TP1 - Pandas, Spark y Visualización de datos

Cátedra Argerich

Ir al inicio

TP1 - Pandas, Spark y Visualización de datos

Para realizar las distintas etapas de este trabajo práctico, utilizaremos los datos de las Apps en Google Play, disponibles en https://www.kaggle.com/datasets/lava18/google-play-store-apps

Dataset

GooglePlayStore.csv (10841 entries)

Details of the applications on Google Play. There are 13 features that describe a given app.

Id	Columna	Descripcion
0	App	Name of app
1	Category	Category the app belongs to
2	Rating	Overall user rating of the app
3	Reviews	Number of user reviews for the app
4	Size	Size of the app
5	Installs	Number of user downloads/installs for the app
6	Type	Paid or Free
7	Price	Price of the app
8	Content	Age group the app is targeted at - Children / Mature 21+ / Adult
9	Genres	An app can belong to multiple genres (apart from its main category)
10	Last	Updated Date when the app was last updated on Play Store
11	Current Ver	Current version of the app available on Play Store

Id	Columna	Descripcion
12	Android Ver	Min required Android version

GooglePlayStore_User_Reviews.csv (64295 entries)

This file contains most relevant reviews for each app. Each review text/comment has been pre-processed and attributed with 3 new features - Sentiment, Sentiment Polarity and Sentiment Subjectivity.

Id	Columna	Descripcion
0	App	Name of app
1	Translated_Review	User review (Preprocessed and translated to English)
2	Sentiment	Positive/Negative/Neutral (Preprocessed)
3	Sentiment_Polarity	Sentiment polarity score
4	Sentiment_Subjectivity	Sentiment subjectivity score

Primera parte - Visualización de Datos (**6ptos**)

Para esta primera parte, deberán realizar los siguientes plots:

1. Realizar **tres** visualizaciones que expliquen los datos del TP **utilizando los siguientes tipos de plots**:
   - Histograma (**1pto**)
   - Violin plot o Box plot (**1pto**)
   - Heatmap (**1pto**)
2. Realizar **un plot** sobre los datos, utilizando un tipo de plot que **no sea ninguno de los usados en el punto uno**, que permita mostrar el comportamiento o interacción de **al menos tres variables**. (**3ptos**)

Segunda parte - Pandas (**8 ptos**)

Realizar sus correspondientes consultas en Pandas (*Cada alumno tendrá asignado **2 ejercicios** de ⭐ estrella y tres de estrellas ⭐⭐, las asignaciones están disponibles en la pagina*)

1. Indicar cuántas aplicaciones tienen al menos un review, y listarlas (⭐)
2. Indicar cuántas aplicaciones no han recibido ningún review, y listarlas (⭐)
3. Indicar el nombre y el tamaño de las aplicaciones educativas. (⭐)

