TP1 - Pandas, Spark y Visualización de datos

Cátedra Argerich

🏠 Ir al inicio

TP1 - Pandas, Spark y Visualización de datos

Primera parte - Pandas (8 ptos)

Utilizamos el dump Wikipedia Español al día 01/09 de 178gb, extrayendo los siguientes csvs:

contents.csv

Tabla con datos de todos los contenidos de Wikipedia en su versión más reciente.

Campo	Descripcion
title	Título del contenido
id	Identificador único del contenido
namespace	🤷
revision_id	Id de la última revisión realizada
parent_revision	Id de la revisión anterior a la actual
revision_timestamp	Timestamp de la última revisión
revisor_username	Username del autor de la última revisión
revisor_id	Id del revisor_username
revisor_ip	IP del revisor (en caso de que no estuviera registrado)
revisor_comment	Comentario de la revisión

contents_text_sample.csv

Tiene una muestra aleatoria del 5% de los contenidos de texto de wikipedia

Campo	Descripcion
id	Id del contenido
title	Título del contenido
text	Texto

geo_tags.csv

Campo	Descripcion
gt_id	Id del geo tag
gt_page_id	Id del contenido al que corresponde
gt_globe	En qué globo se encuentra
gt_primary	🤷‍♀️
gt_lat	Latitud
gt_lon	Longitud
gt_dim	🤷‍♀️
gt_type	Tipo de locación
gt_name	Nombre
gt_country	País
gt_region	Región

logs.csv

Todo el log de acciones realizadas.

Campo	Descripcion
item_id	ID del item afectado
timestamp	Timestamp del log
contributor_username	Username que realizó la acción

Campo	Descripcion
contributor_id	ID del user que realizó la acción
contributor_ip	IP (en caso de que no tuviera usuario)
comment	Comentario
logtype	Tipo de log
action	Acción realizada
title	Título del log

languages.csv

Contiene información sobre qué idiomas habla cada usuario

Campo	Descripcion
babel_user	User id
babel_lang	Código de idioma (ISO 639-2)
babel_level	Nivel en el lenguaje

redirect_list.csv

Algunos de los contenidos de Wikipedia son redirecciones a otros contenidos, esta tabla contiene esa información.

Campo	Descripcion
rd_from	ID del contenido que redirige
rd_namespace	🤷
rd_title	Título del contenido al que redirige
rd_interwiki	🤷
rd_fragment	🤷

categorylinks.csv

Campo	Descripcion
cl_from	ID del contenido
cl_to	Categoría a la que pertenece el contenido
cl_sortkey	🤷
cl_timestamp	Timestamp de la asociación de la categoría
cl_sortkey_prefix	🤷
cl_collation	🤷
cl_type	El tipo de contenido que se asignó a esa categoría

pagelinks_sample.csv

Tabla con links que van de una página interna a otra. Es una muestra de dos tercios.

Campo	Descripcion
pl_from	ID del contenido donde está el link
pl_namespace	🤷
pl_title	Título del contenido al cual va el link
pl_from_namespace	🤷

Realizar sus correspondientes consultas en Pandas

1. Para el usuario que más versiones actuales de contenido de wikipedia editó, calcule la fecha promedio, mínima y máxima en que lo hizo (⭐)
2. Qué porcentaje de las versiones actuales son páginas que se editaron una sola vez (⭐)
3. Cual es el porcentaje de títulos de contenidos de wikipedia cuya longitud es menor a 20 (⭐)
4. La probabilidad de que la versión actual de un contenido fuera editada sin dejar comentario para usuarios que están logueados y que no están logueados (⭐)
5. La palabra más común entre los títulos que no sea una stopword del inglés ni español (⭐⭐)
6. El porcentaje de contenidos que están publicados cuya última edición no tiene comentario para los usuarios que realizaron 1, >10 y >100 de las últimas ediciones (⭐⭐)

