Trabajo Práctico 2 : Story Points

Introducción

En este trabajo práctico se va a utilizar un conjunto de datos que contiene una serie de casos de uso (user stories) de distintos proyectos y el número de story points que tiene asignado cada uno. Los story points indican la complejidad de cada tarea. El objetivo será predecir el story point de cada user story dado el texto que lo representa.

Modalidad de entrega

Repositorio

Deberán utilizar el mismo repositorio del TP1: TA047R-2C2024-GRUPOXX
En dicho repositorio deberá estar disponible todo el contenido obligatorio de la entrega (notebooks, modelos entrenados, *datasets*, reportes, presentación) y cualquier archivo que sea necesario para la correcta ejecución del trabajo.

Notebook

El trabajo debe ser realizado en una notebook *Jupyter* de Python, se espera que la misma contenga **todos** los resultados de la ejecución los cuales siempre deben ser **reproducibles**. La notebook debe respetar la siguiente nomenclatura :

TA047R_TP2_GRUPOXX_ENTREGA

En el caso que sea estrictamente necesario entregar más de una notebook las mismas deben contar con una numeración correlativa manteniendo un orden lógico entre ellas (TA047R_TP2_GRUPOXX_ENTREGA_**N1**, TA047R_TP2_GRUPOXX_ENTREGA_**N2**, etc)
Las secciones del trabajo deben estar claramente diferenciadas en la notebook utilizando celdas de *markdown*. Se debe incluir una sección principal con el título del trabajo, el número de grupo y el nombre de todos los integrantes.
Todo análisis realizado debe estar acompañado de su respectiva explicación y toda decisión tomada debe estar debidamente justificada. Cualquier hipótesis que sea considerada en el desarrollo del trabajo práctico debe ser detallada y debe estar informada en la entrega. Cualquier criterio que se utilice basado en fuentes externas (papers, bibliografía, etc.) debe estar correctamente referenciado en el trabajo.

Modelos

Todos los modelos entrenados deben ser guardados en un archivo (joblib / pickle) y deben estar disponibles en la entrega para ser utilizado por el equipo docente.

Reportes

Se deberá confeccionar un reporte (en formato pdf) a modo de resumen de los puntos desarrollados en cada ejercicio. El documento deberá tener la siguiente nomenclatura TA047R_TP2_GRUPOXX_REPORTE y **deberá seguir el template proporcionado por la cátedra**.

Exposición del TP

Cada grupo deberá realizar una exposición de los resultados obtenidos en cada ejercicio del trabajo práctico. La misma no deberá superar los 20 minutos, podrán exponer las notebooks o confeccionar una presentación tipo PowerPoint o similar. **Deberán seguir las pautas mínimas proporcionadas por la cátedra.**

Devolución y comentarios entre grupos

Para cumplimentar la entrega del Trabajo Práctico todos los grupos deberán dar *feedback* sobre las presentaciones del resto de sus compañeros.

Competencia Kaggle

El trabajo práctico estará enmarcado en una competencia de **Kaggle**, dónde todos los alumnos deberán participar. Para unirse a la misma deben acceder con el siguiente enlace y conformar los grupos correspondientes. Pueden elegir cualquier nombre que represente al equipo.

El objetivo de la competencia es predecir con el menor error posible los story points de una user story dado el texto de su descripción y su título. Para saber qué tan bien se desempeña un modelo, cada grupo hará su predicción sobre el conjunto de test y la subirá (submission) a **Kaggle**.

Kaggle verificará las predicciones contra el archivo de soluciones utilizando la **métrica RMSE** y mostrará la posición del equipo en la tabla de puntajes (leaderboard) utilizando el 60% de sus respuestas. El 40% restante se usará también para calcular su puntaje pero en un tablero

privado que sólo pueden ver los docentes y que se revelará al finalizar la competencia (28/11/2024 23:59hs).

Enunciado

Los conjuntos de datos a utilizar **train** y **test** se encuentran disponibles en la competencia de **Kaggle** y deberán descargarlos desde allí. Allí mismo encontrarán también un archivo de ejemplo de cómo se deben subir las soluciones.
El trabajo consiste en construir diferentes modelos de regresión, capaces de analizar una porción de texto en lenguaje natural y predecir los story points. Para ello habrá que realizar un preprocesamiento del texto para que este pueda ser analizado por los distintos modelos. Se utilizará el modelo de ***bag of words*****, o cualquier otro** que permita convertir texto en vectores.

Los modelos que se deben construir son los siguientes:

- Bayes Naïve
- Random Forest
- XGBoost
- Un modelo de red neuronal aplicando Keras y Tensor Flow.
- Un ensamble de al menos 3 modelos elegidos por el grupo.

Para cada uno de estos modelos se debe realizar una búsqueda de hiperparametros que optimicen su desempeño en el conjunto de test local (porción del archivo training).
Una vez encontrados dichos hiperparametros, se procederá a hacer un *submission* a Kaggle. Es decir que habrá al menos 5 submissions (uno por cada modelo).

Reporte

En el reporte se debe explicar de forma clara y sintética los hiperparametros escogidos para cada modelo. En el caso de la red neuronal, se debe indicar la arquitectura y explicar por qué se la eligió. Se deben indicar además las métricas obtenidas en el conjunto de datos de prueba local (porción del archivo training) y el puntaje obtenido en el tablero público de Kaggle. Pueden encontrar un modelo de informe en el siguiente link.

Fechas de entrega

La fecha final de entrega del trabajo práctico es el día 28 de noviembre de 2024.

Condiciones de Aprobación

- **Todos** los integrantes de los grupos deben participar de la competencia.
- Se debe tener **al menos 1 subida** por modelo/ensamble pedido.
- Todas las semanas deben participar en la competencia realizando al menos una subida a Kaggle.
- Deben cumplir con todos los puntos del enunciado
- Deben presentar el informe y realizar una exposición en las fechas estipuladas.