Entrega 1. Minería de Datos

Wenyi He - Leonardo Olivares

28/10/18

1 Propuesta

La propuesta para este proyecto es conocer y visualizar las tendencias en la industria de software, en la actualidad. Del mismo modo, realizar comparaciones con las tendencias del año pasado, para conocer si han ocurrido cambios significativos. Además, se espera implementar modelos de clasificación que permitan identificar atributos de un desarrollador, como su especialidad, a través de su otras características, como por ejemplo los lenguajes de programación que utiliza con frecuencia.

2 Revisión Bibliográfica

En este proyecto se requirieron consultar distintas fuentes bibliográficas, tanto para complementar el estudio planteado y lograr obtener resultados coherentes sobre el contexto abarcado, como para implementar las herramientas correctas de preprocesamiento y manejo de los datos, de tal manera que los resultados aporten la mayor cantidad de información relevante a la propuesta planteada.

A través de una serie de encuestas anuales, realizadas por StackOverflow, se obtuvieron datos de programadores que decidieron participar. En cada una de estas encuestas participaron más de 100.000 programadores, con el fin de conocer las opiniones y contextos de cada uno, para lograr formar una revisión general de las tendencias en la industria de software.

Al obtener los datos, fue necesario realizar una integración de ambos archivos, para tener toda la información en una versión controlada. Para esto, en una etapa de preprocesamiento, se agruparon los datos similares de cada encuesta, y posteriormente se estandarizó cada columna, para trabajar con valores consistentes.

En 2015 se realizó un estudio de los datos obtenidos en la encuesta del año (https://bit.ly/2Ocp042), en el que se presentaron visualizaciones representativas con el fin de acceder a la información de manera rápida y clara. El objetivo del proyecto actual es integrar las encuestas realizadas en los años 2017 y 2018, realizando visualizaciones actualizadas y logrando mostrar posibles comparaciones en los cambios que han ocurrido.

3 Principales Dificultades

Para esta entrega, se presentaron dos principales deficultades.

La primera, esta relacionada con la integración de los datos. Debido a que se utilizaron los resultados de dos encuestas realizadas en años distintos, los datos en su mayoría no eran los mismos, o tenían diferencias como nombres de feature distintas o posibles valores distintos.

Es por esto que fue necesario realizar una etapa de preprocesamiento, en donde primero se realizó la unión de ambas fuentes de datos, de una manera lógica. Luego, fue necesario realizar un proceso de data cleaning, debido a que muchos features de ambas fuentes se realizaron en escalas distintas, o con clases

distintas, y por lo tanto, fue importante dedicar un tiempo a mejorar la base de datos, de tal forma que en procesos posteriores se pudiera obtener la información correctamente.

La segunda dificultad fue buscar visualizaciones ideales para observar la información contenida en la gran cantidad de datos, de tal manera que un usuario sea capaz de entender lo más importante de manera rápida y sin necesidad de requerir los datos crudos.

4 Aprendizajes Adquiridos

Entre los principales aprendizajes adquiridos para esta entrega, destaca conocer la importancia y dedicación que requiere el proceso de pre-procesamiento previo a la implementación de cualquier modelo o visualización sobre los datos. En la vida real, las fuentes de datos tienen muchos errores, como valores nulos o ruido en los datos.

Es por esto que llega a ser complicado realizar integraciones de datos en ocasiones. Es necesario solventar todo tipo de problemas en esta primera etapa, debido a que posteriormente la consitencia y funcionalidad de los modelos necesitan de una fuente de datos limpia, ya que de esta forma nos aseguramos que los resultados sean los más cercanos a la realidad, logrando aportar un verdadero valor al estudio realizado.

5 Bases de Datos

Encuesta 2018
(https://drive.google.com/uc?export=download&id=1_9On2-nsBQIw3JiY43sWbrF8EjrqrR4U)

Encuesta 2017
(https://drive.google.com/uc?export=download&id=0B6ZlG_Eygdj-c1kzcmUxN05VUXM)