Recommender System for Neighbourhoods

Leonardo Ramirez
March 2021

Peer-graded assignment for final module "Applied Data Science Capstone" part of Coursera "IBM Data Science Professional Certificate"

Recommender system for London Neighbourhoods

- Data
- Manipulation and Exploratory Analysis
- Machine Learning Methods
- Results

Data

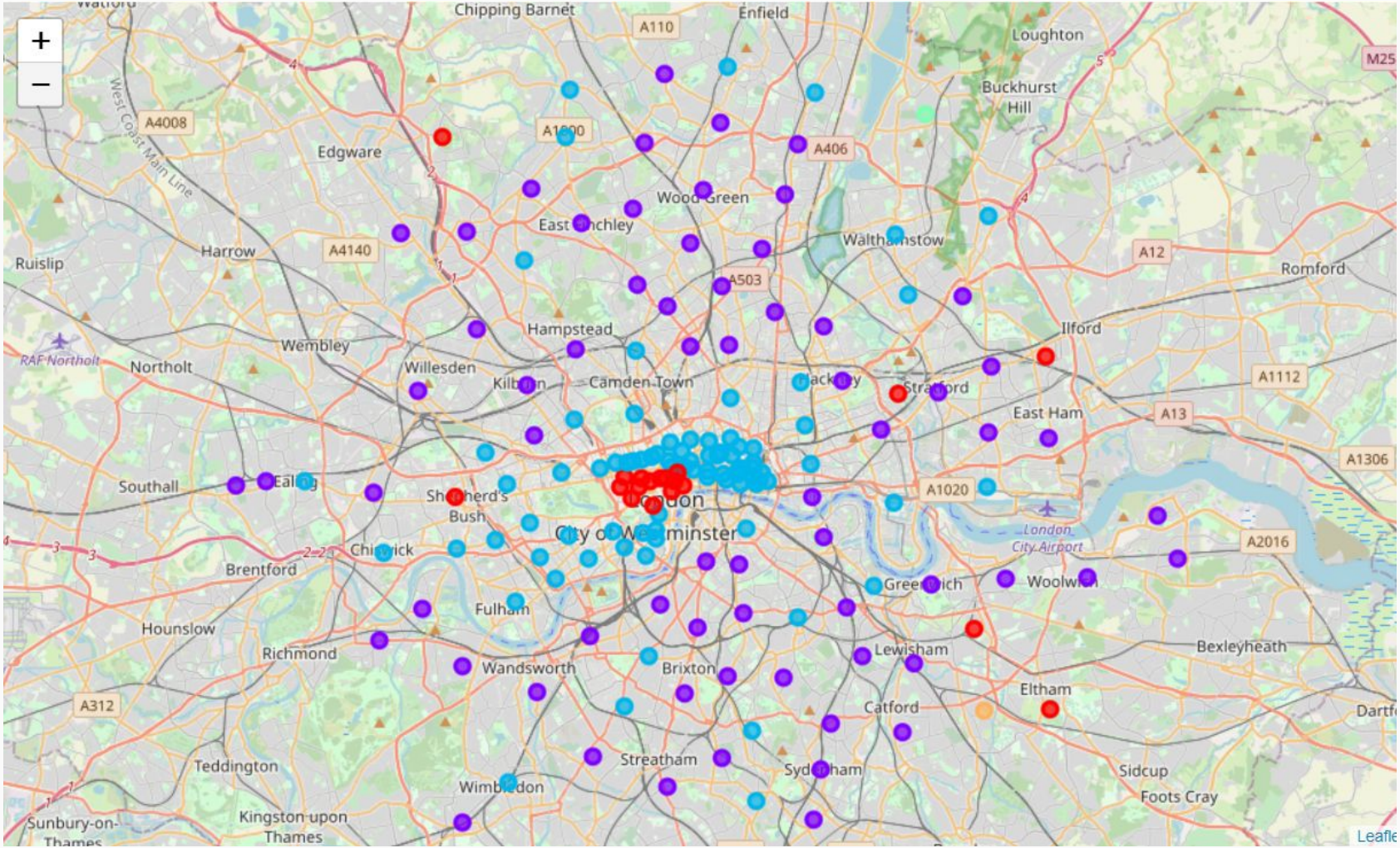
- Three sources of data:
  - Postcode data scraped from Wikipedia pages
  - Postcode coordinates from the Internet
  - Venues from FOURSQUARES API

Manipulation and Exploratory Analysis

- Data cleaning
  - To create London_df table
  - Add coordinates (by joining with coordinates table)
- Manipulation
  - Join with Venue data
- Exploration
  - Display data
  - K-means clustering

Machine Learning Methods

- K-means clustering
  - Used iterative method to find optimum k
  - Found 5 different clusters
- Content-based recommender
  - Based on user ratings
  - Matched neighbourhoods on preference for venues


Chipping Barnet
Enfield
Loughton
Buckhurst Hill
A110
M25
A4008
West Coast Main Line
Edgware
A1000
A406
Wood Green
East Finchley
Harrow
A4140
Ruislip
Walthamstow
A12
Romford
A503
Hampstead
Ilford
Wembley
RAF Northolt
Northolt
Willesden
Kilburn
Camden Town
Stratford
A1112
East Ham
A13
A1306
Southall
Ealing
Shepherd's Bush
A1020
London City Airport
A2016
Chiswick
Brentford
Fulham
Greenwich
Woolwich
Hounslow
Richmond
Wandsworth
Brixton
Lewisham
Bexleyheath
Eltham
A312
Catford
Teddington
Streatham
Sidcup
Foots Cray
Wimbledon
Kingston upon Thames
Sunbury-on-Thames
Leaflet