4. Calcular la probabilidad de que una App tenga vista positiva asignada. (⭐)
5. Calcular el porcentaje de App para adolescentes que no recibieron ninguna review. (⭐)
6. ¿Cuántas aplicaciones tienen un tamaño que varía según el dispositivo? Nombrar cual es la que tiene peor Rating. (⭐)
7. ¿Cuáles son las tres aplicaciones pagas con peor Rating? (⭐)
8. ¿Cuál es la aplicación con el nombre más largo? (⭐)
9. ¿Cuál es la aplicación que generó más dinero? (⭐)
10. Mostrar el largo promedio de las reviews por tipo de sentimiento. (⭐)
11. ¿Cuál es la aplicación con mayor promedio de score de sentimiento subjetivo? (⭐)
12. ¿Qué porcentaje de reviews contienen al nombre de la app en la review? (⭐)
13. Nombrar la aplicación que lleva más tiempo sin ser actualizada. (⭐)
14. Top 5 de categorías con mayor cantidad de reviews. (⭐)
15. Para aquellas aplicaciones que se encuentren en su versión 1.0 (o 1.0.0) y que tengan más de 50000 reviews. ¿Cuál es el top 5 con mejor rating? (⭐)
16. Para el rating promedio de cada categoría que no sea 1.9, ¿Cuáles son los 5 promedios más comunes? (⭐)
17. Para las aplicaciones que no sean para niños (Teen), ¿En qué posición se encuentra el género Communication? (⭐) Hint: está entre uno de los géneros con más aplicaciones.
18. De las apps que tienen en el nombre “FREE” ¿Cuál es la menos puntuada? (Rating) Si hay más de una, mostrar cualquiera (⭐)
19. Calcule la correlacion entre la antiguedad promedio (en dias) de una categoria con el promedio de la polaridad de sus reviews y con el promedio de la subjetividad de las mismas. (⭐⭐)
20. Listar, para cada categoría, cuáles son las tres aplicaciones con mejor Rating (⭐⭐)
21. Para cada categoría, indicar cuál es la aplicación que tiene mayor cantidad de reviews con sentimiento negativo (⭐⭐)
22. Indicar cuáles son las 10 aplicaciones que generaron opiniones más polarizadas (mayor cantidad de opiniones positivas o negativas, pero muy pocas neutras). Por ej, si la app A tiene 50 reviews, 25 positivos y 25 negativos, y la app B tiene 100 reviews, 25 positivos, 25 negativos y 50 neutros, y la app C tiene 10 reviews positivos, 10 negativos, y ninguno neutro, deberíamos listar primero la app A, luego la C y por último la B (⭐⭐)
23. Mostrar el promedio de ratings para las apps de cada categoría de tipo pagas o gratuitas teniendo en cuenta solo a las apps que tengan al menos una review de mas de 20 palabras o alguna que incluya ‘good’/’bad’. (⭐⭐)
24. Indica las 10 apps de categoría Sport con sentimiento positivo y mayor rating.(⭐⭐)
25. Calcular el promedio de rating por tipo de app (paid/free) para las aplicaciones que posean al menos una review donde el modulo de la polaridad (positiva o negativa) sea superior a dos veces la media del modulo de la polaridad de su tipo. **Aclaracion** *Una app con polaridad -0.5 tiene modulo 0.5, asi que si todas las apps pagas tienen 0.5 o -0.5 en polaridad, el promedio del modulo de la polaridad es 0.5 (no 0)* (⭐⭐)
26. Indicar el top 3 de generos con mayor cantidad de reviews. (⭐⭐)
27. Utilizando los textos de las reviews, realizar una consulta mediante técnicas de NLP de modo que la query “love game” devuelva una app, especificando su nombre, rating y la

review.(⭐⭐)

28. Calcule el tamaño promedio de las aplicaciones por versión de Android, sin tener en cuenta las aplicaciones que varían en tamaño según dispositivo. (⭐⭐)
29. Calcular a que cuantil de reviews pertenece cada app y ordenar de forma descendente todas las categorías segun las descargas de las apps separadas en cada cuantil. (⭐⭐)
30. ¿Cuál es la aplicación gratis con mayor ratio de reviews positivas? (⭐⭐)
31. Listar la cantidad de reviews neutras según el año de última actualización de la aplicación. Es decir, si una aplicación tiene 10 reviews neutras y fue actualizada por última vez en 2012, se cuentan 10 reviews para ese año. Además listar el nombre de las aplicaciones cuyo año de última actualización corresponde al año de menor cantidad de reviews neutras. (⭐⭐)
32. ¿Cuáles son las 5 palabras más utilizadas para las reviews, sin tener en cuenta las stopwords? (⭐⭐)
33. Para cada categoría, calcular el promedio de sentimiento de subjetividad de las reviews que recibieron sus aplicaciones. Ordenar las categorías según estos promedios. (⭐⭐)
34. Comprima los textos de las reviews que no sean nulas, obteniendo un ratio de compresión mejor que 2 y responda las siguientes preguntas: (⭐⭐)
    - ¿De cuánto es el ratio de compresión?
    - ¿Cuánto tarda en comprimir y descomprimir (por separado)? (use el magic %%timeit).
    - ¿Cuánto ocupa (en bytes) cada carácter en promedio una vez comprimido?
    - Si tomamos la entropía base dos para los caracteres ¿Cuánto da? ¿Cuántos bytes por carácter son esos?
    - Si utilizaramos un compresor aritmético por caracter, aproximadamente ¿Cuál sería el ratio de compresión en el caso más optimista?
    - ¿Cuál de los dos algoritmos de compresión sería mejor?
35. Correlación entre la polaridad y subjetividad promedio de los comentarios de los juegos cuya última actualización haya sido durante el 2018 y cuyo tamaño sea mayor al tamaño promedio de los juegos de ese año. Ignorar las aplicaciones cuyo tamaño varía con el dispositivo (Varies with device). (⭐⭐)
36. Devolver las categorías que tengan una app dominante de nivel K una app es dominante a nivel K si la cantidad de descargas es mayor al número de de descarga de las k siguientes apps ordenadas según el número de descargas. (⭐⭐)
37. Devolver las categorías que utilicen más caracteres distintos para sus apps, y las que menos caracteres distintos usan. ¿Cuáles son los caracteres que se comparten entre ambas? (⭐⭐)
38. Para cada categoría, calcular cuál es el caracter predominante en la misma. El caracter predominante en una categoria esta dado por el caracter que es dominante en mayor número de apps, por ej para aaaaaaax, xxb, xxs el caracter dominante de ap1 es a, pero para las otras 2 x, por lo que x es el caracter dominante en la categoría. Si no hay caracter dominante, es el primer caracter. (⭐⭐)
39. Calcule la correlación entre la cuota de mercado de una app y su sentimiento promedio. (⭐⭐)