7. La antigüedad promedio de la última edición de los artículos cuyo título contenga tu apellido (si no hay, tu nombre y si tampoco hay usa Cafferata) (⭐⭐)
8. La mediana de la antigüedad para las últimas ediciones vigentes agrupado por el primer carácter del título (⭐⭐)
9. Cuales son los contenidos de wikipedia cuyo título empieza o termina con un emoji (⭐⭐)
10. Para los contenidos visibles en wikipedia, cuales son los artículos que tienen la máxima y mínima distancia entre ids de su revisión actual y la anterior (⭐⭐)
11. Para todos los comentarios de revisión de contenido que tengan más de 20 ocurrencias realice una matriz cuyas columnas sean esos comentarios y de índice los usuarios/ips con valores: True si ese usuario realizó ese comentario, sino False (⭐⭐)
12. Cuantos comentarios de revisión de artículos usan la palabra “mejor” (sin incluir sus variaciones) (⭐⭐)
13. Realice una consulta en los contenidos actuales que le permita identificar algún artículo que este vandalizado utilizando los datos de la revisión (⭐⭐)
14. Qué porcentaje de contenido geolocalizado de wikipedia NO está en la tierra (⭐)
15. Obtenga la matriz de distancias euclídeas para todos los contenidos que están en Marte. ¿Cuáles son los dos contenidos que están a menor distancia? (⭐⭐)
16. Calcule la probabilidad de las palabras para los textos, luego encuentre el documento que más se desvie de esas probabilidades utilizando la divergencia de Kullback-Leibler (⭐⭐)
17. Utilice los textos del contenido para realizar consultas por texto utilizando las técnicas vistas en la clase de NLP (BOW o TF-IDF) de modo que la query “retablo iglesia” devuelva alguna página acerca del retablo de alguna iglesia (⭐⭐)
18. Divida la tierra en bloques de latitud y longitud de 5x5, ¿Cuál es el bloque con menos (o ninguna) referencias? (⭐⭐)
19. Calcule la latitud y longitud promedio de los contenidos con referencias en la tierra y diga dónde está eso (⭐)
20. ¿Cuál es el segundo contenido con más referencias geográficas asignadas? (⭐⭐)
21. ¿Dónde está la referencia geográfica más repetida en la tierra de toda la Wikipedia Español? (⭐)
22. Elija su lugar favorito en el mundo y tome su latitud y longitud, ¿cuál es el título de la página de wikipedia más cercana? (⭐⭐)
23. ¿Qué porcentaje de los contenidos contienen a su mismo título en el texto? (⭐⭐)
24. Calcule el porcentaje de nulos para todas las columnas de geo_tags.csv (⭐)
25. ¿Quién es el usuario que más ha bloqueado a otros? (⭐)
26. ¿Cuál es el usuario o IP más bloqueado? (⭐)
27. ¿Cuál es el mínimo que ha durado desde su registro un usuario bloqueado en la plataforma? (⭐⭐)
28. ¿Cuál es la antigüedad promedio para cada usuario según su última actividad? (⭐⭐)
29. Utilice los logs para crear una matriz cuyas columnas sean los logtypes, los índices los actions y las celdas la cantidad de la intersección de ambas (⭐)
30. La 3-upla de palabras más común en los comentarios de los logs (⭐⭐)
31. El día con más y menos actividad que tuvo el sitio (⭐)
32. El usuario que más agradece y el que más agradecimientos tiene (⭐)

33. La primera discusión creada (⭐)
34. ¿Cuántos usuarios son nativos en un idioma que no sea español? (⭐)
35. Para los usuarios nativos (o superior) en español obtenga una serie cuyo índice sea cada uno de los otros idiomas que sabe y valor sea el nivel promedio (tomando N=4.5) (⭐⭐)
36. Quien es el usuario que más idiomas domina con un nivel de 2 o superior (⭐)
37. Obtenga un dataframe que tenga como índice al user_id, como columnas a los idiomas y el nivel de cada usuario para cada idioma como valor con -1 en caso de no tenerlo cargado. (⭐⭐)
38. Obtenga la matriz de correlación para saber idiomas distintos considerando que un usuario sabe un idioma si indicó un nivel de 1 o superior (⭐⭐)
39. ¿Cuál es el contenido al que más se hacen redirecciones? (⭐)
40. Para los contenidos geolocalizados: ¿Cúal es el contenido más cercano del que fue editado más recientemente? ¿Y la diferencia entre sus tiempos de edición? (⭐⭐⭐)
41. Para los contenidos geolocalizados, según la última versión de cada contenido: ¿Cuál es la latitud y longitud promedio del contenido editado según qué idioma sabe el editor? (⭐⭐⭐)
42. Si la experiencia de un usuario es la cantidad de logs en los que participó, ¿cuál es la tasa de contenidos cuya última revisión no tiene comentario en función de la experiencia de su revisor? (⭐⭐⭐)
43. ¿Cuántos usuarios o ips han sido bloqueados al menos una vez y la vez son los revisores de una última versión de un contenido? Calcule la diferencia entre la primera fecha de bloqueo y el promedio de las fechas de revisión correspondientes para cada usuario. (⭐⭐⭐)
44. Si decimos que la ubicación de un usuario es el promedio de la latitud y longitud de los contenidos geolocalizados para los cuales editó la última versión (ignorar usuarios que no editaron contenido geolocalizado). ¿Cuáles son los dos usuarios más cercanos? (⭐⭐⭐)
45. ¿A qué contenido se asignó por primera vez una categoría? (⭐)
46. Si decimos que la ubicación de una categoría es el promedio de la latitud y longitud de sus contenidos geolocalizados que son miembros de ella (si es que tiene): ¿Cuales son las dos categorías más cercanas? (⭐⭐⭐)
47. La mediana de cantidad de links internos que tienen todos los contenidos que existen (⭐⭐)
48. Si decimos que la ubicación de una página linkeada por otra es el promedio de la latitud y longitud de los contenidos geolocalizados que la referencian: ¿Cuales son las dos páginas que están más cerca? (⭐⭐⭐)
49. Si decimos que un usuario sabe un idioma cuando tiene un nivel de babel mayor o igual a 1, para aquellos que editaron una de las versiones actuales del contenido, ¿Cuál es la tasa de revisiones sin comentario que realizan en función de los idiomas que saben? (⭐⭐⭐)
50. Si decimos que un usuario sabe un idioma cuando tiene un nivel de babel mayor o igual a 1 consiga un dataframe cuyas columnas son tipos de logs, el índice es la cantidad de idiomas que sabe un usuario y las celdas la probabilidad de que esos usuarios generen ese tipo de log. (⭐⭐⭐)

