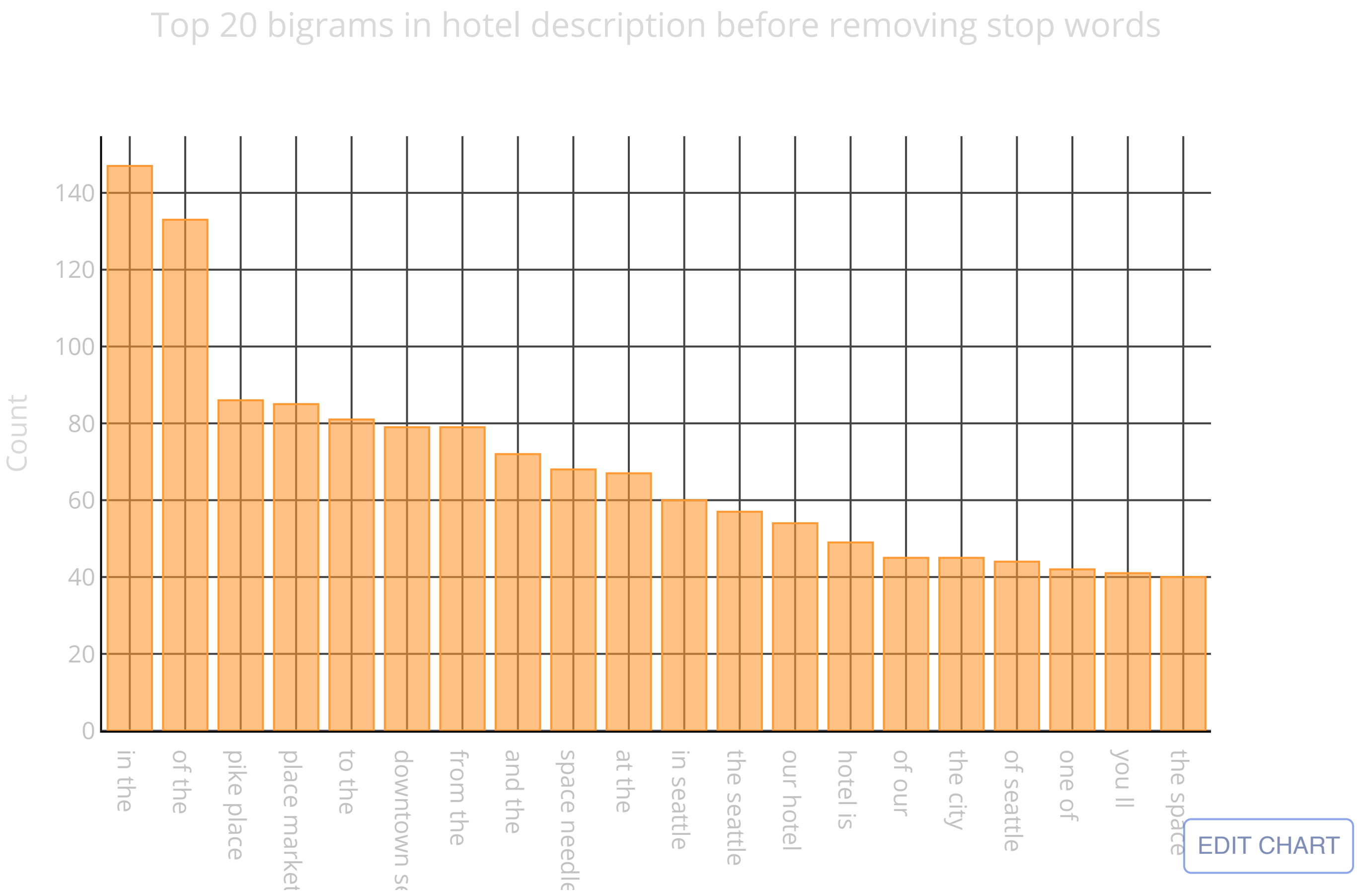


Figure 5

Bigrams Frequency Distribution After Removing Stop Words

```
def get_top_n_bigram(corpus, n=None):
    vec = CountVectorizer(ngram_range=(2, 2), stop_wor
    bag_of_words = vec.transform(corpus)
    sum_words = bag_of_words.sum(axis=0)
    words_freq = [(word, sum_words[0, idx]) for word,
    words_freq =sorted(words_freq, key = lambda x: x[1
    return words_freq[:n]
common_words = get_top_n_bigram(df['desc'], 20)
```

