FINA8823 Group Project: Rapid Miner

Presented by Alex, Christos, Dan, Ramin, Yuchen

University of Minnesota, Carlson School of Management

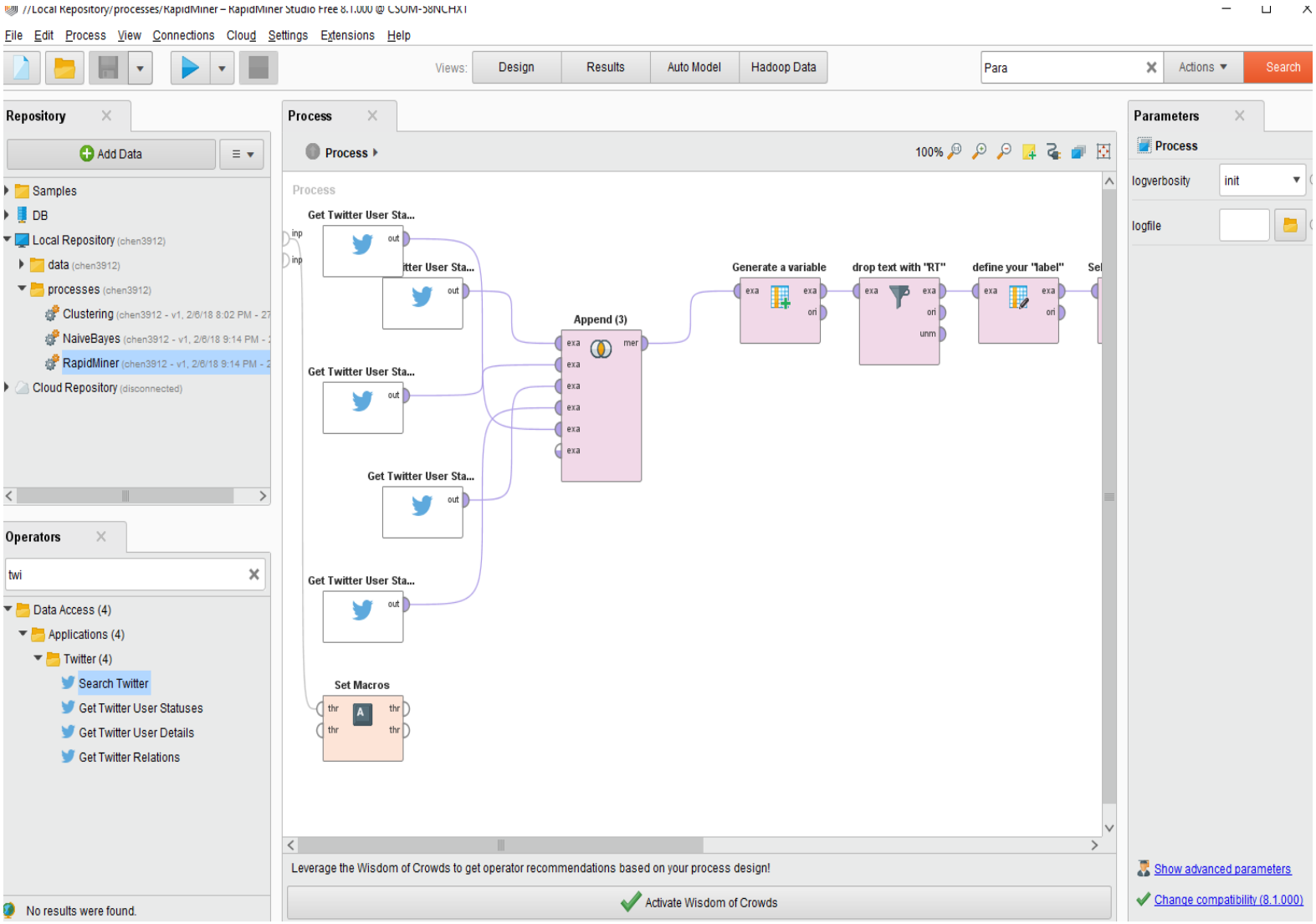
February 11, 2018

Introduction: What is Rapid Miner?

