Práctica final. II Parte de Minería de Datos y Texto.

MADM. Curso 2021-22

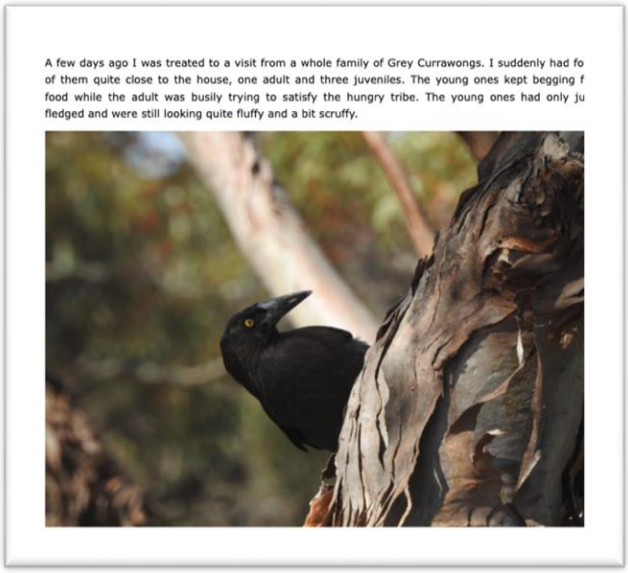
El objetivo de esta práctica es demostrar la adquisición de los conceptos y el uso de herramientas tratados en la segunda parte de esta asignatura.

Contexto:
El avistamiento de aves (*birding*) es una actividad para el estudio de aves silvestres. Muchos aficionados registran con notas sus avistamientos y algunos de estos aficionados comparten sus notas y experiencias en blogs. Imaginad el potencial de digitalizar dichas observaciones para ubicar aves, sus características, metainformación de la observación como la época del año, características del ecosistema, etc.

Ejemplos de blogs:
https://www.shorebirder.com/
https://www.trevorsbirding.com/
https://dantallmansbirdblog.blogspot.com/

Objetivo:
A través de un proceso de tratamiento de texto hay que identificar las aves que los observadores han identificado y registrado en sus blogs.
El **resultado** será una ontología donde queden registradas las aves identificadas, pero estas aves han de estar vinculadas con la Wikipedia (Dbpedia). Así, podríamos tener más información experta de las observaciones.

Por ejemplo, del texto extraído del blog de: https://www.trevorsbirding.com/ (adjunto imagen) el resultado mínimo sería tener un individual de la clase https://es.dbpedia.org/page/Strepera_fuliginosa.



https://es.wikipedia.org/wiki/Strepera_fuliginosa

- La información contextual gracias a ese hecho -de la DBpedia- es que dicha ave es un ave endémica de Australia (category-es:Aves_endémicas_de_Australia) e incluso podríamos inferir con la ontología que dicha persona ha visitado el continente Australiano. En cualquier caso, toda esta información contextual no la usaremos. La práctica se centrará en obtener y etiquetar dicha información para que en futuros trabajos la manipulación de los datos sea más eficiente como por ejemplo el descubrimiento de información.

Proceso (e *Ideas*)

- **Fuente de datos:** Lo ideal sería realizar procesos de *crawling* para automatizar la obtención de texto de esos blogs. Como es un proceso relativamente costoso y no lo hemos tratado, simplificamos este punto y al mismo tiempo, la envergadura de esta propuesta.
  Mi propuesta para este punto es la de copiar y guardar diferentes entradas de los blogs en diferentes ficheros de texto. **Un mínimo de 10 entradas.** No serán válidas aquellas entradas en las que vosotros filtráis el texto de interés.
  Ejemplo de entrada válida usando el ejemplo anterior:

  A family visit

  A few days ago I was treated to a visit from a whole family of Grey Currawongs. I suddenly had four of them quite close to the house, one adult and three juveniles. The young ones kept begging for food while the adult was busily trying to satisfy the hungry tribe. The young ones had only just fledged and were still looking quite fluffy and a bit scruffy.

- **Tratamiento:** Pensad que procesos y que librerías necesitáis utilizar para identificar conceptos relevantes a partir de esos ficheros de texto.

- ***Matching*:** Existe un proceso crítico que consiste en equiparar texto relevante con nuestro dataset (la Wikipedia). Pensad en como podéis realizarlo, reducir la carga o el número de consultas que hacéis sobre la Wikipedia/DBpedia para que no sea "lento/pesado".

Entregable:

- **Informe** que contenga una explicación del proceso, las suposiciones tomadas, una valoración crítica sobre la influencia de las suposiciones en la calidad del proceso
- **Código del proyecto.** Realmente, una url del repositorio donde está publicado.
- **Ontología** poblada con los "datos" identificados

Opcional:

- Se valorará positivamente la asociación de más hechos en la ontología (fechas, lugares, etc.)