bigrams_distribution_stopwords_removed.py

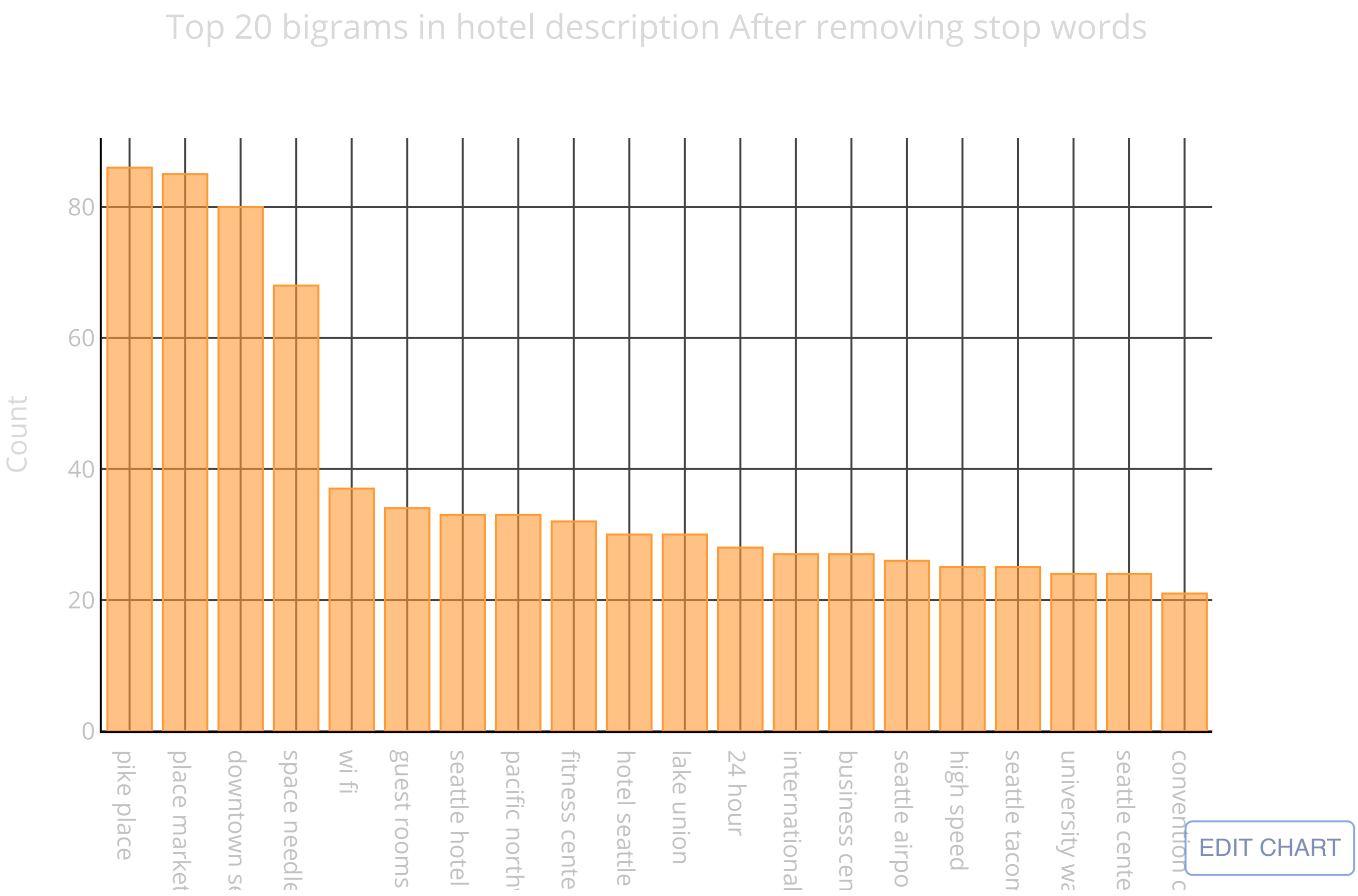


Figure 6

Trigrams Frequency Distribution Before Removing Stop Words

```
def get_top_n_trigram(corpus, n=None):
    vec = CountVectorizer(ngram_range=(3, 3)).fit(corp
    bag_of_words = vec.transform(corpus)
    sum_words = bag_of_words.sum(axis=0)
    words_freq = [(word, sum_words[0, idx]) for word,
    words_freq =sorted(words_freq, key = lambda x: x[1
    return words_freq[:n]
common_words = get_top_n_trigram(df['desc'], 20)
```

