



Examen Practico

27/05/2021

Nombre: Walter Bau

Objetivo:

- Consolidar los conocimientos adquiridos en clase para los métodos de búsqueda y bases de datos orientadas a grafos.

Enunciado:

- Diseñe y desarrolle un sistema recopilador que permita obtener las noticias, Facebook, twitter de los alcaldes dentro de una base de datos orientados a grafos:

Webscraping es la técnica de extraer datos contenidos en un formato no estructurado en una página web y llevarlos a una estructura fácil de usar.

Es por ello, que se desea crear nuevos métodos que permitan la recopilación masiva de información para su posterior estudio y correlación en forma de big data.

En base a ello, vamos a obtener los datos de lo que esta hablando las noticias de los candidatos dentro del Ecuador y almacenar los datos dentro de una base de datos orientadas a grafos.

- Generar un modelo que permita obtener y almacenar los datos en los grafos.
- Vincular los datos con el alcalde seleccionado.
- Se debe tener al menos 1000 nodos generados.
- Obtener de la noticia: el Link, mensaje, fecha
- Facebook: Comentarios, Publicaciones, Amigos, Likes, Seguidores, etc.
- Twitter: Usuario, mensaje, fecha, etc.
- No se debe repetir los alcaldes.
- Se puede utilizar cualquier herramienta o procesamiento para el WebScarping.
- Generar sus análisis, conclusiones y recomendaciones en base a los datos



27/05/2021

Examen Practico

```
In [1]: #IMPORTAR neomodel
from neomodel import StructuredNode, StringProperty, RelationshipTo, RelationshipFrom, config, IntegerProperty, UniqueIdProperty

#URL CONECCION CON LA BASE DE DATOS DE NEO4J
config.DATABASE_URL = 'bolt://neo4j:cuenca@localhost:11005'

#CREAR Object Browsers
class Browsers(StructuredNode):
    nombre = StringProperty(unique_index=True)
    alcalde = RelationshipTo('AlcaldeNoticias', 'BROWSERS ALCALDE')
    alcaldeFacebook = RelationshipTo('AlcaldeFacebook', 'BROWSERS ALCALDE FACEBOOK')

#CREAR Object AlcaldeNoticias
class AlcaldeNoticias(StructuredNode):
    url = StringProperty(unique_index=True)
    nombre_Pagina_WEB = StringProperty(unique_index=True)
    titulo = StringProperty(unique_index=True)
    mensaje = StringProperty(unique_index=True)
    fecha = StringProperty(unique_index=True)
    browsers = RelationshipFrom('Browsers', 'BROWSERS ALCALDE')

#CREAR Object AlcaldeFacebook
class AlcaldeFacebook(StructuredNode):
    url = StringProperty(unique_index=True)
    nombre_Pagina_WEB = StringProperty(unique_index=True)
    titulo = StringProperty(unique_index=True)
    mensaje = StringProperty(unique_index=True)
    browsers = RelationshipFrom('Browsers', 'BROWSERS ALCALDE FACEBOOK')
```

```
In [2]: #Guardar Los datos del object Browsers
browsersG = Browsers(nombre = "GOOGLE").save()
browsersE = Browsers(nombre = "ECOSIA").save()
browsersB = Browsers(nombre = "BING").save()
```

```
In [3]: import requests
from bs4 import BeautifulSoup
import json

#Palabras a Buscar

#URL buscar en noticias del alcalde del Canton Paute RAUL DELGADO
noticiasalcalde = 'https://www.google.com/search?q=raul+delgado+alcalde+de+paute&tbm=nws&sxsrf='
#Rango de paginas a buscar por Noticias
rangoNoticiasAlcalde = 60

#Encabezados HTTP Para navegadores
header = {'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/80.0.3987.16'}

cont=0
contID = 1;
contCambioPaginaP=0
for i in range(rangoNoticiasAlcalde):
    url = noticiasalcalde+str(contCambioPaginaP)
    respuesta = requests.get(url, headers = header )
    contenido = BeautifulSoup(respuesta.content, "html.parser")
    contenido = contenido.find_all('div', class_='dbsr')

    contCambioPaginaP=contCambioPaginaP+10
    cont=cont
    for lista in contenido:
        cont=cont+1

        #SACAR URL
        url = str(lista.find_all('a'))
        url = url.split('href="')
        url = str(url[1])
        url = url.split('" ping="')
        url = str(url[0])
```