RapidMiner provide an advanced analytical solution through template-based frameworks that speed delivery and reduce errors by nearly eliminating the need to write code. It provides data mining and machine learning procedures including

- data loading and transformation
- data preprocessing and visualization
- predictive analytics and statistical modeling
- evaluation, and deployment

Introduction: What is Rapid Miner?



Introduction: What is Rapid Miner?

- Repository: where you import and store datasets
- Operators: all the models and algorithms are presented
- Process view: for analysis and editing
- Help window

Group Project: Classify Companies' Twitter News

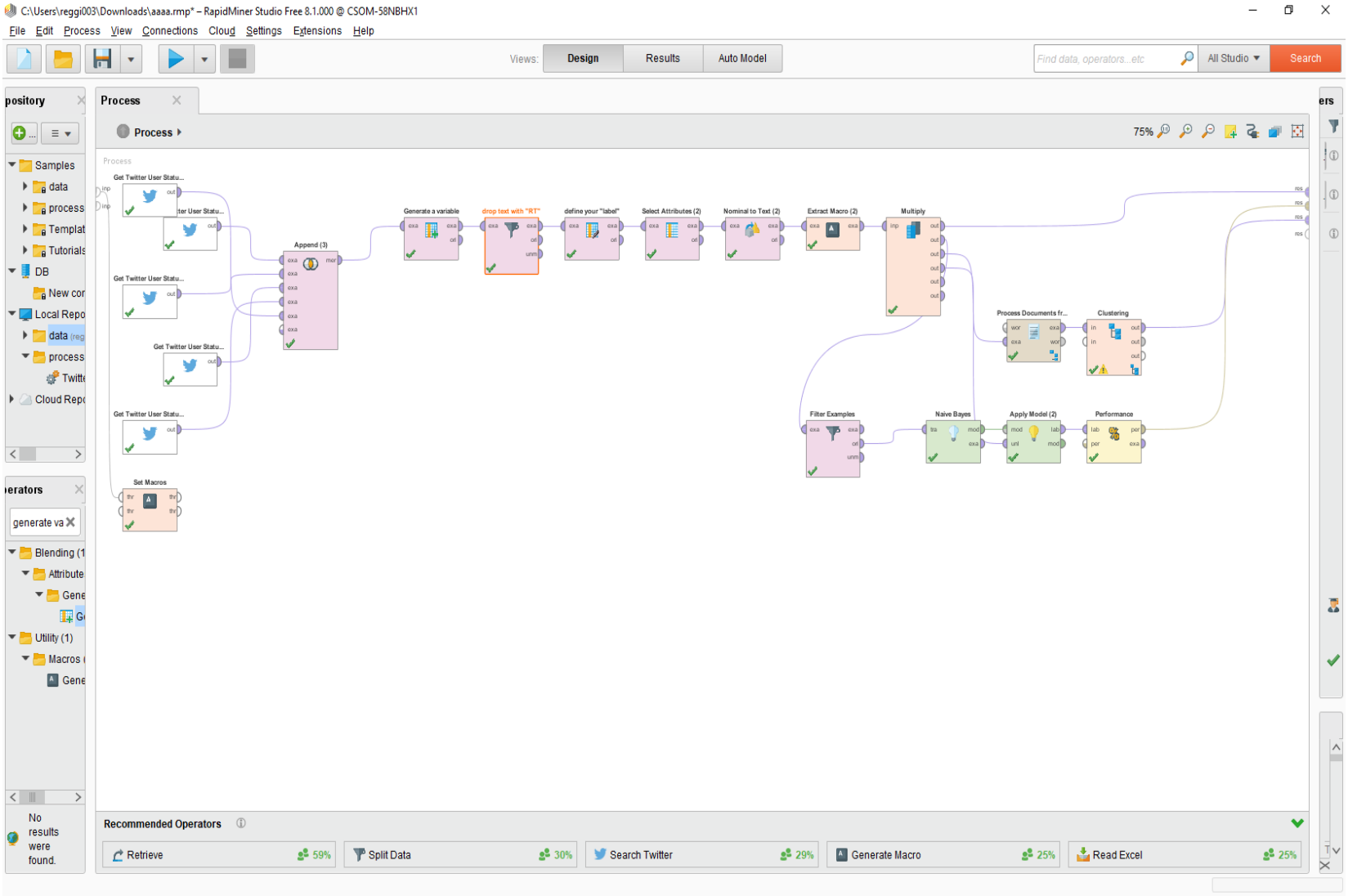
1. Web Scrapping
2. Text Analysis
   2.1 Unsupervised Learning: Clustering
   2.2 Supervised Learning: Naive Bayes

