**Ranking Retrieval with PostgreSQL**

Profesor Heider Sanchez

El objetivo de este laboratorio es poner a prueba las técnicas de indexación de textos en PostgreSQL (full-text search index) mediante tres experimentos.

**P1. Sequential Scan vs GIN:**

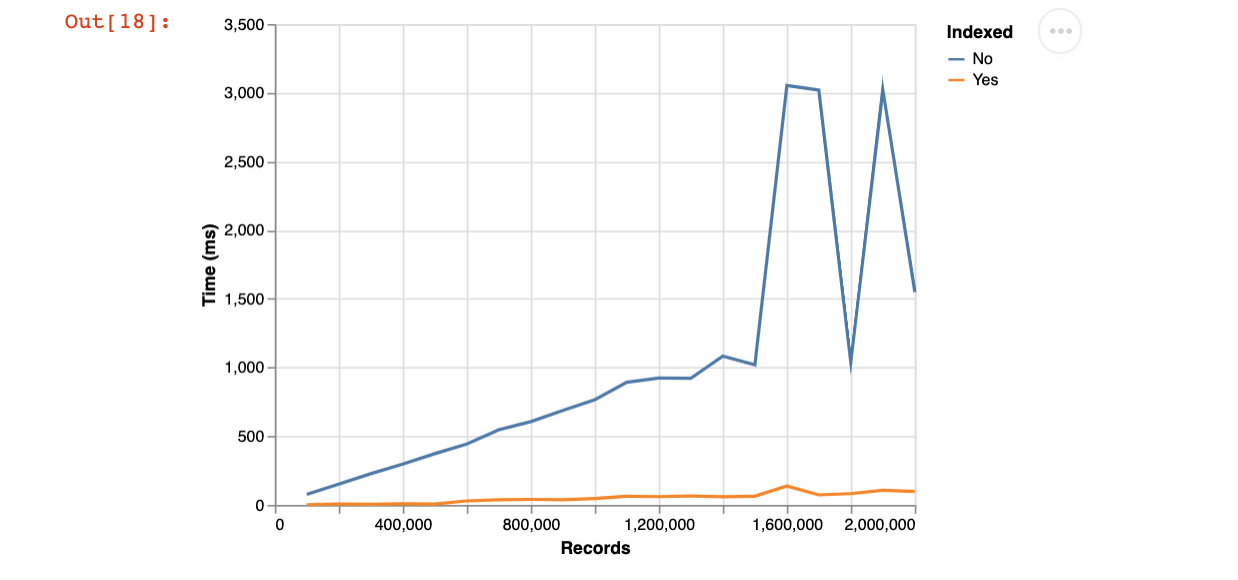
El primer experimento consiste en probar el índice invertido GIN representando el texto con **trigramas**. Un trigrama es un grupo de tres caracteres consecutivos tomados de una cadena. Ejemplo, los trigramas de la palabra “amor” son “amo” y “mor”. Indexar un atributo tipo texto con trigramas es eficaz en la mayoría de lenguajes naturales mejorando considerablemente las búsquedas textuales.

https://www.postgresql.org/docs/13/pgtrgm.html

Tomando como base el script dato en clase, se le pide realizar lo siguiente:

* Crear una tabla con dos atributos textuales, uno sin indexar y el otro indexado.
* Llenar datos aleatorios para diferentes cantidades.
* Ejecutar consultas sobre ambos atributos y tomar los tiempos

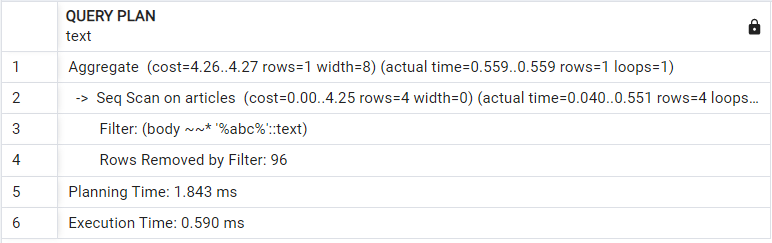
**Mostrar el plan de ejecución y un gráfico como resultado de la experimentación (ver gráfico de referencia).**

****

**Resultado**

No Indexed:

* 100



* 1000

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

* 10 000

Texto, Aplicación

Descripción generada automáticamente

* 100 000

Imagen que contiene Aplicación

Descripción generada automáticamente

* 1 000 000

Imagen que contiene Aplicación

Descripción generada automáticamente

* 10 000 000

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Indexed:

* 100

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

* 1000

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

* 10 000

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

* 100 000

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

* 1 000 000

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

* 10 000 000

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

**Gráfico, Gráfico de líneas

Descripción generada automáticamente**

**P2. Full-text search on Films**

El segundo experimento consiste en aplicar el índice invertido GIN sobre los atributos textuales de la tabla “film” ([dvdrental](https://www.postgresqltutorial.com/postgresql-sample-database/)).

* Restaurar la base de datos en su servidor PostgreSQL
* Crear un nuevo atributo indexado compuesto por el titulo y la descripción de la película.
  + El tipo de dato corresponde al vector de pesos de los términos
* Ejecutar consultas sobre los atributos sin indexar y sobre el atributo indexado
  + Tomar los tiempos para diferentes rankings (top k)

**Mostrar el plan de ejecución y un gráfico como resultado de la experimentación**

No Indexed:

-- 32.352 ms

explain analyze

select title, description

from film

where   (title || description) ilike '%man%'

        or (title || description) ilike '%woman%';

Indexed:

-- 0.331 ms

explain analyze

select title, description

from film

where to\_tsquery('english', 'Man | Woman') @@ full\_text;

Indexed (top rank):

set enable\_seqscan = false;

explain analyze

select title, description, ts\_rank\_cd(full\_text, query) as rank

from film, to\_tsquery('english', 'man | woman') query

where query @@ full\_text

order by rank desc

limit 64;

* Limit 1:

**Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente**

* Limit 2:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

* Limit 4:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

* Limit 8:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

* Limit 16:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

* Limit 32:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

* Limit 64:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Gráfico, Gráfico de líneas

Descripción generada automáticamente

≈ Constante

**P3. Full-text search on News**

El tercer experimento consiste en aplicar el índice invertido GIN sobre los atributos textuales de la tabla “articles” ([all the news](https://www.kaggle.com/snapcrack/all-the-news)).

* Crear la tabla Articles y llenar los datos desde los archivos CSV
* Crear un nuevo atributo indexado compuesto por el titulo y el contenido de la noticia.
  + El tipo de dato corresponde al vector de pesos de los términos
* Ejecutar consultas sobre los atributos sin indexar y sobre el atributo indexado
  + Tomar los tiempos para diferentes rankings (top k)

**Mostrar el plan de ejecución y un gráfico como resultado de la experimentación**