Ejercicios a realizar con ElasticSearch

Notas previas

- Todos los ejercicios deben resolverse usando el lenguaje de programación Python, atacando un servidor ElasticSearch donde se habrá indexado la colección usada durante las prácticas de laboratorio.
- La única excepción a esta norma serán aquellos casos en los que pudiera resolverse el ejercicio con **una única consulta** enviada mediante el interfaz de Cerebro.
- Las prácticas se realizarán en equipos de 3 personas.
- Se entregará un archivo comprimido con el código desarrollado, los ficheros de salida en caso de que se soliciten y un documento describiendo el modo en que se afrontó cada ejercicio y cómo se usaron las diferentes funcionalidades de ElasticSearch para resolver cada problema. En caso de que algún ejercicio haya involucrado el uso de Cerebro se indicará URL, método HTTP utilizado y carga JSON de cada petición.

Primer ejercicio (2.5 puntos)

Obtener un archivo con todos las publicaciones (incluyendo autor, fecha de creación y texto; puede usarse formato TSV o JSON) relativas a una temática especificada mediante una consulta.

Dicho archivo debe ser lo más exhaustivo posible por lo que la consulta inicial deberá ser expandida; para ello se deberá usar la funcionalidad de términos significativos en agregaciones y, muy importante, también se deberán eliminar palabras vacías[1].

Por ejemplo, si la temática escogida fuera alcoholismo se emplearía como keyword inicial *alcoholism*, y se obtendrían términos asociados como *drinking*, *alcoholic*, o *drunk* (que son relevantes) pero también otros como *blamey*, *collaborative* o *foray* (que no lo son tanto).

Queda a elección del estudiante la elección de la temática así como el número máximo de términos asociados a utilizar. Es obligatorio usar varias métricas de similitud. Debe razonarse la elección de la consulta base para la temática escogida, además de reflexionar críticamente sobre los resultados obtenidos con distintas métricas y distintos números de términos.

[1] http://snowball.tartarus.org/algorithms/english/stop.txt

Se valorará muy positivamente la realización de una evaluación sistemática de los resultados, informando sobre la precisión (porcentaje de documentos relevantes) alcanzada con cada configuración.

Segundo ejercicio (2 puntos)

Estudiar la documentación de las consultas *"More Like This"*. ¿Existe alguna configuración de dicha consulta o su generación vía scripting que permita emular la expansión de términos mediante NGD (parámetro gnd en las agregaciones) hecha en el ejercicio anterior?

Tercer ejercicio (2.5 puntos)

Se desea obtener una lista exhaustiva de los medicamentos utilizados por los usuarios que aparecen en la colección.

Describir de manera detallada los pasos seguidos para explorar la colección, cómo se ha construído la consulta o consultas que han permitido llegar a una lista potencial y qué recursos externos se han empleado para validar la información obtenida.

¡Atención! No es necesaria la validación automática de la lista de medicamentos obtenida pero sí se valorará positivamente que se explore qué posibilidades de automatización existen al respecto.

Si se usa algún conocimiento experto de partida debe indicarse claramente en la documentación.

Cuarto ejercicio (3 puntos)

La comorbilidad es un término médico que hace referencia a dos conceptos: (1) la presencia de uno o más trastornos (o enfermedades) además del trastorno (o enfermedad) primaria, y (2) el efecto de dichos trastornos o enfermedades adicionales.

Se desea obtener una lista con el mayor número posible de factores comórbidos relativos a dos problemáticas diferentes: la ideación suicida y las conductas autolesivas.

En este ejercicio **no se puede utilizar ningún conocimiento experto como punto de partida**:

- En el caso del suicidio sólo se podrán usar (como punto de inicio) los siguientes términos o frases: *suicide*, *suicidal*, *"kill myself"*, *"killing myself"*, o *"end my life"*.
- En el caso de las conductas autolesivas se podrá usar la frase *"self harm"*.

Para resolver este ejercicio es probable que se tengan que generar subconjuntos de documentos y hacer uso de agregaciones; por otro lado, se recuerda que además del campo

`selftext` se dispone de otros potencialmente interesantes como `title`, `author` y `subreddit`.

En la evaluación de este ejercicio se valorará muy positivamente la sistematización del proceso de exploración mediante el uso de scripts.