Group Project: Data

Twitter News from

- GE
- IBM
- Amazon
- Yahoo
- Google

Group Project: Process



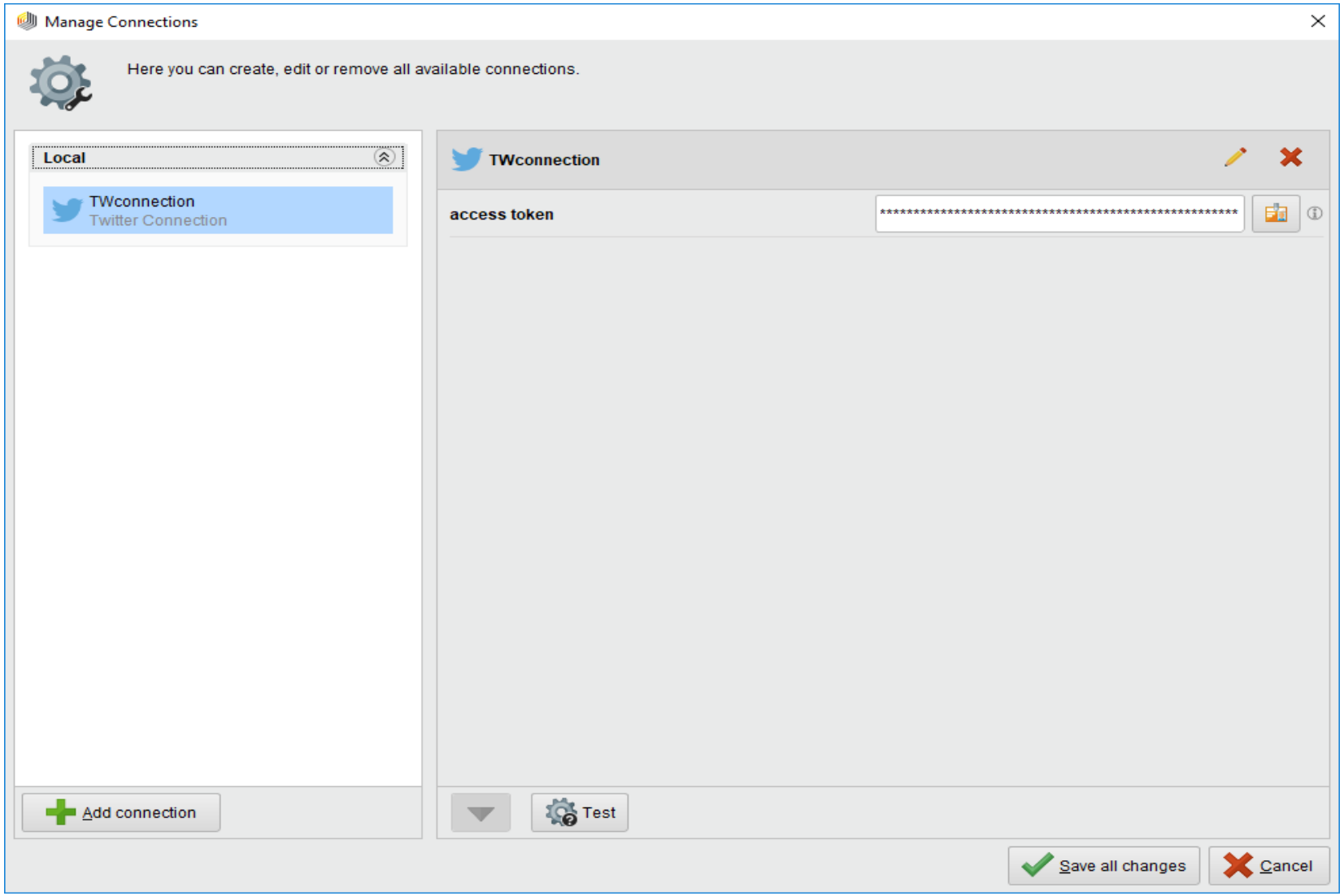
Group Project: Web Scraping

1. Data Access $\Rightarrow$ Search Twitter $\Rightarrow$ Scrape the twits from the user account
2. Create five similar boxes
3. Information: user id, geo latitude, longitude of the ip, language, retweet count, text, source...
4. Get 100 twits from Yahoo, GE, IBM, Amazon and Google