trigrams_distribution.py

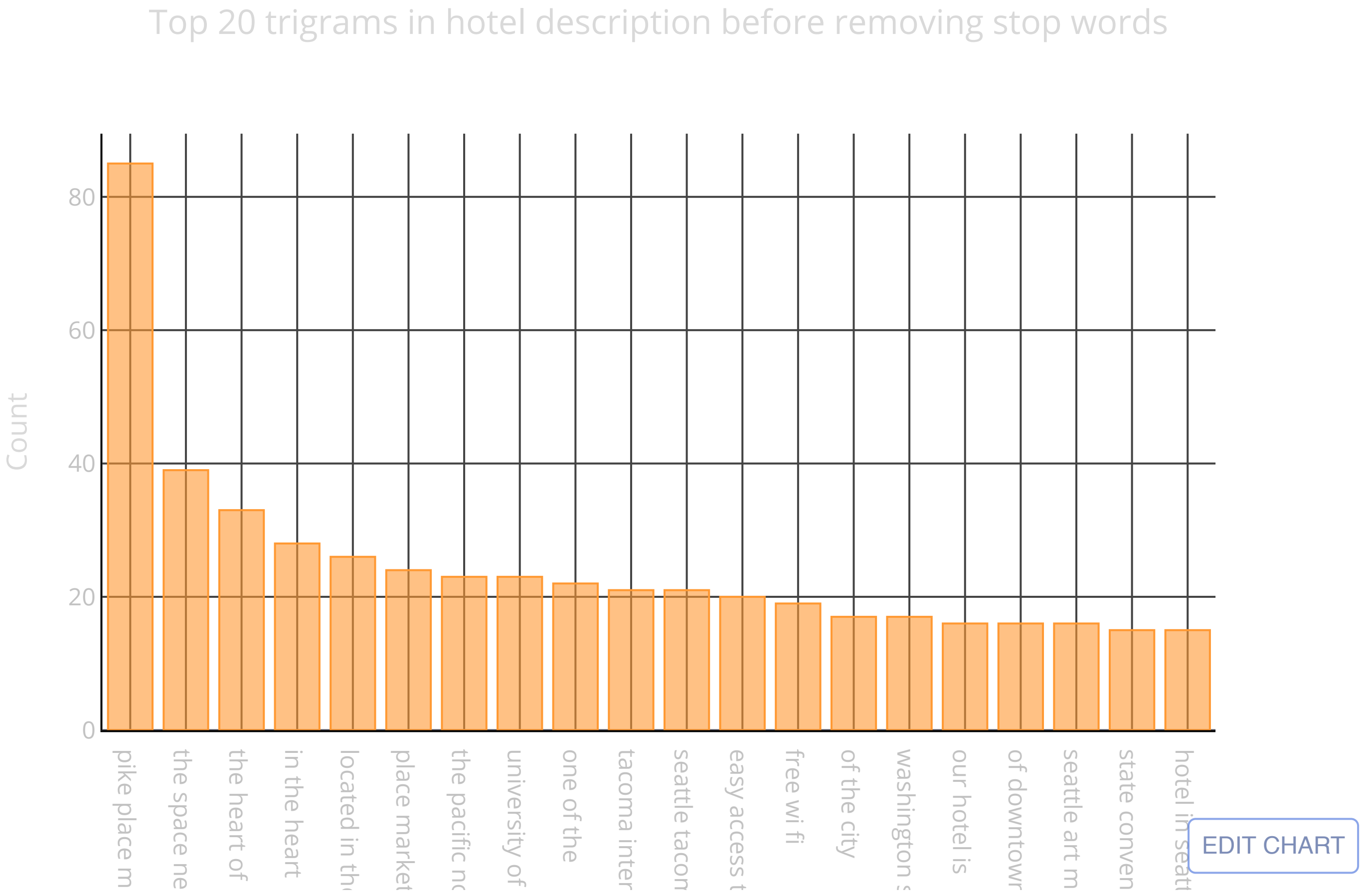


Figure 7

Trigrams Frequency Distribution After Removing Stop Words

```
def get_top_n_trigram(corpus, n=None):
    vec = CountVectorizer(ngram_range=(3, 3), stop_wor
    bag_of_words = vec.transform(corpus)
    sum_words = bag_of_words.sum(axis=0)
    words_freq = [(word, sum_words[0, idx]) for word,
    words_freq =sorted(words_freq, key = lambda x: x[1
    return words_freq[:n]
common_words = get_top_n_trigram(df['desc'], 20)
```