Results

- Seems to recommend similar neighbourhoods
- However recommendation limited by user input
  - So will not suggest anything different from inputs
- Could be improved by adding user characteristics
  - And making recommendations based on neighbourhoods liked by similar users
- Top 5 recommendations all belong to same cluster

neighbourhood_ratings

	Neighbourhood	rating
0	Hammersmith	3
1	Clapham	9
2	Earls Court	9
3	Brixton	1
4	Marylebone	10

```
userProfile.sort_values(ascending=False).to_frame().head(10)
```

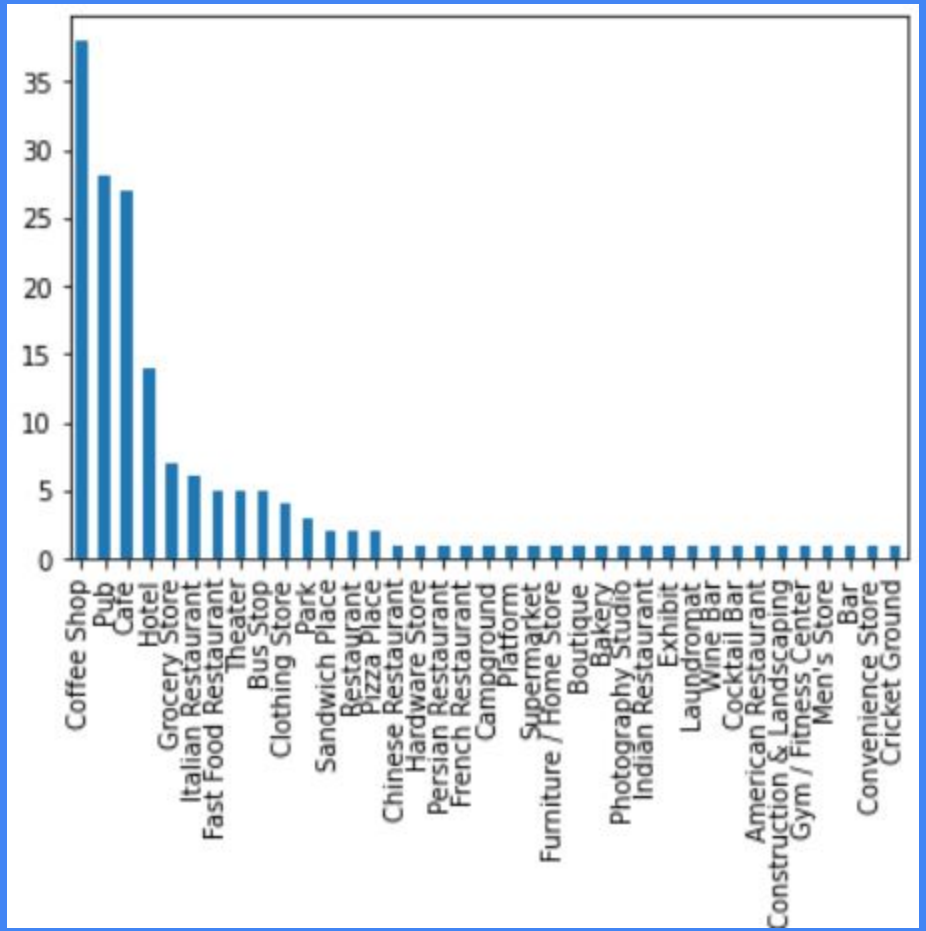
	0
Hotel	3.649130
Pub	1.908446
Café	1.417239
Coffee Shop	1.070655
Italian Restaurant	0.896722
Burger Joint	0.894313
Sandwich Place	0.871069
Cocktail Bar	0.740517
Grocery Store	0.715562
Indian Restaurant	0.680655

```
LDN_Neighbourhoods_venues_sorted.sort_values(by='Relative Rating',ascending= False)
```

	Cluster Labels	Relative Rating	Ranking	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
36	2	0.050112	1.0	Earls Court	Hotel	Pub	Garden	Café	Coffee Shop	Italian Restaurant
160	2	0.035441	2.0	West Kensington	Hotel	Pub	Sandwich Place	Pizza Place	Indian Restaurant	Grocery Store
108	2	0.032388	3.0	Paddington head	Hotel	Café	Coffee Shop	Pub	Garden	Italian Restaurant
152	2	0.031219	4.0	Victoria	Hotel	Pub	Indian Restaurant	Café	Italian Restaurant	Pizza Place
124	2	0.029621	5.0	South Bank	Pub	Coffee Shop	Park	Grocery Store	Hotel	Sandwich Place

More and better data = better recommendations

FOURSQUARES data seems to have a lot of venues aimed at tourists (schools are missing for example)



“The model is as good as the input data”

- Overhead on the bus

Doing this exercise has been a fun way to learn more about machine learning and to practise using Python

Thanks!

Fun Fact:

According to FOURSQUARES API Data…

London has more cafés/coffee shops than pubs!

The most common venues in London are:

	Venue Category
Coffee Shop	665
Pub	523
Café	497
Hotel	393
Italian Restaurant	365
Sandwich Place	232
Gym / Fitness Center	223
Grocery Store	218
Bakery	213
Restaurant	191
Cocktail Bar	183
Clothing Store	175
Pizza Place	175
Indian Restaurant	173
French Restaurant	171
Park	152