BIG DATA AND TEXT ANALYTICS

Module Outline

Dr. Stephen Thomas
613.533.3390
stephen.thomas@queensu.ca

Description

In this module, we discuss the discipline of big data. The term *big data* is overloaded and simultaneously refers to the description of the data itself, the IT techniques necessary for handling massive data sets, the analytics and AI techniques that are possible with large data sets, and the business decisions/problems that we can address by leveraging the data. We investigate all of these areas.

We review examples of massive data sets, such as credit card transactions, social networks, shopping journeys, call center transcriptions, and online shopping data. We survey key IT technologies, such as enterprise data warehouses, enterprise data lakes, Hadoop, Spark, NoSQL databases, the cloud, and more. We investigate how unstructured textual data – one of the largest and least utilized data sources– can be analyzed and structured for business gain. In particular, we investigate tools and techniques for handling textual data, including preprocessing and representation, extracting structure, topic modeling, classification, and sentiment analysis.

On completion of this module, you will be able to answer the following questions:

- What is big data analytics, and how is different from regular analytics?
- Which big data analytics techniques are most appropriate for a given business problem?
- How will big data analytics change my job and my organization?
- What IT infrastructure components are necessary to implement big data solutions?
- How can organizations leverage unstructured textual data?

The module takes a practical view of big data analytics and uses contemporary case studies to illustrate popular approaches and strategies.

Materials

Case: *Volkswagen Group: Driving Big Business with Big Data*. Ning Su and Naqaash Priani. #W14007-PDF-ENG.

Case: P*redicting Consumer Tastes with Big Data at Gap*. Ayelet Israeli and Jill Avery. HBS 9-517-115.

Topics

- Big data overview and landscape
  - The V's of big data, Analytics 3.0, who's using big data?
- Big data IT concepts
  - Cloud, traditional data architectures (OLTP, ETL, EDW, EDL, OLAP), big data architecture and ecosystem (NoSQL, Hadoop, Spark)
- Big data analytics
  - Association rule learning, recommender systems, social network analytics, association rule learning, classification
- Text analytics and Natural Language Processing overview and landscape
- Techniques and best practices for text wrangling and preprocessing
- Text analytics practice areas and case studies
  - Topic models, document clustering, document classification, sentiment analysis

Further Readings

For interested students, below is a list recommended reading material.

- Phil Simon. *Too Big to Ignore.*
- Viktor Mayer-Schonberger and Kenneth Cukier. *Big Data: A Revolution That Will Transform How We Live, Work, and Think.*
- Thomas H. Davenport. *Big Data @ Work.*
- Christian Rudder. *Dataclysm.*
- Bart Baesens. *Analytics in a Big Data World.*