Spring 2018

Big Data Science

Assignment One

Learning Outcomes

- *Learning over 40 concepts around data analytics*
- *Learning the CRISP-DM process on a real analytics use cases*

I. (80pts) Basic Knowledge and High-level Architectural Questions

1. *Forty+ Concepts: Enrich your Analytics Glossary*

Research, briefly explain and define the following concepts in ***your own words***.

Most of the concepts bellow has been discussed during lecture one and two. Provide examples/scenarios. The answer for each term should not exceed a page per concept (and a minimum of a paragraph per concept). For the terms highlighted in yellow you are asked to provide a short/detailed example. (Terms highlighted in Yellow)

Please cite all your sources if you used any references including the required and recommended textbook.

1. Prescriptive Analytics
2. Kafka
3. Apache Spark
4. Neural Networks
5. Big Data
6. Trust Based Recommender Systems
7. Linear Regression
8. Gradient boosting
9. Knowledge Discovery
10. Class Label (in Data Classification)
11. KNN (K nearest neighbor)
12. "Analytic"
13. Hadoop 2.0
14. Deep Belief Networks
15. Deep Learning
16. Convolutional Neural Networks
17. Feature Selection
18. Business Intelligence
19. Cross-validation
20. Graph Database
21. Confusion Matrix
22. Split Validation
23. Sentiment Analysis
24. Feature (in data analytics)
25. Semi-Structured Data
26. Structured Data
27. Unstructured Data

28. Data Clustering
29. Granger Causality
30. Data Classification
31. Supervised Learning
32. Triplestore
33. Unsupervised Learning
34. Training Data vs Test Data (in the context of cross validation)
35. Deep Learning
36. Ensemble Methods
37. ETL Jobs
38. SQL
39. Alternative Data (in the financial investment context)
40. CRISP-DM

II. (50pts) Questions from the readings and usecases:

Hash Joins

This is a fact: Distrusted architectures existed before Google and MapReduce. Hash Joins existed before Hadoop and MapReduce. Nowadays, most Relational Database Management Systems (RDBMS) offer Hash Joins instead of SQL Joins. Hash Joins is method for joining large data sets.

Explain in your own words the concept of Hash Joins. Provide an example. Describe the differences between Hash Joins and Normal SQL Join.

Explain this statement: ***the computational cost of a Hash Join depends on the cost of building the hash table.***

III. (100pts) Real Case Scenario for Data Analytics Lifecycle Project

This exercise will also count as 1.5% of your term paper. You will need to answer these questions in at least a one page each (execluding) the references page.

(50pts) Uplifting Models and the 2012 Obama's Campaign

We discussed in class the use of uplifting predictive models in marketing and politics.

1. Describe in details your own words uplifting models, their advantages and disadvantages. Explain how uplifting models were used in the 2012 political campaign. Cite all your sources.

2. Explain how uplifting models can be applied in Marketing.

Cite all your resources:
e.g: Research papers from: http://ieeexplore.ieee.org/Xplore/home.jsp
or new articles: sample

(50pts) Data Understanding in Behavioral Analytics

Consider the situation where you will help a business identify their customers' cluster type: (1) Loyal Customers, (2) Discount Customers, (3) Lost Causes, (4) Sure Things, (5) Sleeping Dogs and (6) Persuadables.

1. Define each type of these customers (explain how each customer of the above type has a unique behavior)
2. Explain in details the data sources (unstructured vs. structured) you will need to be able to apply data clustering to identify these customer's type (transactional data, social media data, cookies data, click stream…)
3. Bonus (+15pts) provide a sample data sets and a design process (following CRISP-DM) on how you will be using the dataset to cluster customers and later predict a new unknown customers type.

Cite all your resources:
e.g: Research papers from: http://ieeexplore.ieee.org/Xplore/home.jsp
or news articles: sample