Examen Practico

27/05/2021

```
#SACAR NOMBRE DE LA PAGINA WEB
nombreP = str(lista.find_all('div', class_='XTjFC WF4CUC'))
nombrePagina = nombreP.split('</g-img>')
nombrePagina = (nombrePagina[1])
nombrePagina = nombrePagina.replace('</div>]', '')

#SACAR TITULO
titulo = str(lista.find_all('div', class_='JheGif nDgy9d'))
titulo = titulo.replace('<div aria-level="2" class="JheGif nDgy9d" role="heading" style="-webkit-line-clamp:2">', '')
titulo = titulo.replace('</div>]', '')

#SACAR MENSAJE
mensaje = str(lista.find_all('div', class_='Y3v8qd'))
mensaje = mensaje.replace('<div class="Y3v8qd">', '')
mensaje = mensaje.replace('</div>]', '')

#SACAR FECHA
fecha = str(lista.find_all('span', class_='WG9SHc'))
fecha = fecha.replace('<span class="WG9SHc"><span>', '')
fecha = fecha.replace('</span></span>]', '')

#Guardar los datos del object
alcaldeNoticias = AlcaldeNoticias(url = url, nombre_Pagina_WEB = nombrePagina, titulo = titulo, mensaje = mensaje, fecha = fecha)

#Guardar los datos del object browsersG
browsersG.alcalde.connect(alcaldeNoticias)
contID = contID + 1
```

```
In [4]: import requests
from bs4 import BeautifulSoup
import json

#Palabras a Buscar

urlAlcaldeEcosia = 'https://www.ecosia.org/news?q=raul%20delgado&p='
rangourlAlcaldeEcosia = 20

#Encabezados HTTP Para navegadores
header = {'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/80.0.3987.16'}

cont=0
contCambioPagina=0
for i in range(rangourlAlcaldeEcosia):
    url = urlAlcaldeEcosia+str(contCambioPagina)
    respuesta = requests.get(url, headers = header)
    contenido = BeautifulSoup(respuesta.content, 'html.parser')
    conten = contenido.find_all('section', class_='news_results')
    conten = contenido.find_all('div', class_='result_body')
    contCambioPagina=contCambioPagina+1
    cont=cont
    for enlace in conten:
        cont=cont+1

        url = str(enlace.find_all('h2', class_='result-title'))
        url = url.split('data-v-e393cff4="" href=""')
        url = str(url[0])
        url = url.split(' rel="noopener"')
        url = str(url[0])

        nombrePagina = str(enlace.find_all('div', class_='result_info'))
        nombrePagina = nombrePagina.split('cff4="">')
        nombrePagina = str(nombrePagina[0])
        nombrePagina = nombrePagina.split('</div>')
        nombrePagina = str(nombrePagina[0])
        titulo = str(enlace.find_all('h2', class_='result-title'))
        titulo = titulo.split('target="_self">')
        titulo = str(titulo[0])
        titulo = titulo.replace('</a> </h2>]', '')

        mensaje = str(enlace.find_all('p', class_='news-result_description'))
        mensaje = mensaje.split('data-v-e393cff4="">')
        mensaje = str(mensaje[0])
        mensaje = mensaje.replace('</p>]', '')

        fecha = str(enlace.find_all('time', class_='news-result_date'))
        fecha = fecha.split('">')
        fecha = str(fecha[1])
        fecha = fecha.replace('</time>]', '')

        alcaldeNoticias = AlcaldeNoticias(url = url, nombre_Pagina_WEB = nombrePagina, titulo = titulo, mensaje = mensaje, fecha = fecha)
        browsersE.alcalde.connect(alcaldeNoticias)
```