trigrams_distribution_stopwords_removed.py

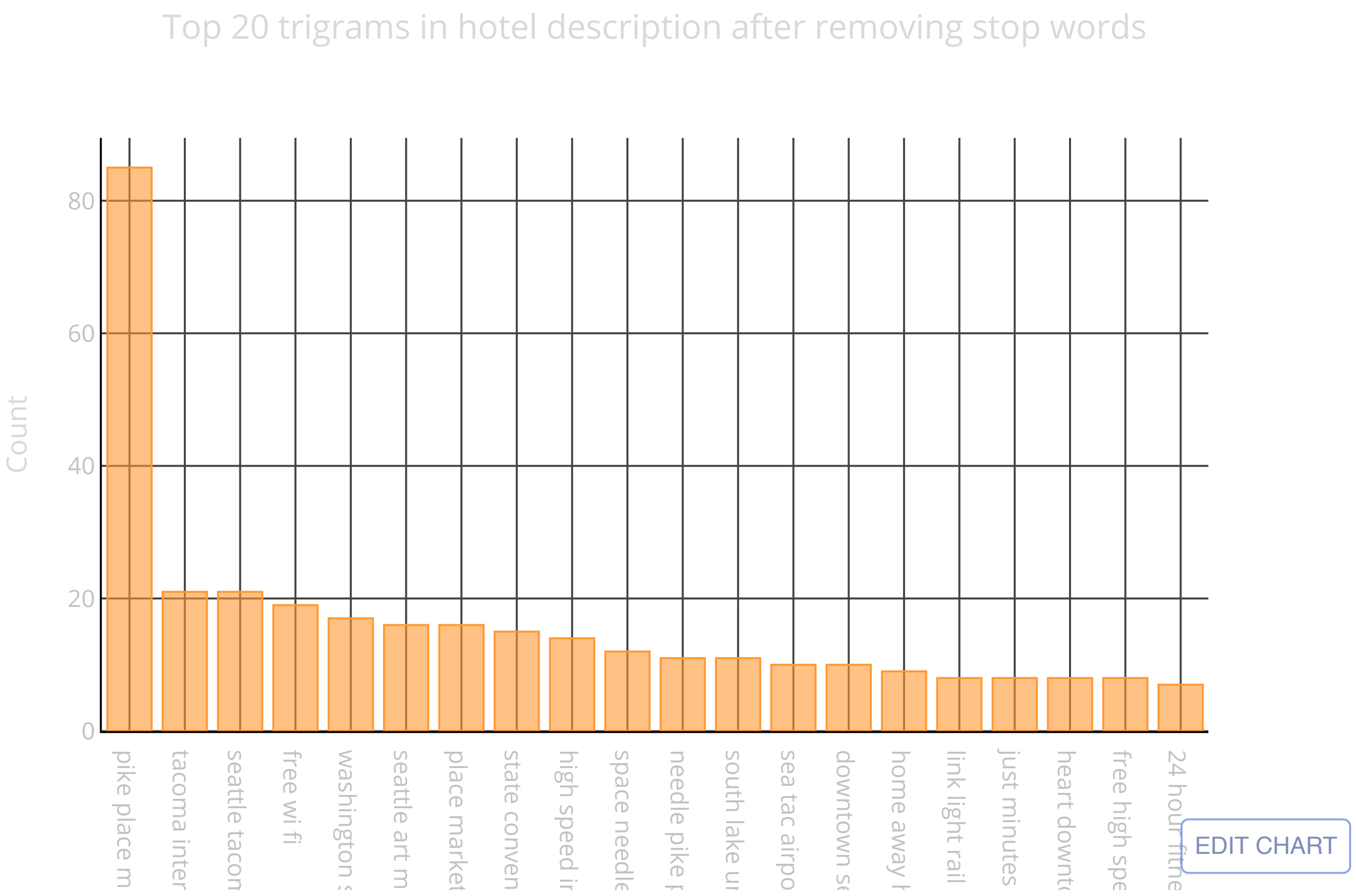
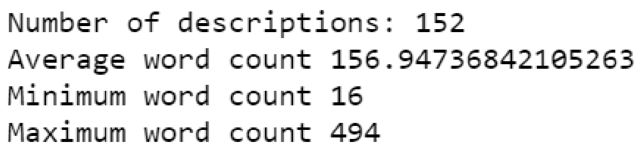


Figure 8

Everyone knows Seattle’s Pike Place Market, it is way more than a public farmers market. It is a historical vibrant tourism attraction comprised of hundreds of farmers, craftspeople, small businesses. The hotel industry thrives on location, tourists look for a hotel that is possibly nearest to downtown and / or must-visit attractions of the city. Therefore, every hotel would brag about it if it is not too far from the hotel.

