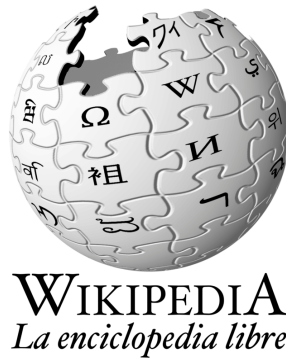


TP2 - Spark

Spark (15 ptos)



Como en nuestro primer trabajo práctico utilizaremos 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.



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

id	Id del contenido
title	Título del contenido
text	Texto

geo_tags.csv

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.

item_id	ID del ítem afectado
timestamp	Timestamp del log
contributor_username	Username que realizó la acción
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

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.

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

categorylinks.csv











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.

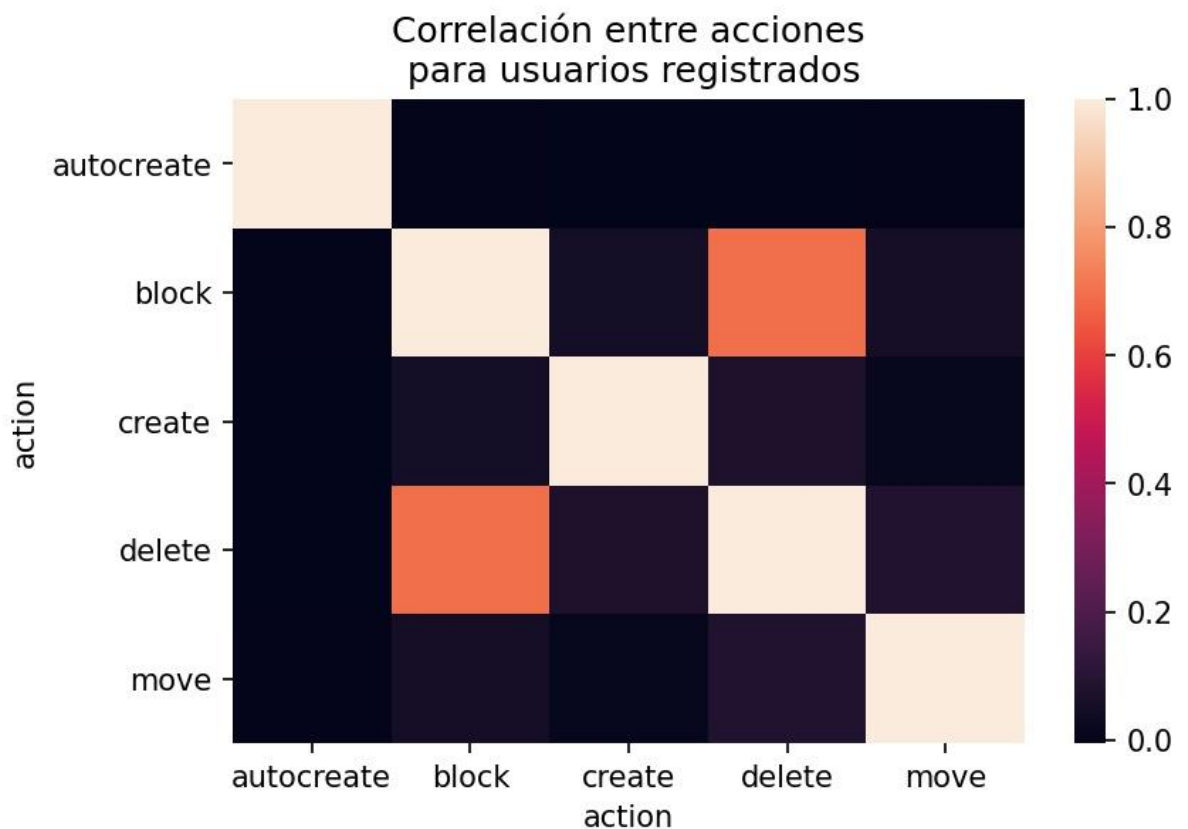
pl_from	ID del contenido donde está el link
pl_namespace	👤
pl_title	Título del contenido al cual va el link
pl_from_namespace	👤


Realizar sus correspondientes consultas en Spark


1. Considerando los logs de acciones realizadas sobre ítems, mostrar el top 10 de ids de ítems que fueron afectados por mayor cantidad usuarios distintos ()
2. Considerando los logs de acciones realizadas sobre ítems, mostrar el top 10 de ids 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 deberían 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 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 trigramma 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:
 - a. key sea una palabra del léxico de la wikipedia
 - b. value sea una lista donde cada elemento de la misma sea una tupla de dos elementos
 - i. identificador de contenido donde aparezca esa palabra.
 - ii. 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 6 puntos de los 10, un 60%. Cada uno va a tener un ayudante asignado, pueden hacer consultas a por slack o piazza o a sus ayudantes.

Spark

- Todos los ejercicios deben realizarse utilizando el API de RDD de Spark.
- Todos los ejercicios valen lo mismo que las “Maurice Gibb” que tienen asignados, a cada uno le corresponde hacer según indiquemos cual les toca:
 - 3 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-tp2, 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: “<NÚMERO DE CONSIGNA> - 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 con su ayudante asignado
- 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!