# HotelMatch

Use your own words to find your perfect hotel

By: Chris Johnson, GA Data Scientist

# Agenda

- Problem Statement
- Data Collection
- Data Cleaning
- Processing
- Final Product
- Conclusions
- Next Steps

# Problem Statement:

**Can hotel recommendations be made based off a descriptive input from a user regarding the nature of their trip?**

Expedia®

MC
Hotels.com MC

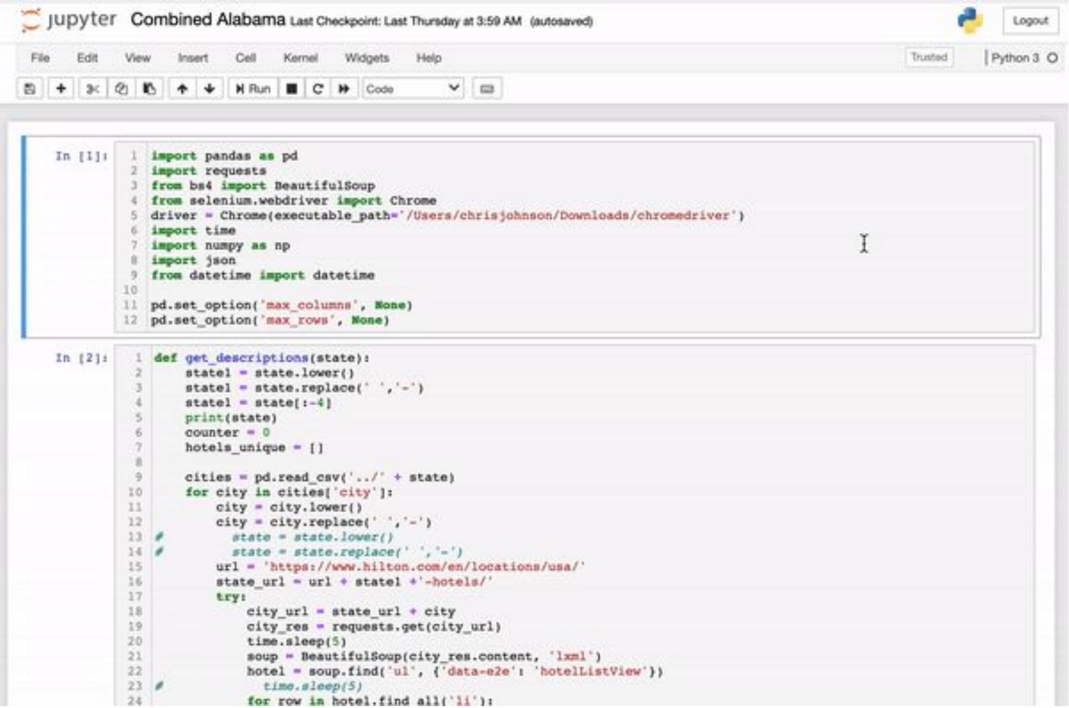
travelocity®

B.
Booking.com

# The Process

- No readily available tools or API’s
- Unique Code for each hotel brand was needed
- Utilized Selenium and Beautifulsoup for scraping
- Encountered issues with various brands which were resolved with VPN’s and wait times

# This 289 line code was to collect Hilton Hotels



In [1]:

```python
import pandas as pd
import requests
from bs4 import BeautifulSoup
from selenium.webdriver import Chrome
driver = Chrome(executable_path='/Users/chrisjohnson/Downloads/chromedriver')
import time
import numpy as np
import json
from datetime import datetime

pd.set_option('max_columns', None)
pd.set_option('max_rows', None)
```

In [2]:

```python
def get_descriptions(state):
    state1 = state.lower()
    state1 = state.replace(' ','-')
    state1 = state[:-4]
    print(state)
    counter = 0
    hotels_unique = []

    cities = pd.read_csv('../' + state)
    for city in cities['city']:
        city = city.lower()
        city = city.replace(' ','-')
#         state = state.lower()
#         state = state.replace(' ','-')
        url = 'https://www.hilton.com/en/locations/usa/'
        state_url = url + state1 +'-hotels/'
        try:
            city_url = state_url + city
            city_res = requests.get(city_url)
            time.sleep(5)
            soup = BeautifulSoup(city_res.content, 'lxml')
            hotel = soup.find('ul', {'data-e2e': 'hotelListView'})
#             time.sleep(5)
            for row in hotel.find_all('li'):
```

Alex, who happens to work for a certain hotel company:
'Hey, did you crash the website?'
Chris:
Down
DOWN
It is no
Chro
A

IT'S NOT YOU, IT'S ME.