27/05/2021

Examen Practico

```
In [5]: import requests
from bs4 import BeautifulSoup
import json

#Palabras a Buscar

urlBingalcalde = 'https://www.bing.com/news/search?q=raul+delgado+alcalde+de+paute'

#Encabezados HTTP Para navegadores
header = {'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/80.0.3987.16'}

for i in range(1):
    url = urlBingalcalde
    respuesta = requests.get(url, headers = header )
    contenido = BeautifulSoup(respuesta.content, 'html.parser')
    conten = contenido.find_all('div', class_='news-card newsitem cardcommon b_cards2')
    print("Enlace: ")
    for enlace in conten:

        url = str(enlace.find_all('a'))
        url = url.split('href="')
        url = str(url[1])
        url = url.split('tabindex="-1"')
        url = str(url[0])

        nombrePagina =str(enlace.find_all('div', class_='source'))
        nombrePagina = nombrePagina.split(">")
        nombrePagina = (nombrePagina[2])
        nombrePagina = nombrePagina.replace('</a><span><span class="news-separator', '')

        titulo = str(enlace.find_all('a',class_="title"))
        titulo = titulo.split('blank">')
        titulo = str(titulo[1])
        titulo = titulo.replace('</a>]', '')

        mensaje = str(enlace.find_all('div', class_='snippet'))
        mensaje= mensaje.split('')
        mensaje= str(mensaje[3])

        fecha=str(enlace.find_all('span'))
        fecha=fecha.split('tabindex="0">')
        fecha=str(fecha[1])
        fecha=fecha.replace('</span>]', '')

    alcaldeNoticias = AlcaldeNoticias(url =url,nombre_Pagina_WEB=nombrePagina,titulo = titulo, mensaje = mensaje, fecha =
    browsersB.alcalde.connect(alcaldeNoticias)
```




Examen Practico

27/05/2021

```
In [6]: import requests
from bs4 import BeautifulSoup
import json

#Palabras a Buscar

#URL buscar en noticias
urlfacebookAlcalde = 'https://www.google.com/search?q=https://www.facebook.com/alcaldiadepute&start='
#Rango de paginas a buscar por Noticias
rangourlfacebookAlcalde = 8

#Encabezados HTTP Para navegadores
header = {'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/80.0.3987.16'}

cont=0
contID = 1;
contCambioPaginaPartidoA=30
for i in range(rangourlfacebookAlcalde):
    url = urlfacebookAlcalde+str(contCambioPaginaPartidoA)
    respuesta = requests.get(url, headers = header )

    contenido = BeautifulSoup(respuesta.content, "html.parser")

    contenido = contenido.find_all('div', class_="g")
    contCambioPaginaPartidoA=contCambioPaginaPartidoA+10
    cont=cont
    for lista in contenido:
        cont=cont+1

        #SACAR URL
        url = str(lista.find_all('a'))
        url = url.split('href="')
        url = str(url[1])
        url = url.split('" ping="')
        url = str(url[0])

        #SACAR TITULO DE LA PAGINA WEB
        titulo =str(lista.find_all('div', class_='TbwUpd NJjxre'))
        titulo = titulo.split('tjvcx">')
        titulo = (titulo[1])
        titulo = titulo.split('<span')
        titulo = (titulo[0])

        #SACAR NOMBRE DE LA PAGINA WEB
        nombrePagina=str(lista.find_all('h3', class_='LC20lb DKV0Md'))
        nombrePagina = nombrePagina.split('<span>')
        nombrePagina = (nombrePagina[0])
        nombrePagina = nombrePagina.split('</span>')
        nombrePagina = (nombrePagina[0])

        #SACAR MENSAJE
        mensaje = str(lista.find_all('span', class_='aCOpRe'))
        mensaje = mensaje.split('<span>')
        mensaje = (mensaje[0])
        mensaje=mensaje.replace('</span></span>]', '')
        mensaje=mensaje.replace('<em>', '')
        mensaje=mensaje.replace('</em>', '')

        alcaldeFacebook = AlcaldeFacebook(url =url,nombre_Pagina_WEB=nombrePagina,titulo = titulo, mensaje = mensaje).save()
        browsersG.alcaldeFacebook.connect(alcaldeFacebook)

    contID = contID + 1
```



