Sistemas de recuperación de Información Web, 2019-2

Trabajo 2 (45%). Integrantes: Máximo 4

Objetivo: Desarrollar un sistema de recomendación utilizando información web, extraída a través de las técnicas web *scraping* y web *crawling*

Cada equipo deberá seleccionar un dominio elegido de los 6 posibles indicados al final del archivo, y deberá anotar los nombres de los integrantes en el archivo de Drive, el cual tiene los temas ya establecidos (Deberá anotarse en la casilla correspondiente al tema que seleccionó. Únicamente se permite repetir los temas 1 y 2)

https://docs.google.com/spreadsheets/d/1txzrEudgCgm7fh-UdTNYkWKKpkKVFl6oeJfbXYRjOJQ/edit?usp=sharing

Se recomienda leer toda la especificación del trabajo antes de iniciar a desarrollarlo.

Se desarrollará un aplicativo que deberá realizar una consulta en diferentes páginas web (Las páginas web que deben consultar se encuentran al final, según el tema que seleccione) para obtener la información de los ítems, y ofrecer a los usuarios recomendaciones de los mismos.

El aplicativo debe tener una interfaz gráfica y los datos serán almacenados en un sistema de gestor de bases de datos. Puede elegir el lenguaje que desee para la programación, pero debe tener presente que realizarlo en Python será mucho más sencillo. Si no desea realizar todo el software en Python, puede realizar ciertas partes como el web *crawling* y la recomendación en Python, y todo lo demás en otro lenguaje, pero debe realizar un método de comunicación entre ambas partes.

Puede elegir el sistema de gestión de bases de datos que desea (relacional o no relacional).

Existirán dos aplicaciones: La primera será el software de recomendación como tal, y al cual pueden acceder los usuarios. La segunda será una aplicación de escaneo de las webs y no será accesible por los usuarios. Únicamente el primer numeral se refiere a la aplicación de escaneo de webs, mientras que todos los demás se refieren al software de recomendación como tal. Existirá una única base de datos a la cual ambas aplicaciones acceden.

1. **(40%) Recopilación de ítems**: Debe existir una aplicación independiente, encargada de obtener la información de los ítems de las páginas web. Dicha aplicación será lanzada cuando el administrador desee y se encargará de escanear las páginas web

a través de las técnicas de web *crawling* y web *scraping*, para posteriormente almacenar los datos de los ítems en la base de datos.
Las características que se deben extraer de cada ítem se encuentran al final, según el tema escogido. Debe existir una tabla (o equivalente en el caso de una no relacional) que almacene la información de cada ítem, cuyas características serán almacenadas en las columnas (o equivalente). También debe guardar el link para acceder de manera directa al ítem en cada una de las páginas.
Tenga presente que debe realizar el escaneo de todas las páginas posibles en cada sitio web.

El proceso de escaneo de webs y guardado de información debe cumplir que:

A. Si en varios sitios webs se encuentran ítem con la misma identificación única (indicada en el tema), se deberá tratar como el mismo ítem, por lo tanto, se debe generar una única fila (o equivalente) en la base de datos, y se guardará el link de acceso de cada una.
   Ejemplo: El televisor indicado en https://www.tiendasjumbo.co/televisor-led-65-samsung-uhd-un65ru7100kxzl/p?idsku=20048688 y https://www.falabella.com.co/falabella-co/product/3863566/Televisor-Samsung-43-Pulgadas-4K-UHD-Smart-Tv-UN43RU7100/3863566 es el mismo, ya que posee el mismo modelo: **UN65RU7100KXZL**. Por lo tanto, deberá existir una única fila en la base de datos del tipo: **UN65RU7100KXZL**, Samsung, 55, ……, link1, link2

   La extracción de las características debe darse de la siguiente forma:

