|  |
| --- |
| **ACTIVIDAD 2.3 – Ejercicio Teórico**  **Anteproyecto de minería de datos web** |
| Imagina que decides montar un sistema de minería de datos web basado en un motor predictivo de Machine Learning.  Este es un ejercicio teórico. Explica en una **extensión máxima de 1 página** tu idea de proyecto, intentando contestar en la medida de lo posible a cuantas puedas de las siguientes preguntas. No es necesario que escribas nada de código, pero si lo ves necesario puedes comentar algunas de las tecnologías que usarías también. |
| Nombre Alumno: Rubén Gómez Olivencia  Otros Alumnos Grupo (si hay): |
| * ¿Cuál es el objetivo de tu proyecto? * ¿Qué tipo de minería web usarás: de contenido, de utilización, de estructura? * ¿Cuáles serían tus fuentes de datos? ¿Qué web o conjunto de webs usarías para extraer datos?   + ¿Usarías una web de un dominio propio o de tu empresa, o usarías webs externas?   + ¿Qué tipo de datos quieres minar: numéricos, textos, imágenes?   + Señala la web concreta y pon ejemplos concretos de datos en esa web querrías minar   + ¿Prevés algún problema en la adquisición de los datos? * ¿Con qué frecuencia tendrías que hacer el minado de datos?   + ¿Qué cantidad de datos esperas manejar para crear tu dataset de entrenamiento?   + ¿Y una vez esté el sistema en funcionamiento? * ¿Qué tipo de sistema predictivo (de Machine Learning) propones?   + ¿Qué cosa te interesa predecir?   + ¿Vas a usar sistemas de clustering, clasificación, regresión, análisis de textos?   + ¿Cuáles crees que son las métricas adecuadas para validar tu sistema de minería web? * Más allá de este ejercicio teórico, ¿cómo de viable crees que sería tu propuesta de sistema en la vida real? |
| **Propuesta de Projecto**  A continuación, describe tu proyecto:  El objetivo del proyecto es intentar predecir quién (organización u empresa) va a recibir subvenciones públicas otorgadas por distintos organismos del Gobierno Vasco.  Actualmente existe una [API](https://www.opendata.euskadi.eus/api-granted-benefits/?api=granted-benefit/) de OpenData Euskadi donde se pueden consultar más de 758.000 subvenciones otorgadas. Si realizamos la consulta más genérica obtenemos datos (devuelve los datos en JSON) de quién ha recibido la subvención, cuándo, cuánto ha obtenido, el departamento/organización/área que la ha otorgado, y la regulación por la que se otorga. Este conjunto de datos es en parte texto (regulación), numérico (cantidades y fecha) y parte que se podría convertir a números (los departamentos/organizaciones/áreas, se podrían convertir a identificadores numéricos).  Para obtener los datos se puede hacer a través de CURL haciendo llamadas a la API. Para no saturar la API habría que realizar un bucle para obtener todos los datos, ya que el sistema hace un uso de paginación. Si por cada petición obtenemos 1000 resultados, con un bucle de 759 repeticiones obtendríamos todos los datos.  Con ello obtendríamos todas las subvenciones otorgadas hasta ahora. Habría que realizar cierta limpieza de datos (valores duplicados por estar en 2 idiomas, castellano y euskera), simplificar datos (convertir departamentos/organizaciones/áreas a números), dar el formato correcto a cada campo... Tras esto, por cada subvención otorgada obtendríamos unos 8 datos, sin contar el texto completo de la propia subvención (para lo que habría que realizar un análisis de texto). Debido a que los datos no se actualizan de manera diaria, los datos nuevos se podrían obtener una vez al mes (existen llamadas a la API para obtener sólo subvenciones entre fechas).  Para el entrenamiento se contaría con algo más de 758.000 subvenciones. Para comprobar la eficiencia del dataset, se podría partir y entrenar con el 80% de los datos y luego realizar pruebas con el 20% restante, para ver la eficiencia obtenida.  La idea es predecir a qué empresa se le va a adjudicar una subvención teniendo en cuenta el histórico que tenemos. Para ello, con los datos analizados podemos hacer uso del sistema de clusterización, y de esta manera, para próximas subvenciones tratar de predecir en qué clúster encajaría.  Para finalizar, es interesante tratar de definir si el proyecto es viable. Debido a que existen subvenciones ofrecidos por muchos organismos (ayuntamientos, diputaciones, Gobierno Vasco…), y que cualquier persona puede obtener subvenciones de menos de 3000€, es posible que incluirlas haga que se pierda eficiencia. Esta cantidad es la mínima para que no tenga que ser concurso público. Si se descartan todas las subvenciones menores a 3000€, nos quedarían algo más de 157.000 subvenciones.  Realizar el análisis de los textos de la subvenciones no será una tarea sencilla, pero teniendo en cuenta la gran cantidad de datos previos que se obtendrían, el sistema obtendría un buen entrenamiento. Habría que tener en cuenta tras realizar los tests oportunos con los datos de test el porcentaje de aciertos se puede considerar lo suficientemente óptimo como para determinar que el sistema predice de manera viable o no. |