Universidad del Valle de Guatemala Mario Barrientos

Programación y Tecnologías Integradoras Carné 13039

Douglas Barrios 16/08/17

**Demostración No Trivial usando Scala**

**Descripción**

Esta demostración es un ejemplo de cómo implementar un algoritmo de machine learning con aprendizaje no supervisado, en este caso K-Means, con un conjunto de datos (dataframe) ya estandarizado en un ambiente de Hadoop. El código de este ejemplo, así como el dataframe se puede encontrar en este github: <https://github.com/mariobsdd/demo-scala>.

La idea de utilizar este ejemplo es que además de aprender cómo se implementan ciertas funciones características del lenguaje de programación de Scala, se implementen otras tecnologías que actúan en conjunto para el funcionamiento de esta demostración. Como se mencionó anteriormente, todo está en un ambiente de hadoop, utilizando tecnologías como Apache Hive (data warehouse en donde estará almacenado el dataframe) y Apache Spark (en donde se ejecutará el código realizado en el lenguaje de programación de Scala). Para esta demostración se utilizaron librerías que provee Spark tanto de machine learning que permiten que no sea muy complicada la implementación del algoritmo, así como también librerías de SQL para poder hacer llamadas de este tipo hacia la data que se encuentra almacenada en Hive, utilizando el mismo lenguaje de programación común para todas las herramientas: Scala.

**Consideraciones**

Para replicar el ambiente que se utilizó en esta demostración, se debe de descargar el conjunto de herramientas de Hadoop. Como es de conocimiento popular, Hadoop se utiliza como un software Open Source como una herramienta de Big Data por todas las ventajas que provee. Este conjunto de herramientas está disponible en un ambiente personal y portable proveído por una compañía de software de Big Data llamada Hortonworks. Este ambiente se llama Sandbox y es un ambiente virtual gratuito que ya posee instaladas todas las herramientas básicas que se utilizan en Hadoop y que incluye una herramienta de administrador llamado Apache Ambari para facilitar el manejo de todo esto para no utilizar la terminal de comandos. Sandbox se puede descargar en el siguiente enlace: <https://es.hortonworks.com/downloads/#sandbox>. Al ingresar al link se debe escoger sobre qué ambiente virtual se desea instalar el Sandbox (VirtualBox, VMWare o Docker), registrarse y se iniciará automáticamente la descarga. Es importante que tome en cuenta los requisitos mínimos de hardware del Sandbox.

La instalación de Sandbox es trivial y si se tuviera algún inconveniente se puede utilizar la siguiente guía de instalación: <https://hortonworks.com/tutorial/sandbox-deployment-and-install-guide/>. Por ende, en esta demostración se toma por hecho que ya se encuentra instalado el sandbox. En el siguiente enlace se puede ver el funcionamiento de esta demostración:

Parte 1>> <https://youtu.be/aOUXT1sHuwE>

Parte 2>> <https://youtu.be/qGrehDeXfyI>