   Se debe obtener el valor de cada característica en cada uno de los dos sitios webs, si el valor coincide en ambos sitios, se guardará el valor, pero si difiere se deberá seleccionar uno de los dos y se deberá guardar en una columna de observaciones un texto de la forma: "Diferencia de valor encontrado para C. Se selecciona el valor X. Posibles valores: X en S1 y Y en S2", donde C es la característica (ejemplo Entradas VGA), X el valor seleccionado (Ejemplo Sin entradas), Y el otro valor encontrado (ejemplo 0), S1 y S2 los sitios web, cortando únicamente el dominio principal (ejemplo tiendasjumbo.com.co y falabella.com.co) **->** *Esto se hace bajo la suposición de que la comparación de igualdad viene dada por igualdad de cadenas: Sin entradas != 0, sin embargo en un caso ideal se debería comprobar que ambos textos tienen el mismo significado, pero no es alcance del trabajo incluir dicha característica (Aunque si quiere hacerlo….)*

Debe tener presente que los nombres de las características se pueden llamar diferente en cada sitio web. Nótese que, para los televisores mencionados anteriormente, en Jumbo se indica en el campo "**CATEGORIA TV**" el valor LED, mientras que en Falabella se indica en el campo "**Tecnología**".
Cuando en una de las páginas no se encuentre información para una característica, se guardará el valor que se encuentra en la otra página, y se guardará una observación de la forma "Único valor de característica para C, valor faltante en S", donde C es la característica (ejemplo: procesador) y S el sitio web donde no se encontró la información (ejemplo: www.tiendasjumbo.com.co)

Tenga presente que se pueden generar varias observaciones ya que se analizan varias características por cada ítem, por lo tanto, se recomienda por simplicidad utilizar un único campo de observaciones e ir concatenando en una nueva línea.

B. La aplicación se puede correr varias veces, por lo tanto, debe analizar la información de los ítems ya almacenada en la base de datos de tal forma que:
    - Los ítems nuevos que se publicaron en los sitios web se deben agregar a la base de datos
    - Los ítems que están en al menos un sitio web y se encuentran en la base de datos se deben actualizar siguiendo el mismo proceso del numeral A
    - Los ítems que fueron eliminados de ambos sitios webs se deben "desactivar" del sistema de recomendación de modo tal que no se visualicen en nuevas recomendaciones ni el explorador de ítems, pero sigan siendo relevantes para el cálculo de la recomendación.
      Puede utilizar una variable booleana que indique cuando el ítem fue desactivado, una variable que indique el estado, o una fecha de desactivación.

Se recomienda utilizar las librerías beautifulsoup o scrapy de Python para realizar el proceso de *scraping* y *crawling*. Tenga presente que, si bien se encuentran diferentes definiciones de ambos conceptos y ambos están estrechamente relacionados, se puede concluir que el *scraping* realiza la recopilación de la información de un sitio web, mientras que el *crawling* navega en varias páginas de un sitio web o en varios sitios web de manera automática para realizar ciertas operaciones, como por ejemplo indexado o *scraping*.

Si desea utilizar *scrapy* puede seguir algunos de los siguientes links para la documentación:
https://docs.scrapy.org/en/latest/intro/tutorial.html
https://medium.com/better-programming/develop-your-first-web-crawler-in-python-scrapy-6b2ee4baf954
https://www.datacamp.com/community/tutorials/making-web-crawlers-scrapy-python

Siguiendo el tutorial de la documentación oficial puede crear un spider que navegue a todas las páginas del sitio *http://quotes.toscrape.com*, y obtendrá un código similar al siguiente, el cual muestra por pantalla el texto de la cita y el autor

```
import scrapy

class QuotesSpider(scrapy.Spider):
name = "quotes"
start_urls = [
  'http://quotes.toscrape.com/page/1/',
  'http://quotes.toscrape.com/page/2/',
]
i = 0

def parse(self, response):
  for quote in response.css('div.quote'):
      print("TEXTO ", quote.css('span.text::text').get())
      print("AUTOR ", quote.css('small.author::text').get())

  next_page = response.css('li.next a::attr(href)').get()
  if next_page is not None:
      next_page = response.urljoin(next_page)
      print("******************PAGINA ", self.i)
      self.i = self.i + 1
             yield scrapy.Request(next_page, callback=self.parse)
```

También puede utilizar Beautifulsoup para navegar entre páginas siguiendo la idea realizada en el Taller.