51. Si la experiencia de un usuario es la cantidad de logs en los que participó, queremos saber que tanto nos sirve para predecir el futuro vandalismo: ¿Cuál es la probabilidad de que un usuario sea bloqueado según experiencias: <10, 10-40, 40-100, >100? Tener en cuenta que esta experiencia debe ser PREVIA al bloqueo del usuario. (⭐⭐⭐)
52. Si decimos que un usuario sabe un idioma cuando tiene un nivel de babel mayor o igual a 1, para cada grupo de usuarios que sabe una determinada cantidad de idiomas, ¿Cuántos de esos usuarios fueron bloqueados al menos una vez? (⭐⭐⭐)
53. Si para un usuario tenemos la cantidad de acciones que realizó para cada tipo de log y la cantidad de veces que fue bloqueado: ¿Cuál es la acción que más y menos correlaciona con ser bloqueado? ¿Qué acción correlaciona más con saber algo (babel>=0) de inglés? (⭐⭐⭐)
54. ¿Cuál es la acción más realizada por usuarios que no están registrados? (⭐⭐)
55. La cantidad promedio de modificaciones históricas que tuvieron los ítems cuya última versión fue editada por un usuario registrado o no registrado. (⭐⭐⭐)
56. Calcule la cantidad de acciones realizadas por usuarios según día de la semana (⭐)
57. Calcule la probabilidad de que una acción en general se realice según día de la semana. Calcule también para los días de la semana la probabilidad de que la última edición de un contenido sea realizada ese día. Calcule la entropía de ambas y la divergencia de Kullback Leibler entre ellas. (⭐⭐⭐)
58. Observe una muestra aleatoria de los comentarios de las acciones realizadas por usuarios o ips antes de ser bloqueados. Observe otra muestra de comentarios de acciones de todos. (⭐⭐)
59. ¿Cuál es el idioma para el cual sus usuarios realizan más agradecimientos en promedio? ¿Y el de menos agradecimientos? Calcule lo mismo para quienes reciben agradecimientos. (⭐⭐⭐)
60. Si decimos que un usuario sabe un idioma cuando tiene un nivel de babel mayor o igual a 1, para aquellos que editaron una de las versiones actuales del contenido, ¿Cuál es la cantidad de agradecimientos promedio que reciben en función de los idiomas que saben? (⭐⭐⭐)

Segunda parte - Visualización de datos (7 ptos)

61. (3 ptos) Elegir uno de los siguientes datasets:
    - Proyectando el comportamiento de la soja
    - ¿Llevo paraguas? Pronosticando la lluvia
    - Predicción de éxitos en oportunidades comerciales
    - Clasificación de preguntas de clientes
    - MELI Data Challenge 2021
    - Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines
    - DengAI: Predicting Disease Spread

Realizar tres visualizaciones que expliquen la variable a predecir conteniendo los siguientes tipos de plots:

- Histograma

- Violin plot o Box plot
- Heatmap

1. (4 ptos) Utilice alguna herramienta para realizar diagramas (por ejemplo Google Draw, draw.io, Google Slides, HTML, Illustrator, Photoshop, etc.) para crear una visualización **ORIGINAL** que no pueda realizarse de forma directa con las librerías más comunes de Python, puede utilizar las librerías de Python como paso intermedio. Puede realizar este punto sobre los datos de: cualquier dataset, estadística oficial, paper, estadística no oficial, encuesta, números sin ninguna fuente en un blog, etc. El objetivo es elegir un tema de su interés y comunicarlo de forma efectiva y agradable.