40. Queremos saber cuánto pesaría si quisiéramos bajar todas las apps de un género, para todos los géneros. Para eso se pide: Calcular separado por géneros, cuanto pesarian todas las apps que tienen ese género (Tener en cuenta que si una app tiene acción y arte, su peso cuenta para ambos géneros) (⭐⭐)
41. Calcular el sentimiento promedio de cada categoría en base al promedio ponderado del sentimiento de sus géneros. Si por ej una categoría tiene género A y B, A tiene asociado 1 y B asociado -1, si A tiene el 75% y B el 25%, la categoría tiene un sentimiento promedio de 0.5 (⭐⭐)

Tercera parte: Spark (**8 ptos**)

Realizar sus correspondientes consultas en Spark (*Cada alumno tendrá asignado **2 ejercicios** de ⭐ estrella y tres de estrellas ⭐⭐, las asignaciones están disponibles en la pagina*)

1. ¿Cuál es la categoría con mayor cantidad de reviews promedio en sus aplicaciones? ¿Por qué? (⭐)
2. Teniendo en cuenta las reviews que reciben las aplicaciones, devolver una (al azar) de la aplicación que haya recibido la mayor cantidad de reviews positivas. (⭐)
3. ¿Cuál es el género más común de las aplicaciones? Indique la cantidad de aplicaciones con dicho género (⭐)
4. A Tomás le gustan los juegos de acción y nos pidió que le ofrezcamos el top 3 de juegos de acción, generando el ranking según la cantidad de reviews positivas que hayan recibido. Devolver algunas referencias (al azar) de cada uno de ellos, para ayudarlo a decidir qué juego bajarse.(⭐⭐)
5. Para las aplicaciones cuya última fecha de actualización haya sido en 2017 o en 2018, hayan tenido más de 10M de descargas y su rating sea de, por lo menos, 4.5, ¿Cuál de ellas es la aplicación que tuvo la mayor cantidad de reviews positivas? ¿Y las negativas? ¿Y las neutras? (⭐⭐)
6. Queremos saber cuánto pesaría si quisiéramos bajar todas las apps de un género, para todos los géneros. Para eso se pide: Calcular separado por géneros, cuanto pesarian todas las apps que tienen ese género (Tener en cuenta que si una app tiene acción y arte, su peso cuenta para ambos géneros) (⭐⭐)
7. De las apps que tienen el nombre “FREE” ¿Cuál es la menos puntuada? (Rating) Si hay más de una, mostrar cualquiera (⭐)
8. Devolver la aplicación más polarizante. tener en cuenta el promedio de polarización de las reviews de cada aplicación. (⭐)
9. Cual es la app más cara de cada categoría (⭐)
10. Cual es la categoría con mayor promedio de apps que hayan sido al menos 1 vez calificadas como positivas (⭐⭐)
11. Realizar un análisis de stopwords de las reviews. Dada la frecuencia de los tokens de las reviews, mostrar los 30 tokens más frecuentes y listar del total de tokens cuales son stopwords utilizando nltk. (⭐⭐)