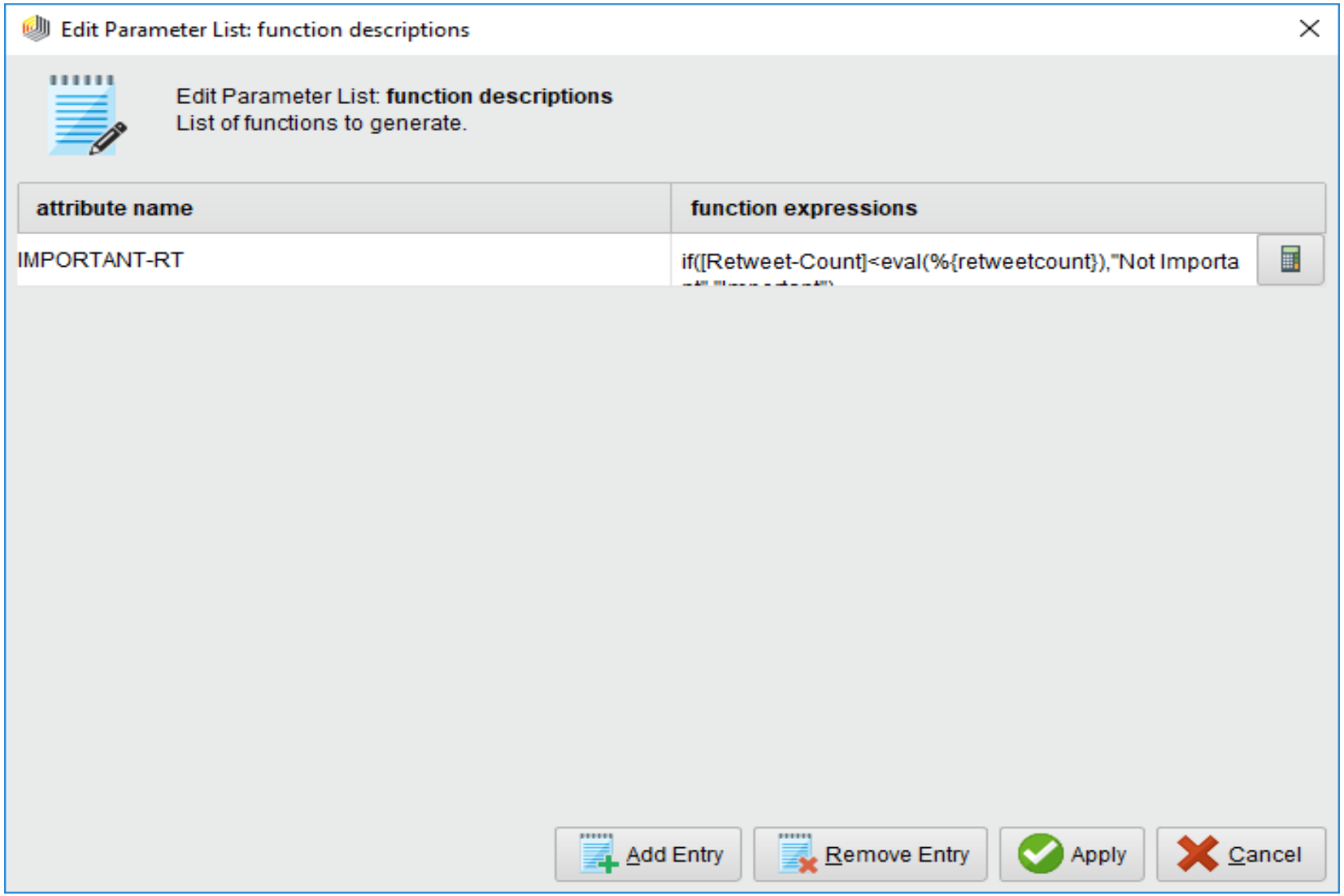
Group Project: Text Analysis

1. Macros (conditions to select/drop/generate variables)
   - desired processing date
   - criteria to classify the importance of the twits
2. Append: combine all the tweets scraped
3. Generate a variable: ”IMPORTANT-RT”
4. Select Attribute: drop/keep variables (”IMPORTANT-RT”,”Text”, ”label”)

