Universidad de Sonora

Reconocimiento de patrones: Introducción

Ramón Soto C. (rsotoc@moviquest.com)

Caso de estudio: *Stack Overflow 2018 Developer Survey*

Como caso de estudio principal en el presente curso hemos seleccionado la encuesta de desarrolladores 2018 de *Stack Overflow* disponible en Kaggle. En este primer análisis, realizaremos las fases de comprensión del negocio y comprensión de los datos.

1. Comprensión del negocio

Descripción del negocio:

stackoverflow

Stack Overflow es un sitio web privado, el sitio principal de la Stack Exchange Network creado en 2008 por Jeff Atwood y Joel Spolsky. Es uno de los sitios más importantes de referencia para dudas específicas de programación, basada en asesoría de pares (las respuestas son proporcionadas por otros programadores). La empresa ofrece también servicios comerciales a través de *Stack Overflow Business*.

Objetivos del negocio:

Stack Overflow fue creado como una herramienta social para programadores con el objetivo de evitar estancamientos al momento de programar. El sitio debía crear una especie de biblioteca activa de preguntas y respuestas técnicas a problemas específicos, inusuales o complejos, sobre cualquier área de programación. Un segundo reto de la empresa fue desarrollar un modelo de negocios que permitiera monetizar la creciente popularidad del sitio. El modelo de negocios de Stack Overflow consta actualmente de tres componentes principales (aquí otra referencia):

- Servicios de reclutamiento de recursos humanos. Stack Overflow maneja listas de empleos, identificación de perfiles asesoría para reclutamientos, etc.
- Servicios de mercadotecnia/ads y reputación. Publicidad "no intrusiva".
- Servicios de comunidad privada (Stack Overflow Enterprise). Versión privada de Stack Overflow, donde los expertos son internos a la empresa.

El éxito de la empresa se mide a través de los siguientes criterios:

- Ofrecer al programador un ambiente amistoso donde buscar soluciones a problemas específicos, posiblemente no bien documentados o solicitar el apoyo de otros programadores y obtener respuestas en tiempos muy cortos.
- Ofrecer a los reclutadores, información confiable de posibles candidatos para cubrir un determinado perfil.
- Ofrecer a las empresas canales de publicidad para productos que resulten atractivos a los usuarios a los que se les presenta cada anuncio, evitando incomodarle con publicidada no deseada.

Evaluación de la situación:

Al día de hoy, Stack Overflow ha alcanzado alrededor de 9 millones de usuarios registrados (actualizar aquí) y más de 16 millones de preguntas que han generado alrededor de 25 millones de respuestas. Sin embargo, desde 2011, Stack Overflow lleva a cabo una encuesta entre los desarrolladores registrados en el sitio, que en la edición 2018 alcanzó más de 100,000 aplicaciones.La integración de estas dos fuentes de datos (actividad y encuesta) son suficientemente interesantes para generar información que permita mejorar las ofertas de Stack Overflow. Los principales riesgos en esta iniciativa de minería sobre los datos disponibles para Stack Overflow, con los posibles planes de contingencia son:

Riesgos	**Medidas de control**
Datos insuficientes para llegar a conclusiones útiles	1. Se establecerán objetivos de diferente alcance a fin de tener al menos resultados parciales. 2. Se identificarán tópicos a incluir en la siguiente encuesta, a fin de repetir el ejercició de minería y complementar las conclusiones. 3. Se generarán indicadores adicionales utilizando datos de fuentes adicionales como el del Índice Tiobe sobre popularidad de lenguajes de programación y la información de empleos en indeed.com.
Provocar cansancio en los usuarios que están respondiendo la encuesta	1. Se compartirán los resultados preliminares con los usuarios que respondieron la encuesta. 2. Se incluirá un elemento de reputación adicional para los usuarios que respondieron la encuesta.

Objetivos de la minería de datos:

Objetivos de negocio	Objetivos de minería de datos
Ofrecer servicios de reclutamiento de recursos humanos	1. Generar una herramienta interna para evaluación de los perfiles de usuarios, a partir de sus preguntas y respuestas y de la calidad de las respuestas. Evaluar polaridad de respuestas y comentarios. 2. Detectar las necesidades de capacitación de los usuarios, en términos de las consultas realizadas. Generar información a los reclutadores sobre potencial de usuarios y necesidades de capacitación para ocupar un perfil laboral. 3. Generar una herramienta interna para planificación de cursos y opciones de capacitación que se puedan ofrecer a los reclutadores registrados.
Ofrecer servicios de mercadotecnia/ads "no intrusiva" y reputación	1. Identificar intereses de productos y servicios como pueden ser ambientes de desarrollo, equipo de cómputo o cursos de capacitación. 2. Identificar grupos de usuarios dados sus perfiles, las respuestas en la encuesta y las conversaciones en las que participan.
Ofrecer servicios de comunidad privada	1. Desarrollar un módulo inteligente para detectar necesidades de capacitación, en términos de las consultas realizadas. Utilizar identificación de tópicos, además de las etiquetas específicas utilizadas. 2. Desarrollo de un módulo inteligente para evaluación de los perfiles de usuarios, a partir de sus preguntas y respuestas y de la calidad de las respuestas. Evaluar polaridad de respuestas y comentarios. 3. Módulo para la aplicación de encuestas derivadas de las encuestas para desarrolladores, aplicadas a nivel interno.

Plan del proyecto:

El análisis preliminar y la identificación de 8 objetivos preliminares de minería de datos confirman la viabilidad de desarrollar un proyecto de ciencia de datos para Stack Overflow. Se propone el siguiente plan de trabajo (las actividades señaladas en rojo son consideradas "hitos" en el desarrollo del proyecto):

Actividad / Justificación	Lección	Indicador
Fase preparatoria del proyecto		
Recopilación y análisis de estructura de los datos Se recopilarán los datos de la encuesta *Stack Overflow 2018 Developer Survey* y se analizarán las columnas disponibles y su significado.	2.2 Preparación de los datos II	Tarea #5
Análisis de datos iniciales Se realizará un análisis preliminar de los resultados de la encuesta a fin de comprender la distribución de los datos e identificar potenciales problemas.	2.2. Preparación de los datos II	Tarea #XXX
Preparación de los datos		

Tarea 3

Realice la fase de comprensión del negocio para su problema.

Fecha de entrega: 7 de septiembre.