Tercera parte: Spark (8 ptos)

Realizar sus correspondientes consultas en Spark

1. Considerando los logs de acciones realizadas sobre ítems, mostrar el top 10 de titulos de ítems que fueron afectados por mayor cantidad usuarios distintos (⭐)
2. Considerando los logs de acciones realizadas sobre ítems, mostrar el top 10 de títulos de ítems que fueron afectados por mayor cantidad de usuarios no registrados (⭐)
3. Realizar un análisis de stopwords del contenido de texto de la Wikipedia. En este punto esperamos que analicen, dada la frecuencia de los términos que hay en la wikipedia cuales deberian ser considerados stop words. (⭐)
4. Considerando el pagelink_sample.csv representar como un grafo en Spark los contenidos de wikipedia (considerando los contenidos como nodos y los links como aristas) como una lista de aristas y mostrar un recorrido en la estructura. (⭐⭐⭐)
5. Considerando el pagelink_sample.csv representar como un grafo en Spark los contenidos de wikipedia (considerando los contenidos como nodos y los links como aristas) como una lista de adyacencia y mostrar un recorrido en la estructura. (⭐⭐⭐)
6. Considerando el pagelink_sample.csv, usando una representación de grafos realizar una función genérica que nos permita calcular los contenidos que se encuentran a un grado de separación (el siguiente del vecino) de cualquier identificador de contenido de la wikipedia. Mostrar el funcionamiento de la implementación con algún contenido incluido en el set de datos (⭐⭐⭐)
7. Considerando el pagelink_sample.csv, usando una representación de grafos realizar una función genérica que nos permita calcular la centralidad de un contenido cualquiera de la wikipedia mediante random walks. Mostrar el funcionamiento de la implementación con algún contenido incluido en el set de datos (⭐⭐⭐)
8. Considerando el pagelink_sample.csv, usando una representación de grafos obtener aquellos contenidos que tienen “relaciones no correspondidas”. Entendemos como funciona una relación correspondida con un ejemplo: Si el contenido A tiene un link al B, pero B no tiene un link a A, podemos decir que B tiene una relación no correspondida con A. (⭐⭐⭐)}

9. Mostrar de forma eficiente el tercer trigrama que tiene mayor frecuencia en los títulos de los contenidos de la wikipedia (⭐⭐)
10. Generar un RDD en el que cada tupla tenga el formato (key, value) donde:
    1. key sea una palabra del léxico de la wikipedia
    2. value sea una lista donde cada elemento de la misma sea una tupla de dos elementos
        1. identificador de contenido donde aparezca esa palabra.
        2. la frecuencia con la que aparece esa palabra en ese contenido. (⭐)
11. Generar una función genérica que dado un n nos permita obtener un RDD con los n-gramas del contenido de texto de wikipedia y su frecuencia (⭐)
12. Obtenga la matriz de distancias euclídeas para todos los contenidos que están en Marte. ¿Cuáles son los dos contenidos que están a menor distancia? (⭐⭐)
13. La región por cada país que tiene la mayor cantidad de contenidos publicados. (⭐)
14. El Top 5 de contenidos que tienen la mayor cantidad de redirecciones que apuntan a ellos. (⭐)
15. Listado en orden de importancia (del más hablado al menos hablado) de los idiomas que manejan aquellos usuarios que hablan por lo menos tres idiomas. (⭐⭐)
16. 10 categorías que tienen la menor cantidad de contenido anónimo publicado. (⭐)
17. Para aquel contenido georeferenciado publicado anónimamente indicar por país, cuántas IPs de usuarios corresponden a IPv4 y cuantas a IPv6. (⭐)
18. Para cada lenguaje indicar cuántos usuarios lo comprenden, cuantos lo manejan a nivel lectura y escritura base, cuantos hacen de él, un uso avanzado. (Para resolver deberá mapear los niveles de babel a esas categorías propuestas y darles un nombre). (⭐)
19. Cantidad de contenido por planeta fuera de la tierra en la Wikipedia. (⭐)
20. Cantidad de Stubs por categoría en la Wikipedia. (⭐⭐).
21. El contenido con mayor cantidad de de acciones realizadas para todos los tipos posibles de acciones (⭐⭐⭐).
22. Top 5 de lenguajes que son usados por usuarios bilingües. (⭐⭐).
23. Cantidad total de contenidos por tipo de locación que pertenecen a la tierra. (⭐)
24. Dado un tamaño de vocabulario parametrizable y una lista de stopwords también parametrizable implemente tf-IDF para los textos de los contenidos de forma distribuida. Debe obtener un vector por cada texto (⭐⭐⭐).
25. Obtenga con spark los datos (de forma ya agregada) que le permitan realizar la siguiente visualización y realice la misma (⭐⭐⭐):

