Practica 2

Infraestructura UPM para publicaciones científicas

El Archivo Digital de la UPM (https://oa.upm.es) gestiona actualmente los trabajos fin de grado, fin de tesis y tesis doctorales de los alumnos de la Universidad. En la siguiente versión de la plataforma se quiere gestionar también las **publicaciones científicas** de sus investigadores.

Es un desafío porque el volumen de datos que se ha de soportar es mucho mayor que con los TFGs, TFMs, y tesis. Por este motivo se ha solicitado a los alumnos de IBD la creación de una **infraestructura Big Data** que soporte las funcionalidades que esperan ofrecer desde su portal.

En concreto, quieren ofrecer **datos estadísticos** sobre los autores de las publicaciones, sus colaboraciones, las áreas de investigación, y además facilitar la **exploración** de su contenido y la **búsqueda** avanzada desde su propio portal web.

Objetivo

- Diseñar una infraestructura TI ***eficiente*** que soporte la gestión **enriquecida** de publicaciones científicas en formato **PDF** y permita un ***escalado horizontal*** sobre ***commodity hardware***.

- Construir un despliegue basado en **Docker** de dicha infraestructura, capaz de levantarse en un entorno local con las siguientes características: **4 CPUs, 10 Ghz RAM, 250GB Disco** (arquitectura AMD o ARM), que permita evaluar la infraestructura.

- Demostrar su **eficacia** sobre **conjuntos de publicaciones** para soportar **análisis estadísticos** y **consultas,** ya sean **simples** o **complejas**, destacando además sus **virtudes**, **limitaciones** y la capacidad de **extensión** o **modificación**.

Entrega

- Fichero **.zip** creado mediante una *'release'* de un repositorio en GitHub.
  - Su nombre contiene la versión del servicio.
  - Contenido:
    - Archivo/s **Docker** necesarios para el despliegue de la infraestructura.
    - Archivo **README.md** con la descripción del entorno y las instrucciones de despliegue y uso.
    - Archivo **LICENSE** asociado a la licencia de distribución.
    - Archivo **Repository.md** con la ruta completa al repositorio en GitHub

- **Deadline**:
  - 12/05/2023 23:59h CEST

Demo

- Despliegue de la infraestructura a partir del **.zip entregado** en un **entorno virtualizado basado en Docker** proporcionado por el profesor
  - Disponible ARM (Mac) y AMD (Windows/Linux).
- Presentación en **grupo** del diseño y alcance de la infraestructura
  - las preguntas se responderán de forma individual por el alumno que indique el profesor
- Validación del funcionamiento a partir de uno o varios conjuntos de datos de **test**
  - Proporcionado por el profesor en el momento de la demo
- **Duración** de la exposición y demo
  - entre 15min-20 min
- **Fecha:**
  - 18/05/2023 9:00h

Evaluación

- Uso de las siguientes **publicaciones** durante el desarrollo de la práctica.
  - Artículos aceptados en la conferencia SEPLN 2022: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/issue/view/286
  - Se recomienda el uso de **'Grobid'** para parsear los documentos.

- Generación de los siguientes **datos estáticos**:
  - **Documents.csv**: contiene meta-información de las publicaciones
    - Columnas: 'file_name', 'title', 'num_pages', 'creation_date', 'modification_date'
  - **Authors.csv**: contiene el número de publicaciones por autor
    - Columnas: 'author' y 'publications'

- Generación de los siguientes **datos dinámicos**:
  - **Keywords.csv**: contiene el número de apariciones de un término concreto (e.g. 'virus'), o cualquiera de sus sinónimos en inglés (e.g. 'infection', 'microbe'..), en el corpus
    - Columnas: 'word' y 'frequency'

- Soporte para las siguientes **consultas simples**:
  - **Articles**: listado ordenado de artículos en los que un *autor específico* ha participado.
    - La relevancia viene determinada por el número de autores (menor número de autores, mayor relevancia del autor concreto)
  - **Texts**: listado ordenado de párrafos, junto con el título del artículo al que pertenecen, que contienen un *término específico*.
    - La relevancia viene determinada por el tamaño del párrafo y la frecuencia del término.

- Soporte para las siguientes **consultas complejas**:
  - **Collaborators**: listado ordenado de autores relacionados con un *autor específico*.
    - La relación entre autores viene determinada por su colaboración directa en un artículo o indirecta a través de autores comunes
  - **Words**: número de palabras en el corpus cuya longitud es de un *tamaño específico*.
    - Palabras con sólo una letra, o con dos letras, o con tres letras...hasta 20 letras.