Inycom
innovation technologies

UCode

Descripción del reto

El reto propuesto por Inycom toca las temáticas de Big Data y Procesamiento del Lenguaje Natural (PLN). Consiste principalmente en enseñarnos qué sois capaces de hacer cuando disponéis de una gigantesca cantidad de datos. Demostrar que valor añadido sois capaces de darles más allá de mostrarlos simplemente de una manera bonita o vistosa. Tomando esto como premisa, podéis realizar cualquier cosa que se os ocurra.

No hay límite en cuanto a lo que podéis hacer con ellos: obtener datos agregados, cruzar distintos datasets, añadir todavía más información (redes sociales), utilizarlos para hacer test automáticos, etc. La forma de leerlos, tratarlos y almacenarlos también es libre; aunque utilizar las herramientas adecuadas puede facilitaros mucho la tarea ;). Tampoco estáis restringidos/as a utilizar únicamente los datos que os ofrecemos. Existen infinidad datasets de acceso libre disponibles en internet (algunos de los cuales se encuentran listados más abajo).

Estos datos pueden no estar estructurados, provenir de fuentes diversas y/o no contener información útil en un primer vistazo. Sin embargo, utilizando técnicas de Big Data y Data Mining se puede llegar a obtener información muy valiosa a raíz de ellos. Los datos que contienen textos en lenguaje natural (tweets, noticias, comentarios, etc.) suelen ser más útiles si se añade una etapa de preprocesado en la que se suprime el contenido inútil o redundante. Aquí es donde entra en juego el PLN, segundo apartado de la introducción al reto.

En resumen, aplicad todos vuestros conocimientos e ideas combinadas (o no) con la ayuda que os facilitamos para conseguir encontrar algo útil dónde otros no verían más que un amasijo de letras y números. El único requisito es utilizar datos. Muchos datos.

Datasets

Os facilitamos varios conjuntos de datos que podéis explotar y exprimir para la realización de vuestro proyecto. No obstante, sentíos libres de utilizar cualquier otro dataset libre disponible por internet o incluso recopilado por vosotros/as mismos/as para trabajar únicamente con ellos o para mezclarlos/cruzarlos con otros conjuntos.

A continuación listamos una breve explicación de los datasets facilitados:

- *Partidos políticos*: recopilación de información relativa a los partidos políticos de España y su red de seguidores en Twitter. Formato SQL.

- *Sentimiento bancario*: conjunto de opiniones expresiones realizadas por usuarios en lenguaje natural (tweets, noticias, comentarios, etc.) acompañada de una evaluación de dicha opinión. Formato JSON y accesibles mediante SOLR a través de 91.121.92.24:8983/solr/ .

- *Aerolíneas*: registro de los vuelos realizados por distintas aerolíneas desde/a diversos aeropuertos junto con sus tiempos y posibles retrasos. Formato CSV.

- *Internet scans*: recopilación de diversos escaneos masivo a lo largo de todo internet que almacena información de dominios, puertos, certificados y un largo etcétera. Disponible aquí.

- *Ataques terroristas*: listado de los ataques terroristas acontecidos en todo el mundo desde 1970 a 2015 junto a una descripción de sus características y consecuencias. Disponible aquí.

- *Datos del gobierno de EE.UU.*: gran cantidad de datos facilitados por el gobierno de EE.UU. en diversas materias como agricultura, sanidad, consumo y educación. Disponible aquí.

Herramientas útiles

Un listado de herramientas que pueden seros de gran utilidad:

- Weka
- SOLR
- MongoDB
- Hadoop
- Spark
- Natural Language Toolkit (NLTK)
- Twitter API

Criterios de evaluación

La valoración de los proyectos, como la de muchas otras cosas, es un poco subjetiva y queda a la libre interpretación de los jueces. De todos modos, existen varios puntos que suelen ser fijos y en los que se basará la decisión:

- Se valorará la complejidad y la cantidad de los datos que se hayan podido extraer.

- Así mismo, también se valorará en las herramientas utilizadas y su adecuación al problema, solución propuesta y datos tratados.

- No se premiará la naturaleza de los datos utilizados (o en otras palabras: no se penalizará a quien no utilice los datasets facilitados).

En función de lo dicho en el párrafo anterior, se otorgará un *LEGO Mindstoms EV3* al equipo gandor.

Buena suerte y a “hackear”.