# Trabajo práctico

Recuperación de Información usando Elasticsearch

## Parte 1: Recuperación de Información

Completar los ejercicios de esta notebook: https://bit.ly/TP_representación_Diplo2020
La entrega debe hacerse sobre la misma notebook. Usar markdown para comentar la celda anterior a la celda del código, indicando la funcionalidad provista y cualquier asunción realizada. **No eliminar el *output* generado por cada celda.**

## Parte 2: Elasticsearch

Utilizando el archivo G20 provisto durante la clase (https://bit.ly/g20_Diplo2020), indexar los tweets del archivo, asegurándose de definir el mapping de forma correcta para los campos de interés. Notar que el mapping correcto no sólo depende del tipo de datos, sino de las consultas que deben hacerse sobre los mismos. De ser necesario o conveniente, se podría elegir indexar un campo de más de una forma.

Los campos de interés son:

- “text”, “created_at”, “id”, “user.location”, “user.followers_count”, “place.bounding_box”, “source”, “entities.hashtags”, “timestamp_ms”, “retweeted”

Las consultas de interés son:

1. Poder buscar teniendo en cuenta el patrón de mayúsculas o minúsculas, por ejemplo: “President” o “PRESIDENT” en el campo ‘text’, solo debe devolver documentos que tengan la palabra escrita con el mismo patrón de mayúsculas y minúsculas.
2. Poder buscar por palabras que compartan una misma raíz, por ejemplo: “pray”, haría un match con tweets que contengan: “prays”, “prayer”, “praying”, “prayers” en el campo ‘text’.
3. Consulta para poder buscar una ubicación en el campo ‘user.location’ y que el matching sea exacto, por ejemplo para las consultas “York” y “California”, **no debería** devolver documentos en donde user.location sea “New York” o “LA, California” por ejemplo.
4. Consulta para poder buscar tweets cuyos usuarios hayan abierto sus cuentas en un rango de fechas.
5. Consulta para poder buscar tweets que hayan sido posteados en un área específica, por ejemplo, en el área de “Washington DC” usando el campo “place.bounding_box”.

Para esto deberán buscar las coordenadas del sitio en cuestión en algún sitio web (como por ejemplo Wikipedia).

Para esta segunda parte adjuntar un documento (.pdf, .ipynb o archivo de texto) en donde se muestre el mapping usado, las consultas realizadas y el resultado de las consultas.

Las entregas deben hacerse en grupos entre 2 y 5 personas. Usar este link para indicar los grupos https://bit.ly/grupos_Diplo2020 .

Usar el siguiente link para el envío de los archivos: https://www.dropbox.com/request/WBWPV09uuoS3oVwMIocJ
Al mismo tiempo enviar un email a axel.soto@cs.uns.edu.ar (sin adjuntos) para indicar que hicieron la entrega. Incluir en la entrega los datos de todos los integrantes del grupo

**Fecha de entrega: 15 de Febrero**