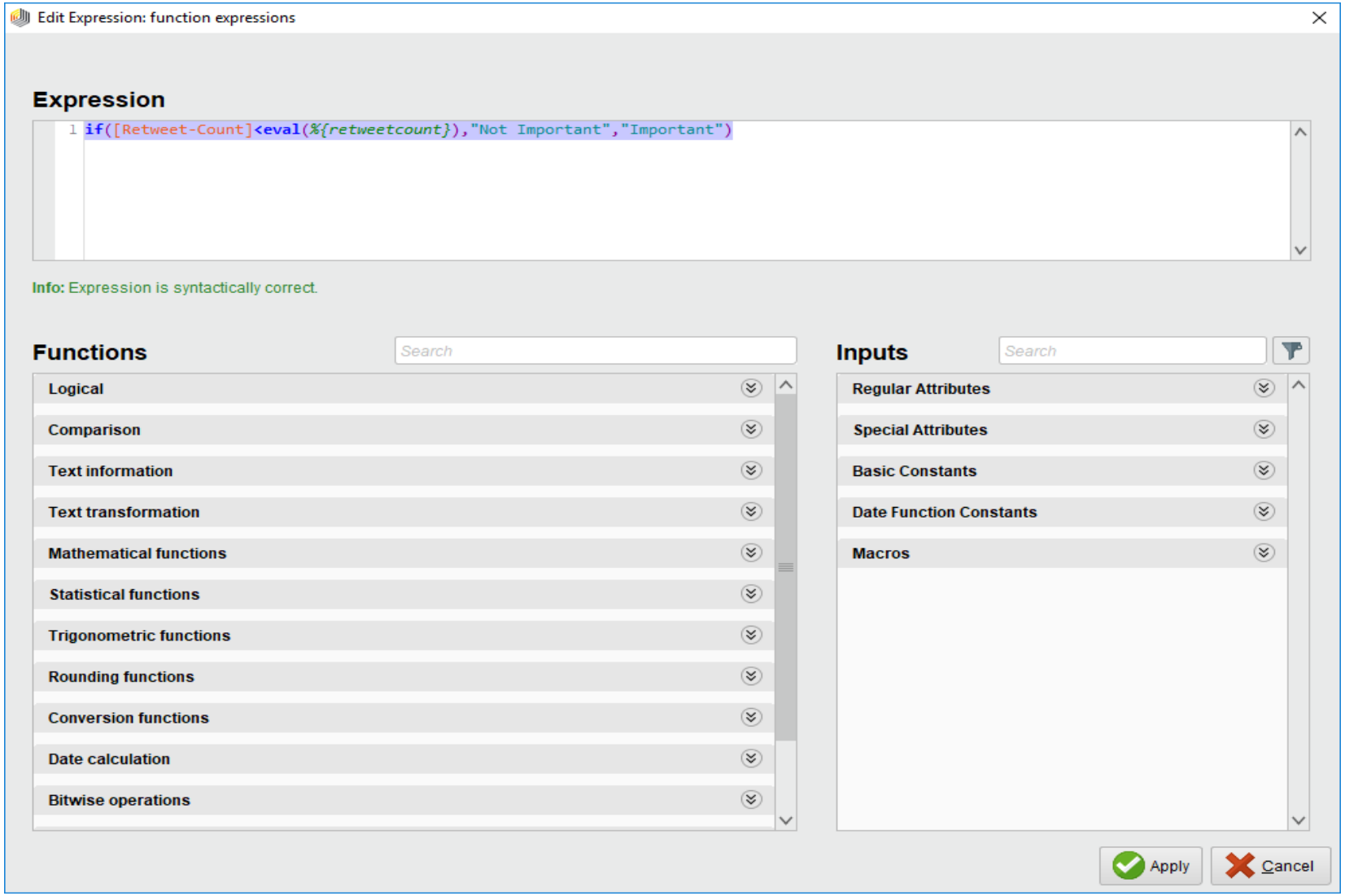
Group Project: Text Analysis



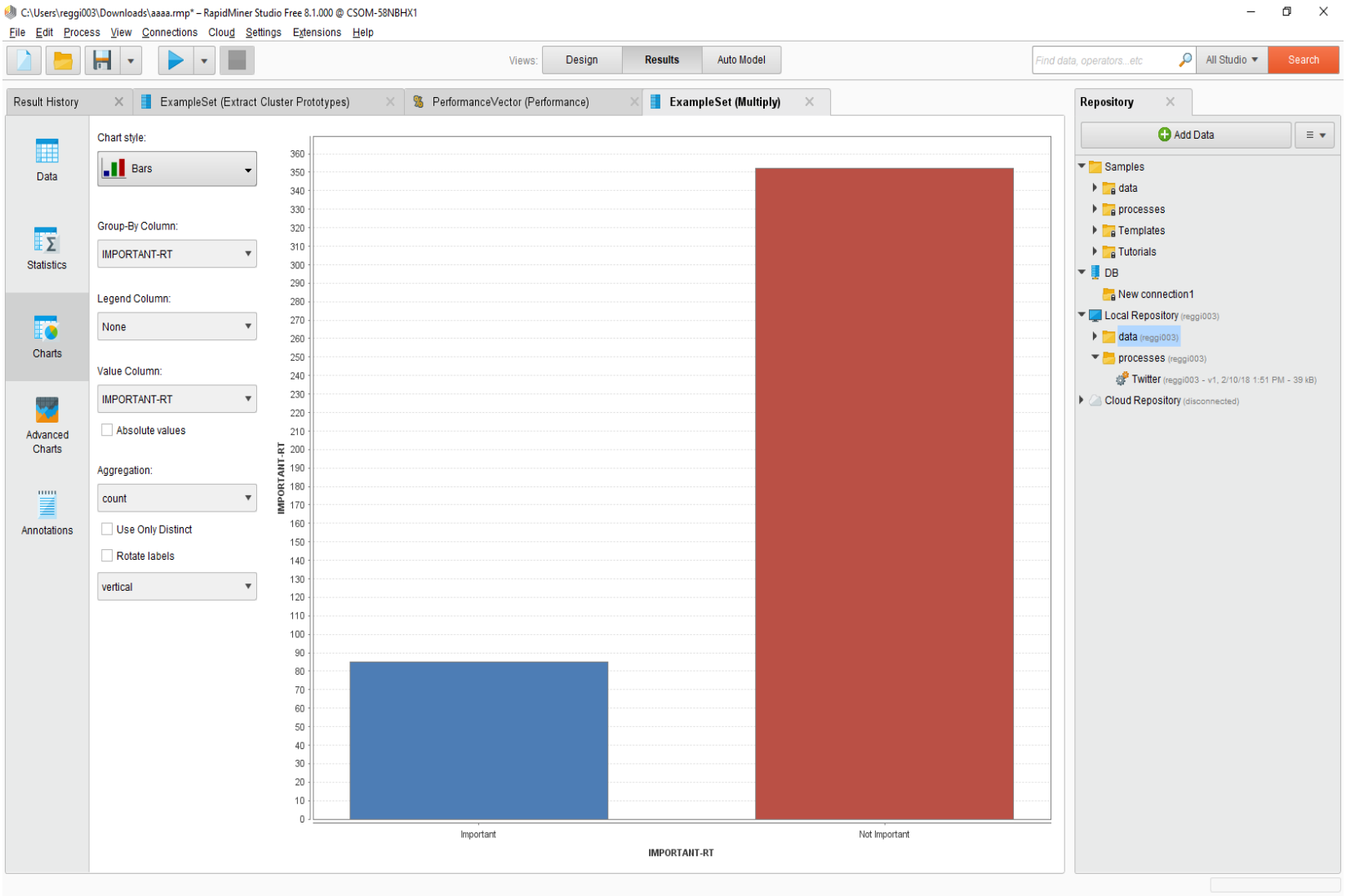
Group Project: Text Analysis



Group Project: Text Analysis



Group Project: Text Analysis