26. Qué porcentaje de las versiones actuales son páginas que se editaron una sola vez (⭐)
27. La probabilidad de que la versión actual de un contenido fuera editada sin dejar comentario para usuarios que están logueados y que no están logueados (⭐)
28. El porcentaje de contenidos que están publicados cuya última edición no tiene comentario para los usuarios que realizaron 1, >10 y >100 de las últimas ediciones (⭐⭐)
29. Para los contenidos visibles en wikipedia, cuales son los artículos que tienen la máxima y mínima distancia entre ids de su revisión actual y la anterior (⭐⭐)
30. Qué porcentaje de contenido geolocalizado de wikipedia NO está en la tierra (⭐)

31. Calcule la latitud y longitud promedio de los contenidos con referencias en la tierra y diga dónde está eso (⭐⭐)
32. ¿Cuál es el segundo contenido con más referencias geográficas asignadas? (⭐⭐)
33. ¿Dónde está la referencia geográfica más repetida en la tierra de toda la Wikipedia Español? (⭐)
34. ¿Quién es el usuario que más ha bloqueado a otros? (⭐)
35. ¿Cuál es el mínimo que ha durado desde su registro un usuario bloqueado en la plataforma? (⭐⭐)
36. La 3-upla de palabras más común en los comentarios de los logs (⭐⭐)
37. ¿Cuál es el contenido al que más se hacen redirecciones? (⭐)
38. Si decimos que la ubicación de un usuario es el promedio de la latitud y longitud de los contenidos geolocalizados para los cuales editó la última versión (ignorar usuarios que no editaron contenido geolocalizado). ¿Cuáles son los dos usuarios más cercanos? (⭐⭐⭐)
39. ¿Cuál es la acción más realizada por usuarios que no están registrados? (⭐⭐)
40. Si decimos que un usuario sabe un idioma cuando tiene un nivel de babel mayor o igual a 1, para aquellos que editaron una de las versiones actuales del contenido, ¿Cuál es la tasa de revisiones sin comentario que realizan en función de los idiomas que saben? (⭐⭐⭐)

Criterio de aprobación

El criterio general es que la totalidad del tp tiene que sumar 14 puntos de los 23, un 60%. Pueden hacer consultas por slack.

Criterio de reentrega

Se podrá reentregar el TP si el puntaje es >=10 y están todos los puntos desarrollados. 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 - Pandas

- Todos los ejercicios valen lo mismo que las estrellitas que tienen asignadas, a cada uno le corresponde hacer según indiquemos cual les toca:
  - 1 ejercicio de ⭐
  - 2 ejercicios de ⭐⭐
  - 1 ejercicio 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.

Segunda parte - Visualización de datos

1. Cada visualización vale un punto, y debe cumplir con las siguientes condiciones:
    1. Debe explicarse por sí misma, sin necesidad de texto aclaratorio.
    2. Debe tener rótulos en los ejes que corresponda y en el título.
    3. Debe mostrar una relación con el target que sea clara.
    4. El uso del color debe ser intencional, elegido por ustedes, no por la librería.
    5. La visualización debe ser legible (Un bar chart de 40 barras por ejemplo es ilegible)
2. Debe cumplir el objetivo propuesto: Les recomendamos preguntar en clases de consultas o por slack, vamos a estar guiandolos en este punto. Dado que la elección de este dataset es personal, pueden ir compartiendo sus ideas/bocetos o consultando cosas en *#consultas-tp1-visu*.

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:
    - 1 ejercicio de ⭐
    - 2 ejercicios de ⭐⭐
    - 1 ejercicio 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 https://www.youtube.com/watch?v=zHYJoCgzDOw (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).

