Unidad 3: Integración de SGBD en la construcción de software

Proyecto: Bases de datos de medios de prensa e información territorial

---

Parte A: Construcción de una base de datos sobre medios de prensa hispanohablantes (Profesor responsable: Matthieu Vernier)

Caso de estudio:

Para facilitar la investigación en ciencias de la comunicación y ciencias políticas, nos interesa construir una base de datos sobre los medios de prensa de distintos paises hispanohablantes.

La lista de los 20 paises hispanohablantes es:

- España, México, Colombia, Argentina, Perú, Venezuela, Chile, Ecuador, Guatemala, Cuba, Bolivia, República Dominicana, Honduras, Paraguay, El Salvador, Nicaragua, Costa Rica, Puerto Rico, Uruguay, Panamá.

La base de datos deberá permitir a investigadores acceder a distintos tipos de información sobre los medios de prensa. En particular es interesante almacenar información sobre los medios de prensa, como por ejemplo:

- su nombre,
- su ubicación geográfica (ciudad, región, país, continente)
- su año de fundación
- su cobertura que puede ser de varios tipos (internacional, nacional y/o local)
- el nombre de su(s) fundadore(s)

Por cada medio de prensa es interesante saber cuál es su sitio web, y cuáles son sus cuentas en Facebook, Instagram y Twitter. Por cada cuenta de redes sociales es interesante saber cuántos seguidores tiene y cuándo fue la última vez que se actualizó el número de seguidores.

Los sitios web de cada medio de prensa se dividen en distintas categorías, por ejemplo “política”, “cultura”, “deporte”, etc. Cada categoria tiene un nombre y una URL principal (por ejemplo “http://www.cuscoenportada.com/category/politica”). Por cada categoria, nos interesa almacenar un ejemplo de URL de una de las paginas de esa categoria (por ejemplo “https://thetimes.cl/category/tecnologia/page/2/”). Finalmente, nos interesa almacenar cuál es la expresión XPATH que permite obtener los enlaces de las noticias en esa página.

Por otra parte, por cada medio de prensa, nos interesa almacenar 1 ejemplo de noticia publicada por este medio. Por cada noticia, sería interesante almacenar:

- la URL de la noticia
- la expresión XPATH que permite leer la fecha de la noticia
- la expresión XPATH que permite leer el título de la noticia
- la expresión XPATH que permite leer el contenido de la noticia

Trabajo

Formando equipos de 2 o 3 estudiantes, diseñarán e implementarán la base de datos del caso de estudio.

Rellenarán su base de datos con datos reales. Cada grupo se enfocará en un solo país, distinto de los otros grupos. Se espera que su base de datos incluye entre 60 y 100 medios de prensa por pais.

Programarán en Python dos scripts Python:

- *insert.py*: este script permite a un usuario insertar un nuevo medio de prensa en la base de datos. A través de un terminal, el script realiza una serie de preguntas al usuario. Por ejemplo:
    1. “Indica la información del nuevo medio que quieres añadir con el formato siguiente: nombre, ciudad, region, pais, continente, año de creación”
    2. “Indica una URL de una noticia de este medio”
    3. “Indica la expresión XPATH que permite leer el título de la fecha”

- *read.py:* este script permite realizar distintas consultas de lectura en la base de datos. Por ejemplo: “¿Cuál es el XPATH para leer la fecha del medio “La Tercera” en Chile?” o “Cuáles son las categorías de un medio”.

- *crawling.py*: este script recibe el nombre de un medio y el nombre de una categoría y devuelve los enlaces de una de las páginas de esta categoría.

- *scraping.py*: este script recibe como input el nombre de un medio de prensa y la URL de una noticia. Devuelve el titulo y fecha de publicación de esta noticia.

Pauta de evaluación de la parte A del proyecto

El grupo realizó un **diccionario de datos** para describir los datos del problema	1 punto
El grupo diseñó correctamente un **diagrama Entidad-Relación** para describir conceptualmente la relación entre los datos del caso	2 puntos
El grupo diseñó un **modelo relacional coherente** e lo implementó en un SGBD relacional	2 puntos
La base de datos incluye datos sobre 50 a 100 medios de prensa	2 puntos
El script Python permite integrar facilmente y sin error nuevos datos en la base de datos	1 punto
El script Python permite leer la base de datos y hacer pruebas de crawling y scraping de datos	2 puntos

---

Parte B: Información territorial basada en las comunas de Chile (Profesor responsable: Luis Veas)

Los objetivos de esta parte del proyecto son:

1) Realizar la exploración, extracción, recolección y análisis de datos territoriales mediante la creación y uso de sistemas informáticos basados en base de datos.

2) Crear un índice de bienestar comunal basado en las variables obtenidas.

3) Trabajar con datos reales que les permitan apreciar la importancia del correcto uso de herramientas computacionales, en este caso particular sql.