Hotel Description Word Count Distribution

```
df['word_count'] = df['desc'].apply(lambda x:
len(str(x).split()))
desc_lengths = list(df['word_count'])


print("Number of descriptions:",len(desc_lengths),
      "\nAverage word count", np.average(desc_lengths),
      "\nMinimum word count", min(desc_lengths),
      "\nMaximum word count", max(desc_lengths))
```



```
Number of descriptions: 152
Average word count 156.94736842105263
Minimum word count 16
Maximum word count 494
```

```
df['word_count'].iplot(
    kind='hist',
    bins = 50,
    linecolor='black',
    xTitle='word count',
    yTitle='count'
```

word_count_distribution.py

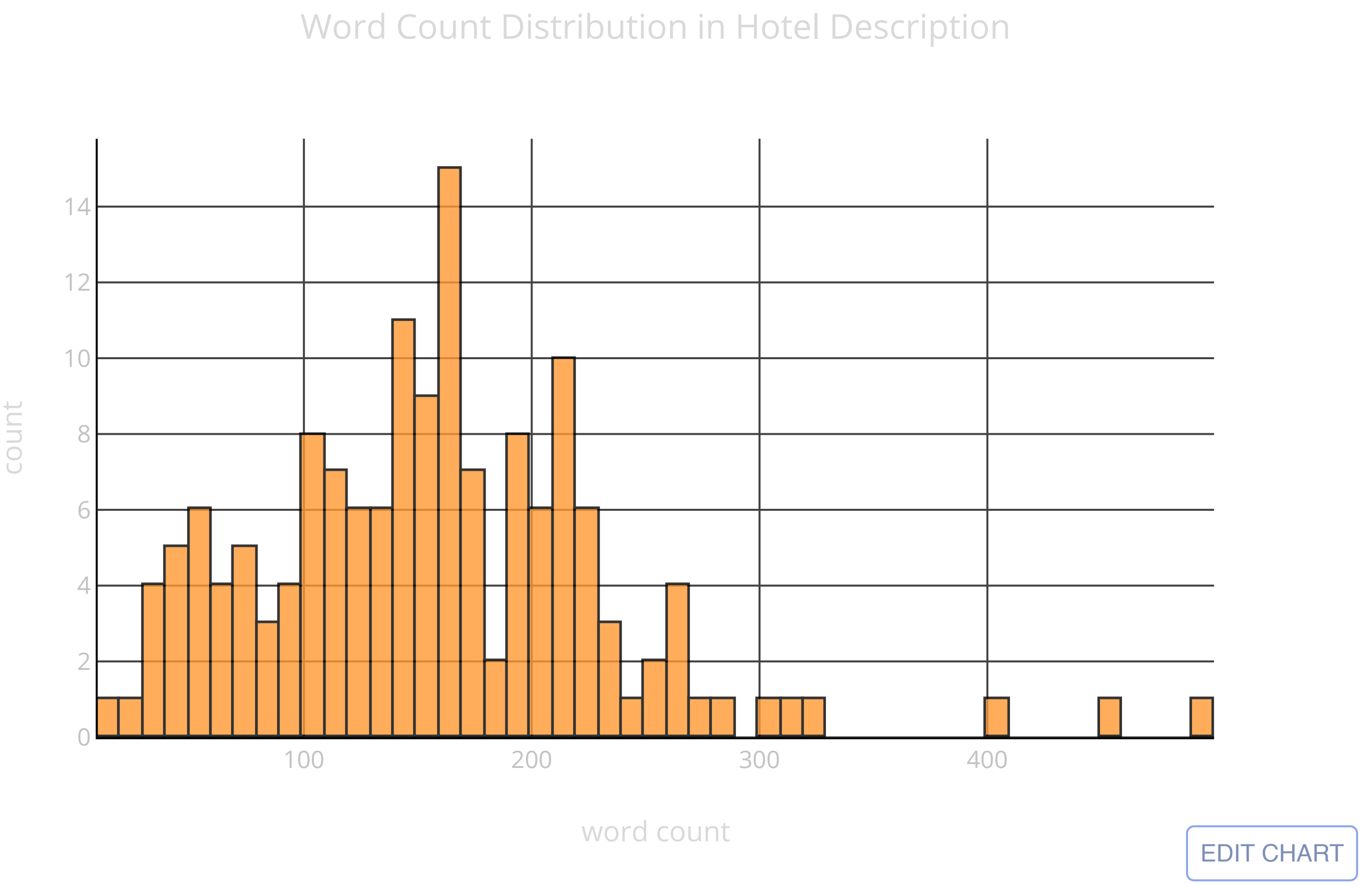


Figure 9

Many hotels use description to their full potential, know how to utilize captivating descriptions to appeal to travelers' emotions to drive direct bookings. Their descriptions may be longer than others.