12. Juan tiene un dispositivo un poco viejo con Android 3.9 y ya no lo puede actualizar. Quiere descargar apps del género de “Arte y Diseño” pero no sabe cuáles son válidas para su versión de Android. Ayúdalo diciéndole cuántas apps puede descargarse y mencionale 3 de ellas con su respectivo Rating. (⭐)
13. Obtenga la matriz de distancias en días entre fechas de actualización de las aplicaciones pagas. ¿Cuáles son las dos aplicaciones con mayor distancia? (⭐⭐)
14. ¿Cuál es la aplicación gratis con mayor ratio de reviews positivas? (⭐⭐)
15. Calcular el promedio de rating por tipo de App.(⭐⭐)
16. Ordenar de forma descendente todas las categorías por cantidad de descargas. (⭐⭐)
17. ¿Cuáles son las 5 palabras más utilizadas para las reviews, sin tener en cuenta las stopwords? (⭐⭐)
18. Calcule el tamaño promedio de las aplicaciones por versión de Android, sin tener en cuenta las aplicaciones que varían en tamaño según dispositivo. (⭐⭐)
19. Para cada categoría, indicar cuál es la aplicación que tiene mayor cantidad de reviews con sentimiento negativo (⭐⭐)
20. ¿Cuáles son las tres aplicaciones pagas con peor Rating? (⭐)
21. ¿Cuál es la aplicación con el nombre más largo? (⭐)
22. ¿Cuál es la aplicación que generó más dinero? (⭐)
23. Indicar cuántas aplicaciones tienen al menos un review, y listarlas (⭐)
24. Indicar cuántas aplicaciones no han recibido ningún review, y listarlas (⭐)
25. Indicar el nombre y el tamaño de las aplicaciones educativas. (⭐)
26. ¿Cuál es la aplicación con mayor promedio de score de sentimiento subjetivo? (⭐)
27. ¿Qué porcentaje de reviews contienen al nombre de la app en la review? (⭐)
28. Nombrar la aplicación que lleva más tiempo sin ser actualizada. (⭐)
29. Calcular la antigüedad promedio de las app de categoría “Family”. (⭐⭐)
30. Para cada categoría, cuáles son las tres aplicaciones con mejor Rating (⭐)
31. ¿Cuál es la app con mayor cantidad de instalaciones? (⭐)
32. Martín tiene un celular de gama baja, y tiene poca capacidad de almacenamiento. ¿Cuál es la app de menor tamaño que puede descargar desde google play? (⭐)

Puntos extra Kahoot

- Los **3 alumnos** que queden en el podio de la clase de Visu tienen un punto extra en el TP de Visu.
- Los **2 alumnos** que logren más podios en los Kahoot de Pandas y Spark (3 puntos por primer puesto, 2 puntos por segundo puesto, 1 punto por tercer puesto) suma 2 puntos extra en el TP de Pandas y Spark, y los siguientes 3 alumnos suman 1 punto extra.

Criterio de aprobación

- La visualización se aprueba con **4 puntos**.
- Pandas y Spark con **10 puntos**.

Criterio de reentrega

- El TP de visualización se puede reentregar sólo habiendo realizado todas las visualizaciones, sin mínimo de puntaje.
- El TP de Pandas y Spark se necesita al menos 6 puntos en total y están todos los puntos para poder reentregar.
- La reentrega consiste en hacer un punto extra y corregir todos los puntos donde tuvieran menos de la mitad de los puntos.
- Se aprueba la reentrega si todos los puntos tienen al menos la mitad de los puntos. En caso de aprobar la instancia de reentrega, la nota es siempre 4.

Primera parte - Visualización de datos

- Cada visualización vale un punto, y debe cumplir con las siguientes condiciones:
- Debe explicarse por sí misma, sin necesidad de texto aclaratorio.
- Debe tener rótulos en los ejes que corresponda y en el título.
- Debe mostrar una relación con el target que sea clara.
- El uso del color debe ser intencional, elegido por ustedes, no por la librería.
- La visualización debe ser legible (Un bar chart de 40 barras por ejemplo es ilegible)
- Debe mostrar adecuadamente el uso de algún plot (distinto a los usados en el punto 1), y cumplir aparte las condiciones enunciadas en el punto anterior.

