Projektna naloga

Analiza ocen evropskih restavracij na TripAdvisor-ju

Avtorji:

Aljaž Božič (63170062)
Anej Lekše (63170174)
Mihael Šinkec (63170277)
Severin Rudolf (63150251)

Opis projekta

Cilj našega projektnega dela je, da analaziramo podatke restavracij, ki so objavljene na TripAdvisor-ju in pridobiti nova znanja na podlagi teh. Poskušali se bomo osredotočiti na restavracije v Sloveniji.

Podatke smo dobili na spletni strani Kaggle.

Podatki so bili že v lepo strukturirani obliki, tako da je bilo predprocesiranja malo.

Iščemo odgovore na vprašanja, kot so sledeča:

- Distribucija vrste kuhinj po mestih - katera mesta imajo podobne tipe uspešnih restavracij?
- Ali obstaja korelacija med cenovnim razredom restavracije in oceno?
- Ali je vrsta kuhinj dober napovednik za oceno?
- Ali je možno na podlagi teh podatkov zgraditi model za napovedovanje uspešnosti restavracije?
- Katerih tipov kuhinj v določenem mestu "primanjkuje" glede na števila te kuhinje v drugih mestih?

Izvedba projekta

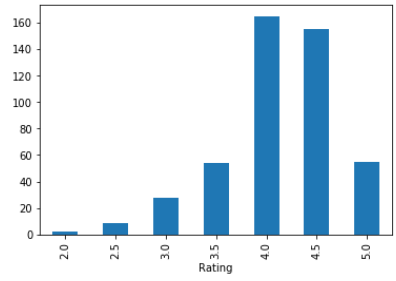
Faza 1

V začetni fazi smo želeli vizualizirati nekaj razmerj med podatki brez pretiranega predprocesiranja, da dobimo malo občutek, kako so ratingi restavracij in njihove druge lastnosti distribuirane.

Za začetek smo samo prebrali in izpisali dataset:

	Name	City	Cuisine Style	Ranking	Rating	Price Range	Number of Reviews	Reviews	URL_TA	ID_TA
0	Martine of Martine's Table	Amsterdam	['French', 'Dutch', 'European']	1.0	5.0	$	136.0	[['Just like home', 'A Warm Welcome to Wintry ...	/Restaurant_Review-g188590-d11752080-Reviews-M...	d11752080
1	De Silveren Spiegel	Amsterdam	['Dutch', 'European', 'Vegetarian Friendly', '...	2.0	4.5		812.0	[['Great food and staff', 'just perfect'], ['0...	/Restaurant_Review-g188590-d693419-Reviews-De_...	d693419
2	La Rive	Amsterdam	['Mediterranean', 'French', 'International', '...	3.0	4.5		567.0	[['Satisfaction', 'Delicious old school restau...	/Restaurant_Review-g188590-d696959-Reviews-La_...	d696959
3	Vinkeles	Amsterdam	['French', 'European', 'International', 'Conte...	4.0	5.0		564.0	[['True five star dinner', 'A superb evening o...	/Restaurant_Review-g188590-d1239229-Reviews-Vi...	d1239229
4	Librije's Zusje Amsterdam	Amsterdam	['Dutch', 'European', 'International', 'Vegeta...	5.0	4.5		316.0	[['Best meal.... EVER', 'super food experience...	/Restaurant_Review-g188590-d6864170-Reviews-Li...	d6864170

Nato smo se usmerili malo v Ljubljano in izrisali graf, ki prikazuje koliko restavracij v Ljublani ima kakšen rating.



Po the uvodnih korakih smo se domislili, da bi lahko naredili wordcloud najbolj popularnih besed v komentarjih. Za odstranitev neuporabnih besed (it, the,…) smo uporabili modul nltk.

Katere so najbolj popularne vrsta kuhinj v Ljubljani?

	Cuisine_style	count
1	European	247
2	Slovenian	174
3	Vegetarian Friendly	129
4	Central European	117
5	Mediterranean	87
6	Vegan Options	65
7	Italian	64
8	Pizza	62
9	Gluten Free Options	61
10	Bar	51
11	Cafe	41
12	Fast Food	37
13	Eastern European	35
14	Pub	34
15	Barbecue	33

