Project Proposal: Beer Data

Una gran cantidad de la población mundial consume cerveza frecuentemente. Para aquellas personas que se quieren introducir más a fondo en el extenso mundo de la cerveza, presentamos un proyecto en el que analizaremos una base de datos y sacaremos patrones y estadísticas importantes. Nos basaremos en las palabras más repetidas de las reseñas para comprobar si existen relaciones entre ellas y sus puntuaciones. También implementaremos la funcionalidad de que el usuario introduzca una palabra o rango de graduación de alcohol y se le muestran de mejor a peor aquellas cervezas que se encuentren en el dataset.

Actualmente existen gran cantidad de marcas de cerveza, algunas de ellas muy conocidas, pero esto no implica que sean las mejores. Con este trabajo queremos aportar un punto de vista menos comercial en el que las cervezas se valoren únicamente por su calidad y no por la popularidad.
Por otro lado, hay una gran confusión a la hora de calificar una cerveza pues cada persona parece tener gustos muy diferentes. Por esto queremos averiguar si existe alguna característica común que provoque mayor o menor agrado y que por tanto repercute a la hora de puntuar. La subjetividad juega un papel muy importante en este ámbito, cada persona puede tener una opinión distinta sobre la misma cerveza. Por ello también, en la medida de lo posible, queremos centralizar mucha información haciendo uso del Big Data, para así generalizar las múltiples opiniones y facilitar la búsqueda de información al usuario.

Hemos encontrado este dataset en www.kaggle.com tras varias búsquedas por Internet pero investigando más a fondo hemos descubierto que el origen de este es de la página www.beeradvocate.com que se dedica al análisis y puntuación de la cereza entre otras cosas.
El formato del dataset que se va a utilizar es JSON y ocupa 1.161.640 KB, más de 1.5 GB.
El origen de nuestro interés sobre este tema nació cuando encontramos este dataset en una página en la cual usuarios cargan datasets de todo tipo. Nos pareció interesante y encontramos mucho potencial cuando lo analizamos. El dataset está formado por críticas, en cada una de ellas se establecen todos los atributos referidos a la crítica: (id, nombre, volumen de alcohol e información de los ingredientes de la cerveza, nombre de usuario, texto de crítica y otros más).

Click aquí para ver la preview del dataset.

Las infraestructuras y herramientas necesarias para llevar a cabo este proyecto serán las citas a continuación:

- Utilización de la herramienta spark para el tratamiento de datos.
- Google Cloud como soporte básico del cual se hará uso de los clusters para el procesamiento y almacenamiento de datos, así como para conseguir una mayor escalabilidad, flexibilidad y obtener un mejor rendimiento.