*Ante cualquier duda, pueden consultar en **#consultas-tp1-visu**.*

Segunda parte - Pandas

- Todos los ejercicios valen lo mismo que las estrellitas que tienen asignadas, a cada uno le corresponde hacer según indiquemos cual les toca:
  - 2 ejercicio de ⭐
  - 3 ejercicios de ⭐⭐
- Cada ejercicio se considera 100% correcto si:
  - Resuelve lo pedido (¡cuidado con casos bordes! ¡revisen todo lo que pueda ser NULL!): Si el ejercicio no resuelve al 100% lo pedido, se considera que vale como máximo la mitad
  - Lo hace de la forma más eficiente posible: Si el ejercicio no está resuelto de la forma más óptima, se considera que vale la mitad

La idea es que no lo hagan solos! Las consignas son complejas de entender en una sola lectura y necesitan pensarse lento, por esto es que es crucial consultar. Para esto hacemos lo siguiente según el tipo de duda:

Dudas de consigna:

- Van a poder consultar en el canal de slack #consultas-tp1-pandas, es MUY importante que antes de consultar vean si su duda no fue resuelta.
- En caso de no haber sido resuelta tienen que publicarla siguiendo el formato: " - La pregunta...". De esta forma todos podemos buscar fácil si ya se resolvió la duda o sumarnos a la discusión. No se debe incluir código de la resolución, ni en la pregunta ni interactuando con otros compañeros.

Dudas para saber si se puede usar alguna librería:

- Se hacen en el mismo formato que las dudas de consigna.

Dudas de código y optimización:

- Si son dudas generales de "*cómo se hace algo en pandas*" se puede consultar en las clases de consulta o en el canal #otras-consultas
- El resto de las dudas se deben consultar con algún ayudante por privado.

Tercera parte: Spark

- Todos los ejercicios deben realizarse utilizando el API de RDD de Spark.
- A cada uno le corresponde hacer según indiquemos cual les toca:
  - 2 ejercicio de ⭐
  - 3 ejercicios de ⭐⭐
- Cada ejercicio se considera 100% correcto si:
  - Resuelve lo pedido (¡cuidado con casos bordes!): Si el ejercicio no resuelve al 100% lo pedido, se considera que vale como máximo la mitad
  - Lo hace de la forma más eficiente posible: Si el ejercicio no está resuelto de la forma más óptima, se considera que vale la mitad. *En este aspecto considerar el buen uso del procesamiento distribuido de spark y potenciales errores que pueda realizar procesando información en el driver*

La idea es que no lo hagan solos! Las consignas son complejas de entender en una sola lectura y necesitan pensarse lento, por esto es que es crucial consultar. Para esto hacemos lo siguiente según el tipo de duda:

Dudas de consigna:

- Van a poder consultar en el canal de slack #consultas-tp1-spark, es MUY importante que antes de consultar vean si su duda no fue resuelta.
- En caso de no haber sido resuelta tienen que publicarla siguiendo el formato: " - La pregunta...". De esta forma todos podemos buscar fácil si ya se resolvió la duda o sumarnos a la discusión. NO SE DEBE incluir código de la resolución, ni en la pregunta ni interactuando con otros compañeros.

Dudas para saber si se puede usar alguna librería:

- Se hacen en el mismo formato que las dudas de consigna.

Dudas de código y optimización:

- Si son dudas generales de “cómo se hace algo en spark” se puede consultar en las clases de consulta o en el canal #otras-consultas
- El resto de las dudas se deben consultar por privado

Todos los ejercicios asignados deben estar resueltos en la entrega

¡También valoramos que se ayuden entre ustedes, debatan y compartan ideas en el canal slack!

Formato de la entrega

La entrega debe subirse a la plataforma Gradescope. Para hacerlo, deben generar un usuario en gradescope.com y buscar la asignación correspondiente al TP1. En youtube pueden encontrar un video mostrando cómo ingresar por primera vez a gradescope (solo deben utilizar el código de este cuatrimestre: Entry Code: N8RG22, el resto es igual). A la plataforma deben subir un único PDF con un link a el/los notebooks con la resolución de cada uno de los puntos de Pandas o Spark (por favor no incluir código en el pdf) y las visualizaciones pedidas (las visualizaciones si deben incluirlas en el documento, para la visu original no es necesario incluir código, solo la imagen)
