Module 8: Deep Dive into Spark MLlib

Case Study I

edureka!

Case Study: Spam Detection
Domain: Telecom

There is a telecom operator forum in which cell phone users make public claims about SMS spam messages. The dataset contains a public set of SMS labeled messages that have been collected for mobile phone spam research. The sample collection is composed by 5,574 English, real and non-encoded messages, tagged according to being legitimate (ham) or spam.

Below is the sample dataset:

- Ham: What you doing? how are you?
- Spam: Sunshine Quiz! Win a super Sony DVD recorder if you can name the capital of Australia? Text MQUIZ to 82277

Tasks:

As a big data consultant, you are provided the sample dataset to generate the word cloud using Spark MLlib
You have to load this dataset in the HDFS and perform:

1. Extract words from the SMS message
2. Removed stop words.
3. Modify the stop words to include your custom words such as '-'
4. Create the features from SMS message using CountVectorizer
5. Split the data into train and test - decide on a strategy
6. Use logistic regression and check the accuracy
7. Try to use a Random Forest classifier and see if it increases the accuracy.
8. Introduce bi-gram and tri-gram and note the change in accuracy.
9. Decide on a strategy and generate a data pipeline.