We're sorry, something went wrong. For customer service or reservations, please call one of these numbers.

Find a hotel

# The Process

Over 21,000 hotels were used for the dataset, out of approximately 54,000 hotel properties in the country. The dataset consists of hotels from 6 of the largest companies.

Hilton

HYATT®

IHG®

MARRIOTT

Radisson® HOTELS & RESORTS

WYNDHAM HOTELS AND RESORTS

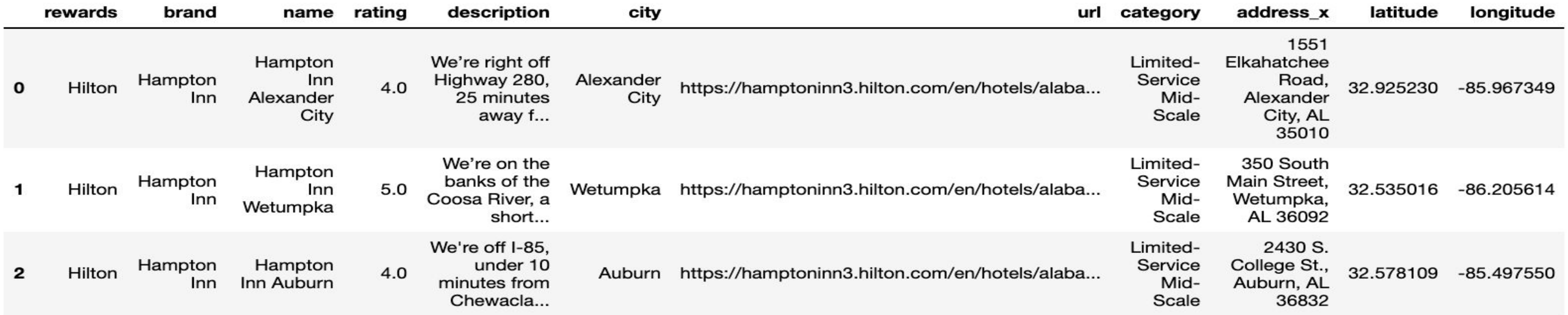
# The Process

Ultimately the following data were collected for each hotel:

- All text description from the page
- Physical Address
- Geo Coordinates (if available)
- Tripadvisor rating (if available)
- Hotel website url

# Data Cleaning

- Each brand was collected separately to make cleaning more manageable
- Cleaning the text of the description
- Creating a separate dataframe for missing info
- Filling in the missing info
- Creating New Feature Columns

