Module 8: Deep Dive into Spark MLlib

Case Study II

edureka!

edureka!

Case Study: Email Analytics
Domain: IT Security Firm

IT International (ITI) is leading the development of new software that could revolutionize how computers support decision-makers.
The IT security team of ITI building this web-based tool for Enron to

- gain insights from the emails in case of fraud and

- identify any abnormal behavior in the email communication to prevent the unexpected.

The dataset contains data from about 150 users, mostly senior management of Enron, organized into folders. The corpus contains a total of about 0.5M messages.

Tasks:

As part of the BigData consultant you are expected to implement the following use cases:

1. Load data into Spark DataFrame
2. Display the top 10 high-frequency users based on weekly numbers of emails sent.
3. Extract top 20 keywords from the subject text for both
   - for the top 10 high-frequency users and
   - for the non-high frequency users.
4. Extract top 10 keywords by identifying removing the common stop words.
5. Extend the stop words dictionary by adding your own stop words such as '__ '
6. Introduce a new column label to identify new, replied, and forwarded messages.
7. Get the trend of the over mail activity using the pivot table from spark itself.
8. Use k-means clustering to create 4 clusters from the extracted keywords.
9. Use LDA to generate 4 topics from the extracted keywords
10. Can you identify top keywords in the spam messages across the organization?