Examen Practico

27/05/2021

```
In [7]: import requests
from bs4 import BeautifulSoup
import json

#Palabras a Buscar

#URL buscar en noticias del alcalde
noticiasalcalde = 'https://www.google.com/search?q=Gustavo+Vera+Arizaga+alcalde+de+gualaceo&tbm=nws&sxsrf='
#Rango de paginas a buscar por Noticias
rangoNoticiasAlcalde = 60

#Encabezados HTTP Para navegadores
header = {'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/80.0.3987.16'}

cont=0
contID = 1;
contCambioPaginaP=0
for i in range(rangoNoticiasAlcalde):
    url = noticiasalcalde+str(contCambioPaginaP)
    respuesta = requests.get(url, headers = header )
    contenido = BeautifulSoup(respuesta.content, "html.parser")
    contenido = contenido.find_all('div', class_='dsr')

    contCambioPaginaP=contCambioPaginaP+10
    cont=cont
    for lista in contenido:
        cont=cont+1

        #SACAR URL
        url = str(lista.find_all('a'))
        url = url.split('href="')
        url = str(url[1])
        url = url.split('" ping="')
        url = str(url[0])

        #SACAR NOMBRE DE LA PAGINA WEB
        nombreP =str(lista.find_all('div', class_='XTjFC WF4Cuc'))
        nombrePagina = nombreP.split('</g-img>')
        nombrePagina = (nombrePagina[1])
        nombrePagina = nombrePagina.replace('</div>]', '')

        #SACAR TITULO
        titulo=str(lista.find_all('div', class_='JheGif nDgy9d'))
        titulo=titulo.replace('["<div aria-level="2" class="JheGif nDgy9d" role="heading" style="-webkit-line-clamp:2">', '')
        titulo=titulo.replace('</div>]', '')

        #SACAR MENSAJE
        mensaje = str(lista.find_all('div', class_='Y3v8qd'))
        mensaje=mensaje.replace('["<div class="Y3v8qd">', '')
        mensaje=mensaje.replace('</div>]', '')

        #SACAR FECHA
        fecha=str(lista.find_all('span',class_='WG9SHc'))
        fecha=fecha.replace('["<span class="WG9SHc"><span>', '')
        fecha=fecha.replace('</span></span>]', '')

        #Guardar Los datos
        alcaldeNoticias = AlcaldeNoticias(url =url,nombre_Pagina_WEB=nombrePagina,titulo = titulo, mensaje = mensaje, fecha = fecha)

        #Guardar Los datos del object browsersG
        browsersG.alcalde.connect(alcaldeNoticias)
        contID = contID + 1
```

```
#SACAR NOMBRE DE LA PAGINA WEB
nombreP =str(lista.find_all('div', class_='XTjFC WF4Cuc'))
nombrePagina = nombreP.split('</g-img>')
nombrePagina = (nombrePagina[1])
nombrePagina = nombrePagina.replace('</div>]', '')

#SACAR TITULO
titulo=str(lista.find_all('div', class_='JheGif nDgy9d'))
titulo=titulo.replace('["<div aria-level="2" class="JheGif nDgy9d" role="heading" style="-webkit-line-clamp:2">', '')
titulo=titulo.replace('</div>]', '')

#SACAR MENSAJE
mensaje = str(lista.find_all('div', class_='Y3v8qd'))
mensaje=mensaje.replace('["<div class="Y3v8qd">', '')
mensaje=mensaje.replace('</div>]', '')

#SACAR FECHA
fecha=str(lista.find_all('span',class_='WG9SHc'))
fecha=fecha.replace('["<span class="WG9SHc"><span>', '')
fecha=fecha.replace('</span></span>]', '')

#Guardar Los datos
alcaldeNoticias = AlcaldeNoticias(url =url,nombre_Pagina_WEB=nombrePagina,titulo = titulo, mensaje = mensaje, fecha = fecha)

#Guardar Los datos del object browsersG
browsersG.alcalde.connect(alcaldeNoticias)
contID = contID + 1
```