Faza 2

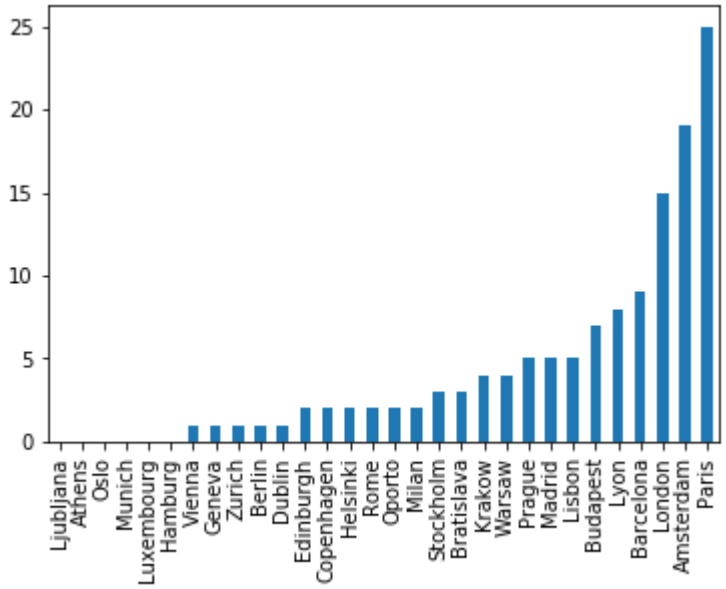
Od relativno preprostih stvari smo se odločili nadaljevati naše delo v smeri zadnjega in predzadnjega vprašanja, saj se je na podlagi naših (že tako informacijsko omejenih) podatkov to zdela najbolj optimalna opcija.

Cilj nam je bil karseda natančno napovedati uspešnost restavracije v določenem mestu.

Najprej smo ustvarili dataframe z distribucijo kuhinj za vsako mesto.

Cuisine_style	Amsterdam	Athens	Barcelona	Berlin	Bratislava	Brussels	Budapest	Copenhagen	Dublin	Edinburgh	...	Munich	Oporto	Oslo	
Caucasian	0.0	0.0	0.0	2.0	1.0	0.0	0.0	1.0	0.0	0.0	...	0.0	0.0	0.0	
Czech	0.0	3.0	0.0	0.0	34.0	0.0	4.0	0.0	2.0	0.0	...	0.0	0.0	0.0	
British	19.0	7.0	12.0	6.0	3.0	13.0	9.0	6.0	57.0	696.0	...	1.0	1.0	3.0	
Seafood	133.0	78.0	293.0	132.0	21.0	81.0	38.0	48.0	51.0	46.0	...	118.0	44.0	40.0	
Canadian	0.0	0.0	0.0	2.0	0.0	0.0	1.0	0.0	1.0	1.0	...	0.0	0.0	0.0	
Italian	405.0	124.0	619.0	871.0	83.0	493.0	283.0	226.0	188.0	163.0	...	441.0	66.0	104.0	1
Brew Pub	4.0	9.0	45.0	12.0	12.0	41.0	10.0	7.0	7.0	5.0	...	8.0	5.0	6.0	
Danish	2.0	0.0	0.0	0.0	0.0	0.0	1.0	508.0	0.0	0.0	...	0.0	0.0	5.0	
Irish	7.0	1.0	11.0	9.0	2.0	6.0	4.0	5.0	658.0	2.0	...	6.0	0.0	1.0	
Swedish	3.0	0.0	1.0	2.0	0.0	0.0	0.0	3.0	1.0	2.0	...	0.0	0.0	3.0	

Spodnji graf prikazuje, v katerem mestu (razen v Bruslju) je največ belgijskih restavracij:



Ko smo imeli podatke v tej obliki smo določili kriterij uspešnosti.

Uspešna restavracija mora imeti:

1. dober rating
2. veliko ocen
3. unikatno ponudbo

Formula ki to kolikor toliko dobro obsega je:

SR = (rating^2 * NumOfReviews) / št restavracij z enako ponudbo v istem mestu

Success rating smo dodali tudi v dataframe kot atribut

ID	Name	City	Cuisine Style	Rating	Price Range	Number of Reviews	NumOfSameType	SuccRating
0	Martine of Martine's Table	Amsterdam	['French', 'Dutch', 'European']	5.0	3	136.0	2605.0	1.305182
1	De Silveren Spiegel	Amsterdam	['Dutch', 'European', 'Vegetarian Friendly', '...	4.5	4	812.0	4243.0	3.875324
2	La Rive	Amsterdam	['Mediterranean', 'French', 'International', '...	4.5	4	567.0	4368.0	2.628606
3	Vinkeles	Amsterdam	['French', 'European', 'International', 'Conte...	5.0	4	564.0	4570.0	3.085339
4	Librije's Zusje Amsterdam	Amsterdam	['Dutch', 'European', 'International', 'Vegeta...	4.5	4	316.0	5155.0	1.241319

Nato smo razdelili dataset v testno in učno množico ter nad njimi pognali linearni regresor.

Na koncu smo ugotovili po računanju MSE, da je kolikor toliko uporabnih ocen le okrog 60%.

```
-----
Procent uporabnih ocen: 61.344900306748464
```

Zaključek

Na podlagi tega smo ugotovili, da je zelo težko napovedovati tako obsežno stvar s tako omejenim datasetom. Podatki, ki bi prišli prav, so recimo povprečna distribucija obiskovalcev po dnevih, itd.