ACTIVITAT
Objectius: - Obtenir i processar dades a partir de documents XML
Instruccions: - Es tracta d'un treball en grups de dos - Responeu a l'espai de cada pregunta, si ho feu amb diapositives enganxeu la diapositiva en aquest mateix espai. - Es valorarà la cura en la presentació del document i que segueixi l'estructura indicada.
Criteris d'avaluació: - Cada pregunta té el mateix pes - Es valorarà la presentació i els comentaris al codi
Entrega: - Aquest document anomenat **memoria.pdf** i el codi corresponent

Repositori de referència: https://github.com/jpala4-ieti/DAM2-MP06-UF03-Base

Noms i Cognoms: Joel Berzal Álamo
Repositori GitHub amb exercicis resolts:
https://github.com/joelberzalgithub/AMS2-MP06-PR3.1-BerzalJoel-ChicaAlex

Preparació de l'activitat

Escull un dels dumps complets amb el tema que més t'interessi entre els disponibles a Stack Exchange i crea una nova base de dades amb BaseX https://archive.org/details/stackexchange

En aquest enllaç trobareu alguns fitxers descarregats:
https://drive.google.com/drive/folders/1uasdYDtDLIju3qfaDERGSkrbvXOSXbrf?usp=sharing

Aquesta llista et pot ajudar a escollir el tema
https://data.stackexchange.com/

IMPORTANT: Per alguns temes Chrome informa que el fitxer conté virus. Podeu usar Firefox si us interessa un tema i esteu d'acord amb els següents riscos.

EXPLICACIÓ DELS RISCOS

BaseX en si està dissenyat per processar dades XML i no executa scripts incrustats ni codi dins del contingut XML durant el procés d'importació. El risc principal associat amb la importació de fitxers XML no és l'execució d'un virus durant el procés de càrrega, sinó més aviat el potencial per a contingut maliciosament elaborat per explotar vulnerabilitats en l'intèrpret d'XML o l'aplicació que processa les dades XML.

Tanmateix, hi ha alguns riscos potencials a considerar:

- **Injecció d'Entitat Externa XML (XXE):** Aquest és un tipus d'atac contra una aplicació que analitza entrada XML. Si l'intèrpret d'XML està configurat per resoldre entitats externes, un atacant pot crear un document XML que faci referència a recursos externs, portant a l'accés no autoritzat a sistemes interns, divulgació de fitxers i atacs de denegació de servei.

- **Atac Billion Laughs:** Això és un atac de denegació de servei que té com a objectiu els intèrprets d'XML. Creant un petit document XML amb una gran quantitat d'entitats niades, un atacant pot esgotar els recursos del sistema.

- **Càrregues Malicioses:** Encara que BaseX per si mateix no executarà un script contingut dins d'un document XML, els scripts o càrregues malicioses podrien ser incrustats dins de l'XML i executats per altres components de l'aplicació que processa les dades XML després que s'han importat.

Exercicis

Exercici 1. Consultes amb xquery i xpath (4 punts)

Quines són les XQuery que permeten resoldre les següents preguntes?

- Quins són els títols de les preguntes que han estat més vistes (ViewCount)? Inclou títol i nombre de vistes en ordre descendent.

- Quins són els usuaris que han realitzat més preguntes? Inclou nom de l'usuari i el nombre de preguntes realitzades en ordre descendent.

- Quins són els tags més usats? Inclou nom del tag i nombre d'usos en ordre descendent.

- Pels 10 tags més usats, retorna les 100 preguntes amb més vistes que en contenen algun.

- Quins són
    - Els títols i els cossos de les preguntes amb més puntuació (Score)
    - La resposta més votada a la pregunta.
    - Cal incloure número de vots de la pregunta
    - Inclou tags

Cada consulta que respon a un dels punts anteriors s'ha de guardar dins el directori **./data/input** en fitxers amb .xquery.

Els resultats generats per les consultes han de ser XML. Exemple:

xquery:

```
declare option output:method "xml";
declare option output:indent "yes";

<posts>{
  for $p in /posts/row[@PostTypeId='1']
  return <title>{$p/@Title/string()}</title>
}</posts>
```

retorna:

```
<posts>
  <title>What is considered surprise</title>
  <title>How does DeGray's Pilebunker chip work when trashing Gems of size 2+?</title>
…
```

Exercici 2. Processat de les consultes amb java i tractament de resultats (3 punts)

Entrada: cadascun dels fitxers .xquery dins el directori **./data/input**

Què és demana:

- Fes una programa java que llegeixi cadascun dels fitxers
- Executi la consulta contra el basexserver
- Guardi el resultat com a fitxer .xml dins el directori **./data/output**

Exercici 3. Tractament de títols processats amb llenguatge natural (3 punts)

Preparació: Executa **DescarregarModelsMain.java** per descarregar automàticament els models necessaris dins del directori **./data/models**

Exemple: Mira't l'exemple public **ExecutarExempleNLPMain.java.**

Tasca:
Processa els títols i els body de les preguntes amb més vistes per extreure determinats noms propis..

Sortida: Emmagatzema el resultat en el fitxer **./data/noms_propis.txt**
Important: Si s'identifiquen dues entitats del mateix tipus seguides es consideren com una sola:

Explicació del processat de llenguatge natural (NLP) i extracció de noms propis

La frase que es processa és: "John Doe, a software engineer at Google, recently visited New York City. He said, \"It's an amazing place!\" The trip made him feel very happy."

Part de la sortida mostra el següent:
Entity Detected: John - Entity Type: PERSON
Entity Detected: Doe - Entity Type: PERSON
Entity Detected: software - Entity Type: TITLE
Entity Detected: engineer - Entity Type: TITLE
Entity Detected: Google - Entity Type: ORGANIZATION
Entity Detected: recently - Entity Type: DATE
Entity Detected: New - Entity Type: CITY
Entity Detected: York - Entity Type: CITY
Entity Detected: City - Entity Type: CITY

Aclariment sobre la sortida:
Entity Detected: John - Entity Type: PERSON
Entity Detected: Doe - Entity Type: PERSON

Cal identificar-lo com a John Doe, PERSON en una única línea del fitxer **./data/noms_propis.txt**