Examen Practico

27/05/2021

```
In [8]: import requests
        from bs4 import BeautifulSoup
        import json

        #Palabras a Buscar

        urlfacebookAlcalde = 'https://www.google.com/search?q=https://www.facebook.com/gustavoveraalcalde&start='
        #Rango de paginas a buscar por Noticias
        rangourlfacebookAlcalde = 8

        #Encabezados HTTP Para navegadores
        header = {'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/80.0.3987.16'}

        cont=0
        contID = 1;
        contCambioPaginaPartidoA=30
        for i in range(rangourlfacebookAlcalde):
            url = urlfacebookAlcalde+str(contCambioPaginaPartidoA)
            respuesta = requests.get(url, headers = header )

            contenido = BeautifulSoup(respuesta.content, "html.parser")

            contenido = contenido.find_all('div', class_='g')
            contCambioPaginaPartidoA=contCambioPaginaPartidoA+10
            cont=cont
            for lista in contenido:
                cont=cont+1

                #SACAR URL
                url = str(lista.find_all('a'))
                url = url.split('href="')
                url = str(url[1])
                url = url.split('" ping="')
                url = str(url[0])

                #SACAR TITULO DE LA PAGINA WEB
                titulo =str(lista.find_all('div', class_='TbwUpd NJjxre'))
                titulo = titulo.split('tjvcx">')
                titulo = (titulo[1])
                titulo = titulo.split('<span')
                titulo = (titulo[0])

                #SACAR NOMBRE DE LA PAGINA WEB
                nombrePagina=str(lista.find_all('h3', class_='LC201b DKV0Md'))
                nombrePagina = nombrePagina.split('<span>')
                nombrePagina = (nombrePagina[0])
                nombrePagina = nombrePagina.split('</span>')
                nombrePagina = (nombrePagina[0])

                #SACAR MENSAJE
                mensaje = str(lista.find_all('span', class_='aCopRe'))
                mensaje = mensaje.split('<span>')
                mensaje = (mensaje[0])
                mensaje=mensaje.replace('</span></span>]', '')
                mensaje=mensaje.replace('<em>', '')
                mensaje=mensaje.replace('</em>', '')

                alcaldeFacebook = AlcaldeFacebook(url =url,nombre_Pagina_WEB=nombrePagina,titulo = titulo, mensaje = mensaje).save()
                browsersG.alcaldeFacebook.connect(alcaldeFacebook)

            contID = contID + 1
```



Examen Practico

27/05/2021

neo4j@bolt://localhost:11005/neo4j - Neo4j Browser

File Edit View Window Help Developer

Database Information

Use database

neo4j - default

Node Labels

(1,382) AlcaldeFacebook

AlcaldeNoticias Browsers

Relationship Types

(1,379) BROWSERS ALCALDE

BROWSERS ALCALDE FACEBOOK

Property Keys

fecha mensaje nombre

nombre_Pagina_WEB titulo url

Connected as

Username: neo4j

Roles: admin, PUBLIC

Admin: server user list

server user add

Disconnect: server disconnect

DBMS

Version: 4.2.1

neo4j\$

neo4j\$ match(n) return count(n)

	count(n)
1	1382

Started streaming 1 records after 2 ms and completed after 3 ms.

\$:play start

Getting started with Neo4j Browser

Neo4j Browser user interface guide

Get started

Try Neo4j with live data

A complete example graph that demonstrates common query patterns.

Actors & movies in cross-referenced pop culture.

Open guide

Cypher basics

Intro to Graphs with Cypher

What is a graph database?

How can I query a graph?

Start querying



Inteligencia Artificial 1.



Examen Practico

27/05/2021



Examen Practico


Details

Plugins

Upgrade

ExamenInt

Click to add description

Version	4.2.1
Edition	enterprise
Status	Active
Labels	3
Nodes	1382
Relationship Types	2
Relationships	1379
DBMS ID	database-1f05fbbe-c1a9-4793-8dc3-f40ff4c8505c
Property Keys	6
IP address	localhost
Bolt port	11005 



Examen Practico