Pueden ver como es el formato de entrega acá.

Puntos extra (hasta tres ⭐)

Utilizamos el promedio del puntaje normalizado de cada kahoot/parcialito para armar un podio. El podio se modificara a medida participen en los Kahoots. Quien esté primero recibira tres ⭐, quienes estén segundos o terceros recibiran dos ⭐ extra. Quienes estén en cuarto y quinto puesto un ⭐ extra.

Asignaciones de Ejercicios

Legajo	Alumno	Ejercicios Pandas	Ejercicios Spark
106053	AAB, LETICIA ISABEL	33, 47, 54, 60	13, 15, 31, 7
107742	ABUIN, AQUILES EZEQUIEL	19, 47, 54, 57	11, 20, 32, 6
106905	AGHA ZADEH DEHDEH, LUCIA	29, 11, 18, 60	10, 12, 22, 5
100685	AGUILAR BUGEAU, PEDRO JOSE	2, 7, 8, 41,	3, 12, 28, 4
104221	AGUILAR, PEDRO	33, 6, 7, 55	2, 20, 22, 40
107539	AGUIRRE ARGERICH, FACUNDO AGUSTÍN	45, 12, 16, 48	1, 15, 31, 38
79558	ALBORNOZ, ROMINA CARLA	2, 23, 8, 60	37, 22, 36, 25
96283	APARICIO ROTERMUND, AXEL	29, 10, 12, 50	34, 31, 35, 24
108229	ARGÜELLES, MAIRA LUCÍA	2, 7, 28, 55	33, 9, 15, 21
108434	AVALOS, VICTORIA BELEN	24, 5, 9, 42	30, 22, 36, 8
108317	BALDI MORALES ALVES, TOMÁS	14, 9, 58, 60	27, 9, 39, 7
109071	BARBALASE, AGUSTIN	45, 20, 15, 55	26, 20, 32, 6
107754	BAT MENTZEL, MARCOS EZEQUIEL	25, 10, 22, 40	23, 12, 36, 5
100862	BENITEZ POTOCHEK, TOMÁS ARI	3, 9, 10, 42	19, 28, 29, 4
106841	BENITEZ, NAHUEL TOMAS	26, 16, 8, 59	18, 22, 36, 40
108100	BENITO, AGUSTÍN	45, 9, 23, 51	17, 12, 15, 38
108921	BIANCHI FERNANDEZ, MARCOS	36, 58, 38, 52	16, 36, 39, 25
106005	BIANCUZZO, JUAN IGNACIO	29, 15, 12, 55	14, 20, 35, 24

Legajo	Alumno	Ejercicios Pandas	Ejercicios Spark
97106	BONASTRE, LUCAS	34, 7, 23, 55	13, 15, 32, 21
101505	BOTTER BRUN, JUAN BAUTISTA	14, 13, 15, 44	11, 20, 28, 8
86088	BOZUNOVSKY, MARCELO	3, 5, 58, 40	10, 20, 22, 7
105288	BRIZUELA, SEBASTIAN	3, 28, 30, 44	3, 12, 20, 6
97640	BRONDO, FACUNDO LUCIO	36, 58, 15, 57	2, 9, 31, 5
103523	BUONO, FERNANDO	31, 47, 54, 52	1, 29, 32, 4
108025	CABIBBO ARTEAGA, NEHUÉN DANIEL	34, 54, 15, 48	37, 28, 36, 40
107143	CALDERON, GONZALO MANUEL	34, 6, 35, 43	34, 9, 31, 38
105161	CALLEBAUT, MELINA	2, 23, 27, 43	33, 9, 20, 25
107662	CIVINI, DIEGO EMANUEL	3, 8, 11, 52	30, 9, 39, 24
108664	CORREA, LUCAS VALENTIN	31, 20, 6, 49	27, 28, 29, 21
102439	CORRIONERO, LUAN SHAIR	26, 30, 35, 50	26, 12, 36, 8
104319	CUETO QUINTO, ALAN RAMIRO	36, 10, 11, 59	23, 15, 32, 7
107923	CWIKLA, MARTIN JUAN	29, 22, 6, 42	19, 31, 39, 6
108645	DALL'ACQUA, DENISE	29, 27, 22, 48	18, 32, 36, 5
106175	DAVILA SANCHEZ, MANUEL JESUS	56, 12, 9, 44	17, 12, 22, 4
101830	DE SANTIS, FEDERICO EZEQUIEL	19, 16, 17, 46	16, 9, 31, 40
108671	DEMARCO, JUAN PEDRO	32, 37, 30, 46	14, 15, 39, 38
93025	FARIÑA, NOELIA NOEMI	24, 12, 37, 46	13, 32, 36, 25
107491	FERNANDEZ, JULIO MATEO	39, 30, 8, 46	11, 31, 35, 24
106829	FIEGL, LUCAS AUGUSTO	24, 20,22, 55	10, 28, 32, 21
108239	FIOROTTO, CAMILA	3, 5, 13, 57	3, 20, 31, 8
87039	FLORES SOSA, ZORAIDA YURICO	14, 54, 58, 42	2, 12, 36, 7

