Olivier Gruss

Organización de datos – TP2-Reentrega

Parte I-Spark

Se deberán realizar las consignas pedidas utilizando la librería Spark de Python sobre los datasets encontrados en la carpeta Dataset.

Ejercicio S9: ¿Cuál es el código postal del local cuyas reviews suman más votos obtenidos de 'cool', 'funny' y 'useful'?

Solución: https://colab.research.google.com/drive/1tMO-DMBahCglL-hHmCgZOigYk2Q4qCi4?authuser=1

Ejercicio S14: Nos vamos a quedar dos días en New Orleans por una meetup de data science. Queremos visitar la ciudad pero no tenemos mucho tiempo así que visitaremos sus mejores lugares. Vamos a calcular el score de review promedio para cada lugar, pero para tener en cuenta la varianza vamos a restarle a cada promedio su desviación estándar y solo usar lugares con más de 10 reviews.

- Nos han dicho que la ciudad tiene un barrio francés con muy buena gastronomía. ¿Cúal es el mejor lugar para comer con la categoría "French"?
- Después de comer queremos ir a un bar a tomar tragos, ¿cuál es el mejor de la categoría "Bars"?
- ¿Cuál es el mejor museo (categoría "Museums")? ¿De qué trata?

Solución:
https://colab.research.google.com/drive/1ZYSFxp5k2xH4AjpLpdVhW_s9JajsVbJY?authuser=1

Ejercicio S27: ¿Quién es el usuario más antiguo en el sitio que tiene exactamente 250 fans?

Solución: https://colab.research.google.com/drive/11Zi8kZKa8U7so7HxN4ULcB0YHI9xohb-?authuser=1

Ejercicio S35: Queremos saber dónde vive alguno de nuestros usuarios, para los usuarios que tiene más de 50 registros a negocios distintos en la tabla reviews obtenga el promedio y desviación estándar de la latitud y longitud de los negocios que calificaron (contando cada negocio una sola vez). Para el usuario que menos desviación estándar sumada tenga de ambas coordenadas muestre ese promedio y dónde está eso (https://www.gps-coordinates.net/) y cómo se llama el usuario para después irlo a buscar a la casa.

Olivier Gruss

Solución:
https://colab.research.google.com/drive/1fUU8kpkLM63f29N2RzH2xnvwABG42SZQ?authuser=1

Ejercicio S42: Queremos crear nuestro propio clasificador de reviews según sean positivas o negativas usando los datos de yelp, para hacer esto vamos a hacer una cosa muy sencilla: asignarle a cada palabra de las 500 más comunes sin contar stopwords el promedio de las stars para las reviews en las que aparece, luego, cuando aparezca un nuevo texto para las palabras que conozcamos del mismo promediamos sus scores. Por ejemplo, si tenemos las palabras “buena” con polaridad 3.4 y “rica” con polaridad 4.3 y tenemos el texto “buena y rica” su predicción será 3.85. Puede usar una muestra para entrenar. ¿Cuál es la salida del predictor para “I loved this place, the food was amazing!”?

Solución: https://colab.research.google.com/drive/1329hyGkiPV_EA6GfO9HgEah17C-svTL3?authuser=1#scrollTo=zYyzZM2IiiVN