2. **(5%) Control de usuarios:** El usuario debe tener una interfaz gráfica para realizar el registro en el sistema. Se debe guardar los nombres, correo y contraseña.
   El ingreso al sistema será a través del correo y contraseña de cada usuario.
   Los datos de los usuarios deben ser almacenados en la base de datos.

3. **(10%) Explorador de ítems y registro de calificación**: Al iniciar sesión el usuario tendrá en una interfaz gráfica principal la lista de todos los ítems, en la cual se deben visualizar el identificador único, las características y la calificación que el usuario ha otorgado a ese ítem (Si no se ha calificado aparecerá vacía). Para cada ítem debe existir una opción que permita visualizar las observaciones y links (del numeral 1. Al dar clic sobre el link debe abrir la página en el navegador), y registrar la calificación (o editar la calificación otorgada).

4. **(30%) Recomendación**: Se debe crear un sistema de recomendación híbrido, el cual proporcionará la recomendación de los ítems a los usuarios basado tanto en contenido como en colaborativo. Los sistemas híbridos permiten obtener una recomendación más precisa al tener en cuenta tanto las características de los ítems como la similitud entre usuarios.
   Se deben desarrollar dos sistemas de recomendación independientes que luego será combinados para dar la recomendación final. El sistema de recomendación por contenido deberá tener en cuenta todas las características de los ítems y deberá almacenar un perfil de usuario con las preferencias para cada característica. Dicho perfil debe ser calculado de manera automática y deberá existir una opción donde el usuario los pueda visualizar por interfaz gráfica.
   El sistema de recomendación colaborativo puede ser de tipo usuario-usuario o de tipo ítem-ítem. Puede seleccionar la técnica que prefiera para la recomendación siempre y cuando sea bien aplicada (kNN, Factorización de matrices, Redes Neuronales, etc.)
   Existen varios métodos para realizar la recomendación híbrida. En el siguiente artículo se realiza una revisión de la literatura y se muestran los posibles métodos para combinar las recomendaciones: https://arxiv.org/pdf/1901.03888.pdf (Sección 3.5). Uno de los métodos más populares es la recomendación por peso, en la cual se le da un peso a la recomendación por contenido y un peso a la recomendación colaborativa. Dichos pesos se deben ir ajustando con una evaluación a sus valores ideales. El siguiente artículo propone una combinación por medio de redes bayesianas: https://core.ac.uk/download/pdf/81120133.pdf. Sin embargo, para efectos del trabajo basta con asignar un peso a cada uno de los sistemas que considere adecuado para iniciar, justificando adecuadamente el por qué de su suposición (En el caso ideal se debería realizar luego una evaluación para ir ajustando los pesos. *Aunque si quiere hacerlo...*)

   Debe existir una opción en la interfaz gráfica que permita visualizar los ítems más recomendados para el usuario logueado, ordenados del más recomendado al menos recomendado, con un límite de 10 ítems. Debe existir una opción para cada ítem de esta sección en la cual el usuario pueda indicar si la recomendación fue acertada o no, es decir, si le gustó o no el ítem que el sistema le recomendó. Se deberá medir y

guardar los aciertos que el sistema ha tenido, con el fin de calcular una precisión o exactitud, la cual debe ser visualizada en la interfaz del perfil de usuario.
También debe existir una opción que le permita ingresar la calificación del ítem (Puede ser la misma interfaz utilizada en el numeral anterior). Tenga presente que la forma de indicar si el ítem le gustó o no al usuario puede ser diferente a la calificación otorgada, según la forma como haya planteado el sistema. Cree el método de evaluación y justifíquelo.

5. **(15%) Cold start problem**: Controle de manera adecuada el problema de inicio en frío, cuando ingresa un usuario nuevo al sistema y cuando ingresa un ítem nuevo. Justifique la forma como manejó dicho problema.