Text Preprocessing

The test is pretty clean, we don’t have a lot to do, but just in case.

```
REPLACE_BY_SPACE_RE = re.compile('[/(){}\[\]\|@,;]')
BAD_SYMBOLS_RE = re.compile('[^0-9a-z #+_]')
STOPWORDS = set(stopwords.words('english'))

def clean_text(text):
    """
        text: a string

        return: modified initial string
    """
    text = text.lower() # lowercase text
    text = REPLACE BY SPACE RE.sub(' ', text) # replac
```

description_preprocessing.py

Modeling

- Create a TF-IDF matrix of unigrams, bigrams, and trigrams for each hotel.
- Compute similarity between all hotels using sklearn’s linear_kernel (equivalent to cosine similarity in our case).
- Define a function that takes in hotel name as input and returns the top 10 recommended hotels.

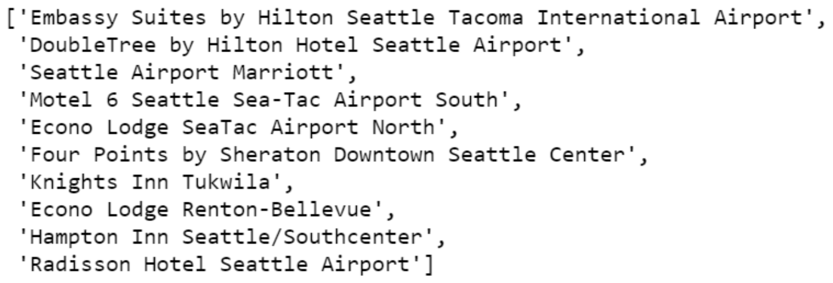
```python
df.set_index('name', inplace = True)
tf = TfidfVectorizer(analyzer='word', ngram_range=(1,
tfidf_matrix = tf.fit_transform(df['desc_clean'])
cosine_similarities = linear_kernel(tfidf_matrix, tfid

indices = pd.Series(df.index)

def recommendations(name, cosine_similarities = cosine

    recommended_hotels = []

    # gettin the index of the hotel that matches the r
    idx = indices[indices == name].index[0]

    # creating a Series with the similarity scores in
    score_series = pd.Series(cosine_similarities[idx])

```

hotel_rec_model.py

Recommendations

Let’s make some recommendations!

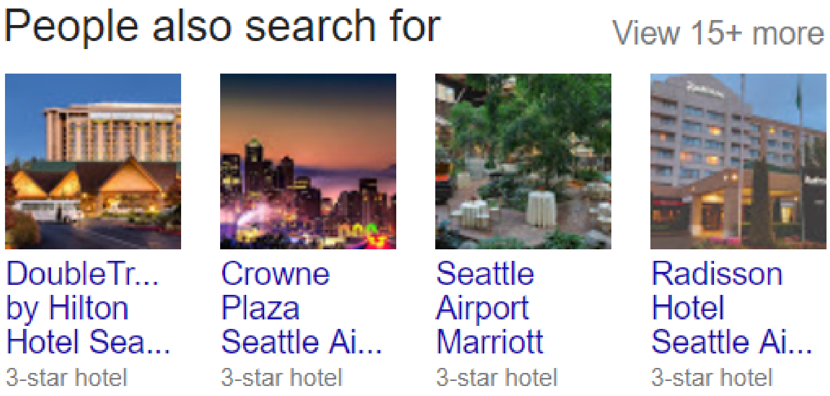
```
recommendations('Hilton Seattle Airport & Conference
Center')
```



```
['Embassy Suites by Hilton Seattle Tacoma International Airport',
 'DoubleTree by Hilton Hotel Seattle Airport',
 'Seattle Airport Marriott',
 'Motel 6 Seattle Sea-Tac Airport South',
 'Econo Lodge SeaTac Airport North',
 'Four Points by Sheraton Downtown Seattle Center',
 'Knights Inn Tukwila',
 'Econo Lodge Renton-Bellevue',
 'Hampton Inn Seattle/Southcenter',
 'Radisson Hotel Seattle Airport']
```

A good test on whether our similarity works is that the content based recommender returns all airport hotels when an airport hotel is a seed.

We can also ask Google. The following are recommended by Google for “Hilton Seattle Airport & Conference Center”:



People also search for View 15+ more

DoubleTr... by Hilton Hotel Sea... 3-star hotel

Crowne Plaza Seattle Ai... 3-star hotel

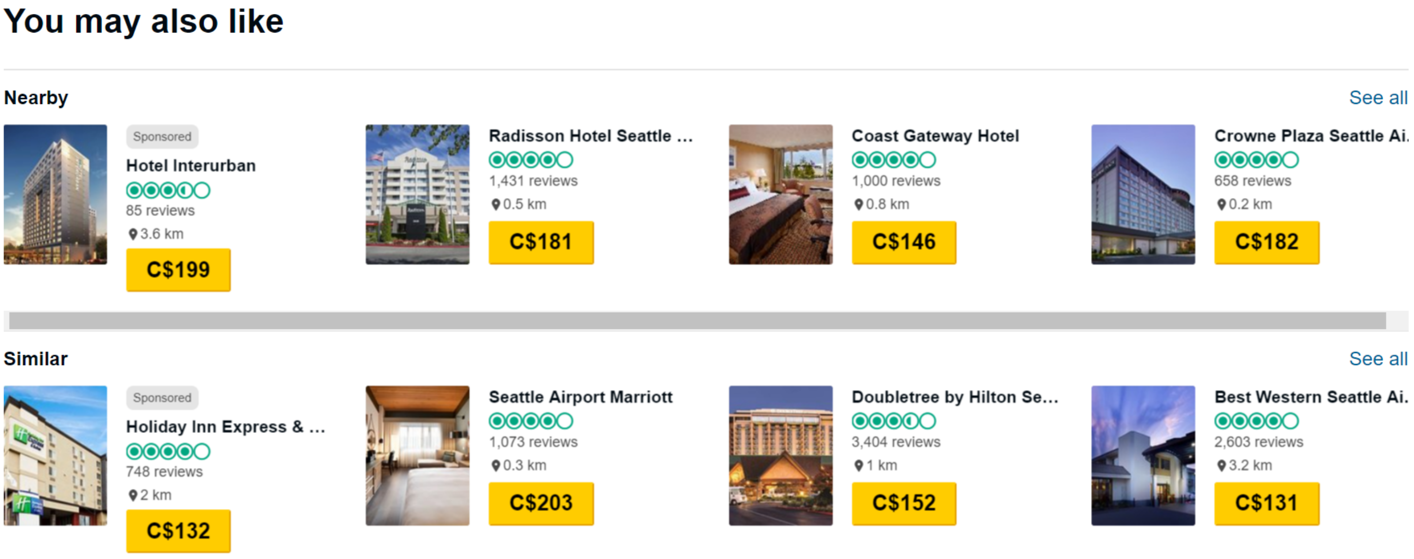
Seattle Airport Marriott 3-star hotel

Radisson Hotel Seattle Ai... 3-star hotel

Figure 10

Three out of four recommended by Google were also recommended by us.

The following are recommended by tripadvisor for “Hilton Seattle Airport & Conference Center”:



You may also like

Nearby See all

Sponsored Hotel Interurban 85 reviews 3.6 km C$199

Radisson Hotel Seattle ... 1,431 reviews 0.5 km C$181

Coast Gateway Hotel 1,000 reviews 0.8 km C$146

Crowne Plaza Seattle Ai. 658 reviews 0.2 km C$182

Similar See all

Sponsored Holiday Inn Express & ... 748 reviews 2 km C$132

Seattle Airport Marriott 1,073 reviews 0.3 km C$203

Doubletree by Hilton Se... 3,404 reviews 1 km C$152

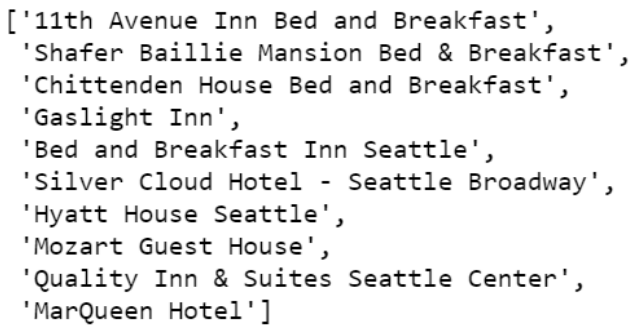
Best Western Seattle Ai. 2,603 reviews 3.2 km C$131

Figure 11

Not bad either.

Try a bed & breakfast.

```
recommendations("The Bacon Mansion Bed and Breakfast")
```



```
['11th Avenue Inn Bed and Breakfast',
 'Shafer Baillie Mansion Bed & Breakfast',
 'Chittenden House Bed and Breakfast',
 'Gaslight Inn',
 'Bed and Breakfast Inn Seattle',
 'Silver Cloud Hotel - Seattle Broadway',
 'Hyatt House Seattle',
 'Mozart Guest House',
 'Quality Inn & Suites Seattle Center',
 'MarQueen Hotel']
```

The following are recommended by Google for "The Bacon Mansion Bed and Breakfast":