Legajo	Alumno	Ejercicios Pandas	Ejercicios Spark
108571	FRANCAVILLA, CANDELA SOFIA	26, 18, 35, 44	1, 32, 35, 6
105658	GALDO MARTINEZ, MARIANA	14, 38, 47, 48	37, 12, 36, 5
107587	GALLINO, PEDRO	2, 54, 27, 51	34, 28, 39, 4
105892	GAMBERALE, LUCIANO MARTIN	25, 22, 35, 52	33, 28, 36, 40
109667	GEMETTO, VALENTINA MARIA	1, 8, 30, 59	30, 12, 29, 38
106998	GHOSN, LAUTARO GABRIEL	32, 9, 18, 41	27, 20, 32, 25
108937	GRIN, PEDRO	39, 10, 54, 53	26, 22, 31, 24
107985	GÜLDEN, JUAN FRANCISCO	31, 13, 10, 43	23, 12, 39, 21
105711	HERNANDEZ, JUAN CRUZ	19, 54, 58, 49	19, 12, 15, 8
108344	JANAMPA SALAZAR, MARIO RAFAEL	19, 20, 6, 40	18, 15, 29, 7
106079	JANON, SANTIAGO IGNACIO	34, 20, 7, 40	17, 28, 39, 6
106136	LABOUR, VALENTIN	39, 7, 5, 42	16, 9, 36, 5
108257	LANZILLOTTA, VALENTINA	4, 12, 7, 59	14, 31, 35, 4
108068	LEDESMA, MARTÍN	36, 54, 58, 49	13, 12, 31, 40
105993	LEVI, DOLORES	26, 17, 9, 53	11, 9, 29, 38
107552	LLORENS, IÑAKI	56, 22, 15, 49	10, 9, 35, 25
100566	LOPEZ, SANTIAGO	1, 5, 6, 40	3, 22, 28, 24
108460	MARTINEZ, FRANCISCO EZEQUIEL	25, 47, 9, 43	2, 12, 29, 21
104889	MAZZARO, FRANCO DARIO	56, 17, 18, 41	1, 12, 31, 8
106438	MIGUEL, THEO	4, 30, 23, 50	37, 29, 32, 7
105876	MINELDIN, RAMIRO	24, 54, 9, 51	34, 9, 31, 6
106999	MORALES, JULIAN LISANDRO	33, 17, 30, 42	33, 32, 35, 5
108091	MORILLA, MARTIN	39, 17, 38, 50	30, 9, 32, 4

Legajo	Alumno	Ejercicios Pandas	Ejercicios Spark
106248	MOYANO, ANDRES RICARDO	1, 37, 6, 46	27, 20, 39, 40
107752	MURSELI, AGUSTIN	24, 18, 13, 60	26, 9, 39, 38
99479	MUTCHINICK, JULIAN	45, 27, 38, 60	23, 29, 31, 25
107690	OJEDA, DANIELA	4, 22, 17, 53	19, 9, 22, 24
108397	ORDOÑEZ, ALEJO	21, 35, 28, 41	18, 31, 35, 21
108013	ORONA, IGNACIO	33, 22, 27, 46	17, 28, 39, 8
87622	OURA, JACQUELINE JUDIT OLGA	19, 23, 7, 43	16, 9, 36, 7
108755	PALOPOLI, MAXIMO	34, 58, 17, 51	14, 20, 32, 6
108215	PANDOLFI, JOAQUIN	1, 35, 9, 53	13, 9, 35, 5
106249	PAPA, FRANCO	2, 18, 47, 48	11, 28, 39, 4
102340	PAULOZZI MOLINA, GERONIMO	25, 27, 28, 49	10, 22, 36, 40
105600	PAZ BLANCO, PILAR	4, 35, 37, 46	3, 28, 35, 38
101947	PERALTA, FEDERICO MANUEL	24, 22, 23, 49	2, 12, 39, 25
107997	PEREZ GOLDSTEIN, JULIETA	32, 47, 12, 44	1, 28, 32, 24
105867	PIÑANGO RAMOS, JULIO CESAR	21, 28, 11, 50	37, 29, 31, 21
91076	PORRAS CARHUAMACA, SHERLY KATERIN	21, 18, 27, 44	34, 9, 12, 8
107788	QUIROGA, BRUNO MARTIN	26, 9, 47, 41	33, 12, 36, 7
106007	RAIMONDI, LUCAS NAHUEL	31, 30, 17, 57	30, 28, 31, 6
93751	RAMIREZ, JOSE ISRAEL	26, 37, 27, 49	27, 15, 28, 5
99770	REA, MATIAS ABRAHAM	56, 58, 15, 40	26, 29, 32, 4
106716	REIMUNDO, MARTIN	19, 27, 47, 52	23, 32, 36, 40
108127	RICO, MATEO JULIÁN	56, 16, 30, 52	19, 12, 31, 38
86601	RIPETOUR CHAIMAN, DIEGO	4, 17, 5, 41	18, 12, 35, 25