Ítems a entregar:

- Base de datos

- Código y ejecutable de la aplicación de web *scraping*

- Código y ejecutable de la aplicación de recomendación

- Archivo de Word con los integrantes y las justificaciones solicitadas

Nota: El trabajo deberá ser sustentado en la fecha indicada. La asistencia de todos los integrantes del equipo es obligatoria y la única excusa para no asistir es aquella validada por la universidad. Se seleccionará aleatoriamente un integrante de cada equipo para sustentar, así que asegúrate de incluir en el trabajo solo aquellos que realmente trabajaron. Si la sustentación se realiza de manera adecuada, la nota final es la nota del trabajo según la calificación de los parámetros. Si la sustentación no se realiza de manera adecuada, la nota final del trabajo se puede ver disminuida hasta en 1.5

Temas

1. **Muebles armables**
   Sitios web:
   https://inval.com.co/col/
   https://www.homecenter.com.co/homecenter-co/

En la página de inval se encuentran los muebles fabricados por ellos. En la página de Homecenter se debe filtrar los muebles marca inval.
Tenga presente que en la página de inval los muebles se visualizan por categorías, por los que se deben escanear todas.

Identificador único: Referencia
En Homecenter se encuentra como Referencia proveedor mientras que en inval se encuentra como el nombre del producto. Para efectos del trabajo considere la referencia en Homecenter hasta el guion, en el caso que lo tenga. Por ejemplo, si en Homecenter aparece CC741-A considérelo únicamente como CC741

Características:
Tipo o categoría (Mesa, centro computo, armario, etc.)
Precio
Peso
Ancho
Alto
Profundidad o Fondo

2. **Libros**
   Sitios web:
   https://www.planetadelibros.com.co/
   https://www.panamericana.com.co/

   En cada sitio se deben tener en cuenta los libros de la sección literatura y sus sub secciones.

   Identificador único: Nombre del libro, Autor
   La identificación única de cada libro es su nombre y el nombre del autor. Tenga presente que la comparación la debe realizar ignorando mayúsculas, minúsculas, tildes, espacios y guiones, para reducir la probabilidad de que dos ítems iguales se guarden como diferentes

   Características:
   Autor
   Editorial
   Número de páginas
   Precio

3. **Televisores**
   Sitios web:

https://www.walmart.com/browse/tv-video/all-tvs/3944_1060825_447913
https://www.bestbuy.com/site/tvs/all-flat-screen-tvs/abcat0101001.c?id=abcat0101001

Se deben buscar los televisores en cada uno de los dos sitios.

Identificador único: Modelo

Características:
Marca
Precio
Tamaño de pantalla
Resolución
Tipo de display

4. **LEGO**
   Sitios web:
   https://www.target.com/b/lego/-/N-56h5n?lnk=snav_rd_legos
   https://www.lego.com/en-us/themes

   Se deben considerar los ítems en las categorías: Architecture, City, Friends, Batman, Minecraft

   Identificador único: Número de modelo
   Nótese que en la página de target se encuentra en el título, mientras que en la página de lego se encuentra como # ítem

   Características:
   Categoría, tema o colección
   Precio
   Número de piezas

5. **JUEGOS PS4**
   Sitios web:
   https://www.gamesmen.com.au/video-games/ps4/games
   https://www.buygames.ps/en/ps4-games

   Identificador único: Nombre de juego
   Se deben extraer únicamente el nombre del juego para su comparación. Nótese que en el caso de buygames el texto "PS4" se debe eliminar. En algunos ítems de

gamesmen se relaciona texto en paréntesis o corchetes que tampoco se debería tener en cuenta. Tenga presente que la comparación la debe realizar ignorando mayúsculas, minúsculas, tildes, espacios y guiones, para reducir la probabilidad de que dos ítems iguales se guarden como diferentes

Características:
Publicador
Desarrollador
Género
Edad recomendada

6. **IMPRESORAS HP**
   Sitios web:
   https://www.provantage.com/~67PPRNT_.htm
   https://www.bestbuy.com/site/computers-pcs/printers/abcat0511001.c?id=abcat0511001

   Se deben tener en cuenta únicamente las impresoras marca HP

   Identificador único: Número de modelo

   Características:
   Tipo de color (Monocromática, Color)
   Tiene Wifi
   Posee escáner
   Precio