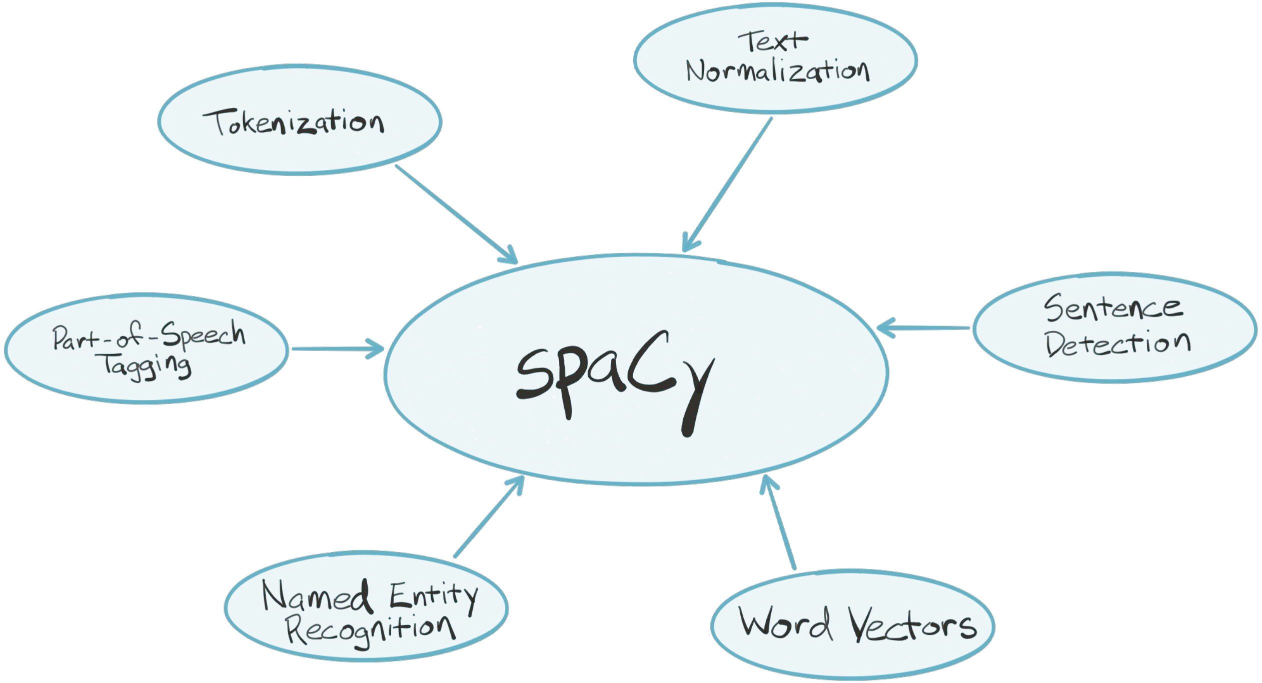


| | rewards | brand | name | rating | description | city | url | category | address_x | latitude | longitude |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Hilton | Hampton Inn | Hampton Inn Alexander City | 4.0 | We're right off Highway 280, 25 minutes away f... | Alexander City | https://hamptoninn3.hilton.com/en/hotels/alaba... | Limited-Service Mid-Scale | 1551 Elkahatchee Road, Alexander City, AL 35010 | 32.925230 | -85.967349 |
| **1** | Hilton | Hampton Inn | Hampton Inn Wetumpka | 5.0 | We're on the banks of the Coosa River, a short... | Wetumpka | https://hamptoninn3.hilton.com/en/hotels/alaba... | Limited-Service Mid-Scale | 350 South Main Street, Wetumpka, AL 36092 | 32.535016 | -86.205614 |
| **2** | Hilton | Hampton Inn | Hampton Inn Auburn | 4.0 | We're off I-85, under 10 minutes from Chewacla... | Auburn | https://hamptoninn3.hilton.com/en/hotels/alaba... | Limited-Service Mid-Scale | 2430 S. College St., Auburn, AL 36832 | 32.578109 | -85.497550 |

# Let’s Go to the Recommender

# Processing

- spaCy
- Creating vectors
- Why spaCy
- Features of spaCy



# The Recommender

To produce recommended hotels for each user the application takes in the following information:

- Description of trip
- Destination and Search Radius
- Types of Hotels
- Minimum Tripadvisor Rating
- Hotel Rewards Programs

# The Recommender

The recommender utilizes all options other than the description to filter down the dataset. Once filtered down the same NLP steps are performed on the user input that was performed on the dataset. Once vectors are created for the user input, cosine similarity is used to find the post similar description and the associated hotel. The recommender then returns the top 20 results in descending order, with an associated map.

# Conclusions

- There is no “accuracy score” for this, however without the inclusion of radius as a filter the results are not perfect. For example the first search for a hotel in Washington, DC on the entire dataset resulted in only 2 of the top 10 results being in Washington, DC area.
- Inherent limitations for sites which do not include a lot of text descriptions.
- Performing NLP and Vectorizing on the filtered data instead of all 21000 hotels
- Sentiment analysis was not useful

# Next Steps

- Collect data on amenities offered by hotels
- Collect data from other hotel chains
- Collect reviews from guests to use in sentiment analysis
- Incorporate use of other features from tokenized spaCy docs
- Deploy a version which can link directly to hotel websites
- See if I can take a user description and predict the type of hotel they are seeking, and see how it compares with the filters they choose

# Questions