K-Means Clustering

"Clustering is the process of dividing the entire data into groups (also known as clusters) based on the patterns in the data." - Pulkit Sharma (https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/)

K-Means in Spark

1. Look at the following example: https://jaceklaskowski.gitbooks.io/mastering-apache-spark/spark-mllib/spark-mllib-KMeans.html
2. Look at the Meteorite Landings dataset from NASA: https://data.nasa.gov/Space-Science/Meteorite-Landings/gh4g-9sfh
3. The dataset is available in the data folder. It contains the latitude and longitude of the landings. Generate clusters based on the coordinates with SparkML KMeans and save the cluster centers. You need to use a VectorAssembler to transform your data to a feature vector before fitting the K-Means model (see link 3 and 4 below).
4. Generate the clusters again without changing the code. Do you get the same cluster centers as before? Why/why not?
5. Predict which cluster a new meteorite would be in if it landed at your current latitude and longitude. How close is this to the cluster center?
6. Change the number of generated clusters k. How does this influence the time it takes to fit to the data? How does it change your proximity to a cluster center?
7. The training cost is defined as "sum of squared distances to the nearest centroid for all points in the training dataset". It is found by getting the training summary from the K-Means model and calling "trainingCost" (see the documentation for KMeansModel and KMeansSummary for details). How is the training cost influenced when changing the number k?

Docker

If you are using Mac/Windows 10 Pro or Educate and have HyperV activated, you should download docker from here: https://www.docker.com/products/developer-tools (for linux users: you should be able to apt-get the docker package)

Otherwise, you should install docker by following the instructions here: https://docs.docker.com/toolbox/toolbox_install_windows/

Links

1. KMeansModel Documentation: https://spark.apache.org/docs/latest/api/scala/index.html#org.apache.spark.ml.clustering.KMeansModel
2. KMeans Documentation: https://spark.apache.org/docs/latest/api/scala/index.html#org.apache.spark.ml.clustering.KMeans

3. K-Means Clustering with Apache Spark: https://medium.com/rahasak/k-means-clustering-with-apache-spark-cab44aef0a16

4. VectorAssembler: https://spark.apache.org/docs/latest/ml-features#vectorassembler