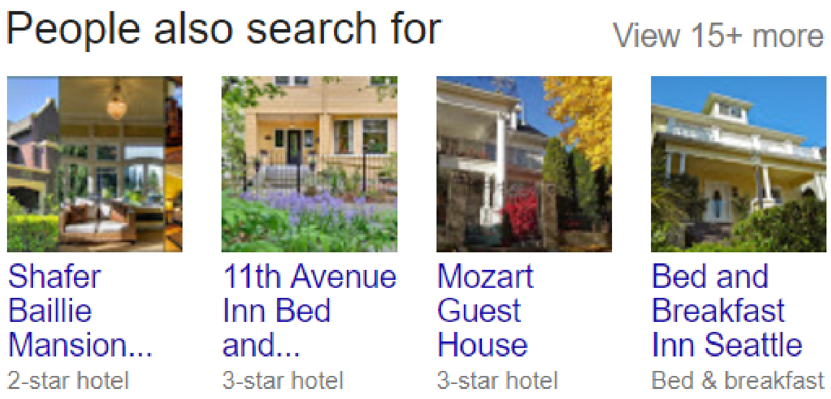
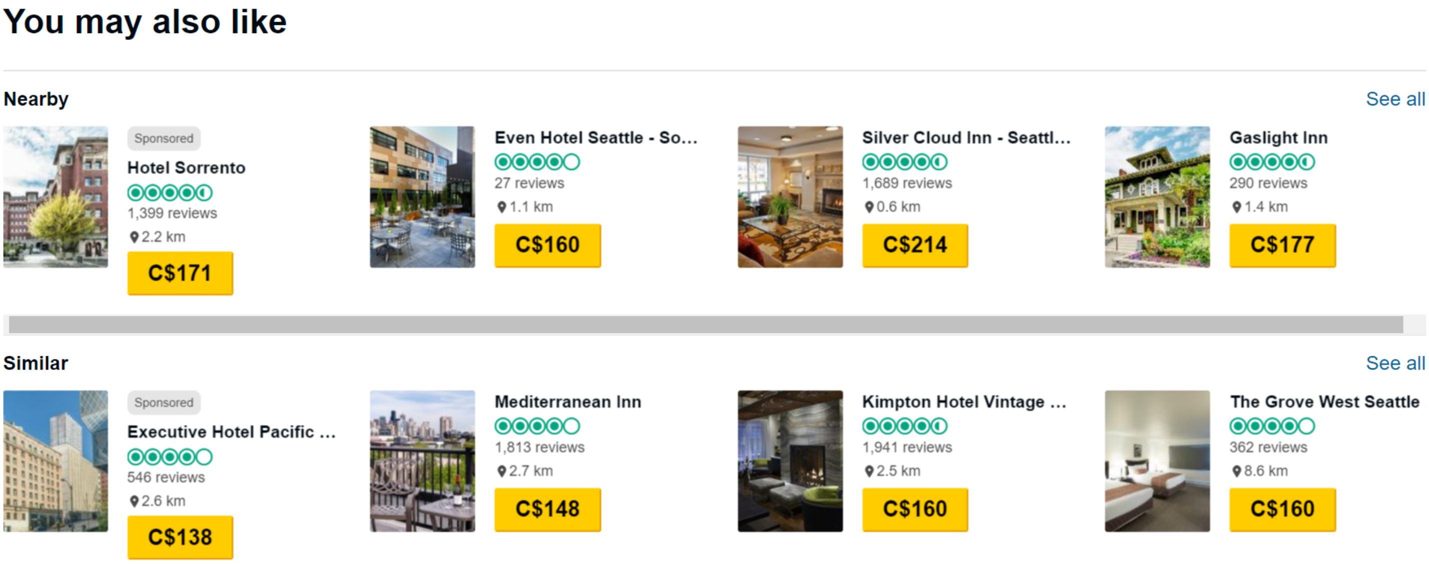


Figure 12

Cool!

The following are recommended by tripadvisor for "The Bacon Mansion Bed and Breakfast", which I was not impressed.



You may also like

Nearby

See all

Sponsored
Hotel Sorrento
●●●●◐
1,399 reviews
2.2 km
C$171

Even Hotel Seattle - So...
●●●●○
27 reviews
1.1 km
C$160

Silver Cloud Inn - Seattl...
●●●●◐
1,689 reviews
0.6 km
C$214

Gaslight Inn
●●●●◐
290 reviews
1.4 km
C$177

Similar

See all

Sponsored
Executive Hotel Pacific ...
●●●●○
546 reviews
2.6 km
C$138

Mediterranean Inn
●●●●○
1,813 reviews
2.7 km
C$148

Kimpton Hotel Vintage ...
●●●●◐
1,941 reviews
2.5 km
C$160

The Grove West Seattle
●●●●○
362 reviews
8.6 km
C$160

Figure 13

Jupyter notebook can be found on Github, if you prefer, this is a nbviewer version.

Have a productive week!