Legajo	Alumno	Ejercicios Pandas	Ejercicios Spark
106041	RIVERA VILLATTE, MANUEL	32, 16, 5, 59	17, 20, 32, 24
106302	RIVERO TRUJILLO, TOBIAS LUCIANO	3, 38, 30, 49	16, 31, 35, 21
101891	RODRIGUEZ, NAZARENO JOSE LUIS	21, 18, 20, 48	14, 20, 22, 8
96713	ROLDAN MONTES, CRISTIAN EDUARDO	31, 38, 11, 51	13, 20, 35, 7
101043	RONCHI, SANTIAGO AGUSTIN	4, 11, 12, 43	11, 32, 36, 6
106835	RUANO FRUGOLI, CLARA	25, 15, 9, 57	10, 15, 32, 5
106768	RUIZ SUGLIANI, SANTIAGO NAHUEL	21, 11, 35, 53	32, 28, 29, 4
106147	SABAJ, GASTON EZEQUIEL	45, 17, 35, 43	2, 35, 32, 40
99131	SECCHI, ANA MARIA	39, 12, 10, 59	1, 12, 31, 38
107185	SHIMABUKURO, GONZALO JOAQUÍN	36, 13, 16, 44	37, 12, 32, 25
104892	SICCA, FABIO	1, 20, 22, 42	34, 12, 20, 24
108679	SILVANO LIMA, BAUTISTA	33, 16, 35, 50	33, 29, 31, 21
93735	SOSA AQUINO, RICARDO ARIEL	25, 38, 10, 48	30, 35, 36, 8
103227	SOTO BERTANI, SEBASTIÁN MATIAS	29, 37, 38, 51	27, 20, 29, 7
106673	SOTO, MARILYN NICOLE	14, 11, 58, 51	26, 15, 28, 6
109200	SOUZA, MARTINA FLORENCIA	56, 28, 23, 57	23, 9, 31, 5
104239	SPRENGER, ROBERTA	34, 8, 9, 57	19, 15, 35, 4
107746	SUAREZ PINO, IMANOL	21, 10, 28, 59	18, 39, 32, 40
107710	SZEJNFELD SIRKIS, TOMAS	14, 6, 13, 55	17, 12, 36, 38
104509	URSINO, IAN MIKA	39, 12, 13, 60	16, 35, 22, 25
107585	VALDEVENITO, ALAN EZEQUIEL	1, 5, 18, 50	14, 12, 22, 24
97076	VARGAS CHAVEZ, RODRIGO IGNACIO	33, 20, 18, 53	13, 28, 29, 21
104115	VERA BENITEZ, SEBASTIAN	32, 58, 5, 53	11, 9, 31, 8

Legajo	Alumno	Ejercicios Pandas	Ejercicios Spark
104734	VERNIERI, ANITA	45, 15, 16, 40	10, 15, 22, 7
106129	VETRANO, IGNACIO EZEQUIEL	36, 16, 20, 41	3, 31, 20, 6
106930	VIAU, IGNACIO	31, 13, 28, 40	2, 29, 12, 5
97023	YBARRA ESCALANTE, DIEGO EMANUEL	32, 47, 23, 52	1, 32, 39, 4
101656	Del Pozo, Francisco Marcelo	4, 10, 28, 46	3, 12, 36, 4
102912	Pucci Romero, Tobias	26, 30, 16, 48	13, 15, 29, 38

This page was generated by GitHub Pages.