La forma de trabajo es la siguiente:

- Deberán explorar los links con información territorial entregados, de ser necesario pueden agregar links nuevos
- Basado en los datos exploradores, deberán crear un modelo entidad relación de los datos, que les permitan contestar, mediante queries sql, una serie de preguntas relacionadas con los dos datos del estudio.
- Basado en el diagrama entidad-relación crear el diagrama relacional
- Crear el diccionario de datos
- Crear su base de datos
- Deberán crear un proceso manual o automatizado de extracción de datos desde distintos sitios web de gobierno y otros que disponibiliza datos reales (getData.py). El resultado de este proceso debe ser un archivo csv por tabla de su base de datos.
- Deberá crear un sistema (lo llamaremos metrics.py) que les permita:
    - Cargar a su base de datos la información de los archivos csv
    - Con la base de datos poblada (cargada de datos), deberán contestar un total de 14 consultas sql que definiremos según los datos
    - Crear tabla con indicadores de bienestar comunal, esto basado en su base de datos. en esta tabla se deben ingresar datos referentes a cuán bueno es vivir en x comuna del país
    - Poblar tabla con los indicadores de bienestar
    - Mostrar los resultados de su indicador de bienestar

Nota: las sentencias SQL serán determinadas al momento de seleccionar las variables a estudiar, cada grupo debe seleccionar a lo más 4 que sean distintas de comuna/ciudad/provincia/país

Evaluación:

Se deben conformar grupos de 2 a 3 personas y se evaluará:

Item	Descripción	puntuación
1	diagrama modelo entidad-relación	10
2	diagrama modelo relacional	10
3	diccionario de datos	10
4	script de base de datos	10
5	seleccionar las variables a estudiar (a lo menos 4)	5
6	documentar el proceso de descarga (puede ser uno o varios programas o un proceso manual)	15
7	programa de carga de datos	10
8	programa que entregue el resultado de las 14 sentencias sql	40
9	explicar su modelo de indicador de bienestar	10
10	creación/poblar/mostrar de indicador de bienestar	20

Deberán entregar un informe, un script.sql y los programas desarrollados. En el informe debe estar presente los ítems: 1, 2, 3, 5, 6 y 9. El script.sql debe contener el script de creación de la base de datos.

Links de referencia

información de (país/región/comuna)
https://repositoriodeis.minsal.cl/ContenidoSitioWeb2020/uploads/2019/11/DPA2018.xls
https://es.wikipedia.org/wiki/Anexo:Comunas_de_Chile
https://www.bcn.cl/siit/mapoteca/comunas
https://www.subdere.gov.cl/sites/default/files/documentos/articles-73111_recurso_1.pdf

Estadísticas comunales por año
https://www.bcn.cl/siit/reportescomunales/comunas_v.html?anno=2017&idcom=5602

Censo 2017 (les entrega información poblacional y de viviendas en el país)
http://www.censo2017.cl/descargue-aqui-resultados-de-comunas/

Antenas Celulares, indicador de conectividad. (Autorizadas/ En Trámite)
https://antenas.subtel.gob.cl/leydetorres/mapaAntenasAutorizadas.html

carabineros (seguridad)
https://www.carabineros.cl/detalleUnidad.php
https://datoscomunales.pazciudadana.cl/

Información de centros de salud
https://reportesdeis.minsal.cl/ListaEstablecimientoWebSite/
https://repositoriodeis.minsal.cl/DatosAbiertos/Establecimientos_ChileDEIS_MINSAL%2019-05-2023.xlsx
https://deis.minsal.cl/#datosabiertos

farmacias (acesso a medicamentos)
https://seremienlinea.minsal.cl/asdigital/index.php?mfarmacias

panorama laboral
https://www.observatorionacional.cl/panorama-laboral/datos-comunales

educación (párvulo/básica/media/profesional)
https://mi.mineduc.cl/mime-web/mvc/mime/listado
https://datosabiertos.mineduc.cl/
https://datosabiertos.mineduc.cl/planes-y-programas-de-estudio/

Entretención
https://teatroamil.cl/articulos/teatro-a-mil-en-comunas/
https://es.wikipedia.org/wiki/Anexo:Estadios_de_f%C3%BAtbol_de_Chile
https://www.fichajes.com/chile/primera-division/estadios

Transporte
https://datos.gob.cl/dataset?organization=subsecretaria_de_transporte&groups=transporte
http://www.sectra.gob.cl/encuestas_movilidad/encuestas_movilidad.htm

Ejemplo de índice de bienestar
http://www.supersalud.gob.cl/664/w3-article-7627.html
https://estudiosurbanos.uc.cl/documento/indice-de-calidad-de-vida-urbana-icvu-2021/

Ejemplo de diagrama de información comunal