Group Project: Clustering

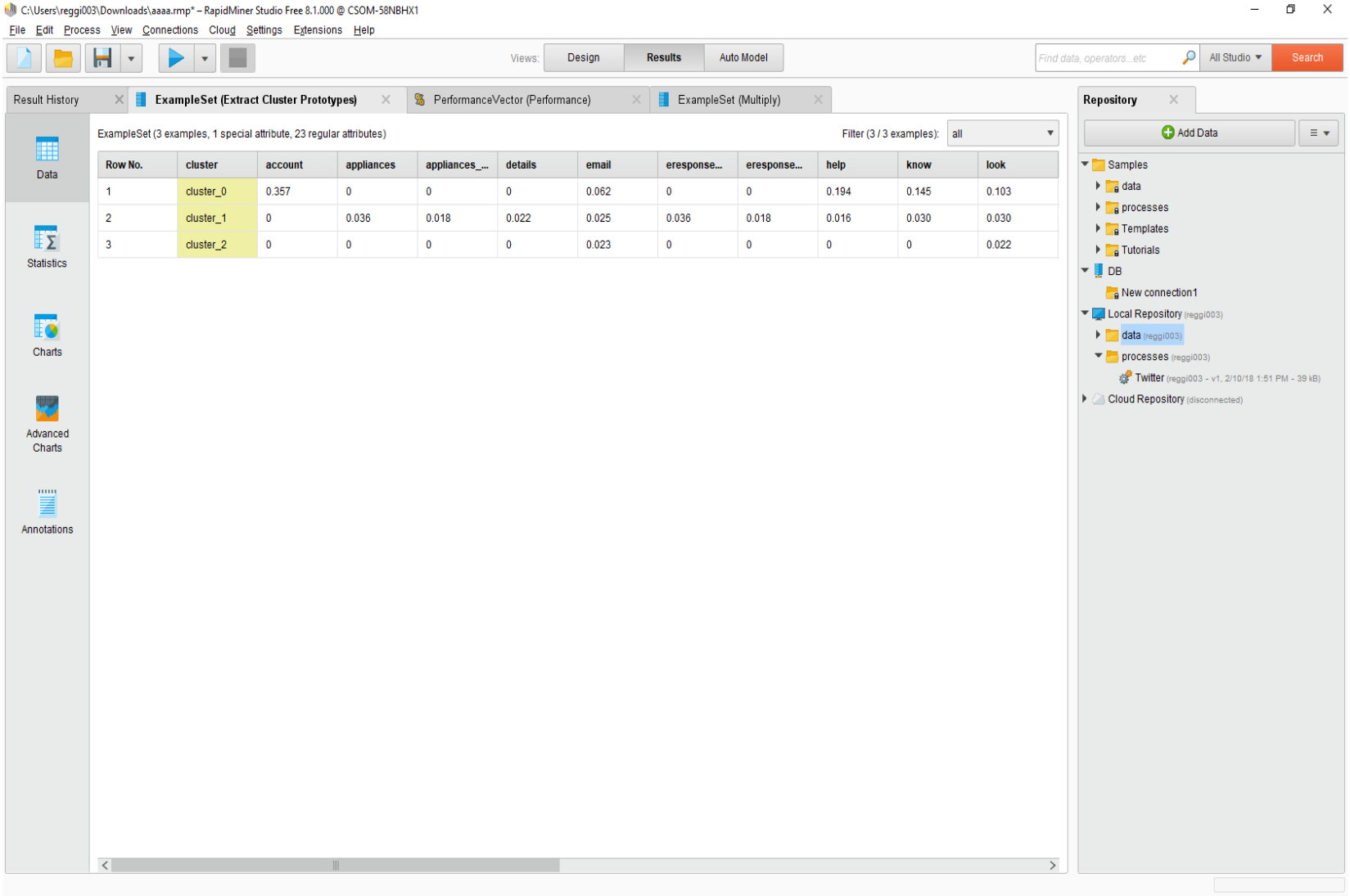
Algorithm: Kmean

k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster

1. Assignment step: Assign each observation to the cluster whose mean has the least squared Euclidean distance (”nearest” mean)
2. Update step: Calculate the new means to be the centroids of the observations in the new clusters.

Group Project: Clustering

Result



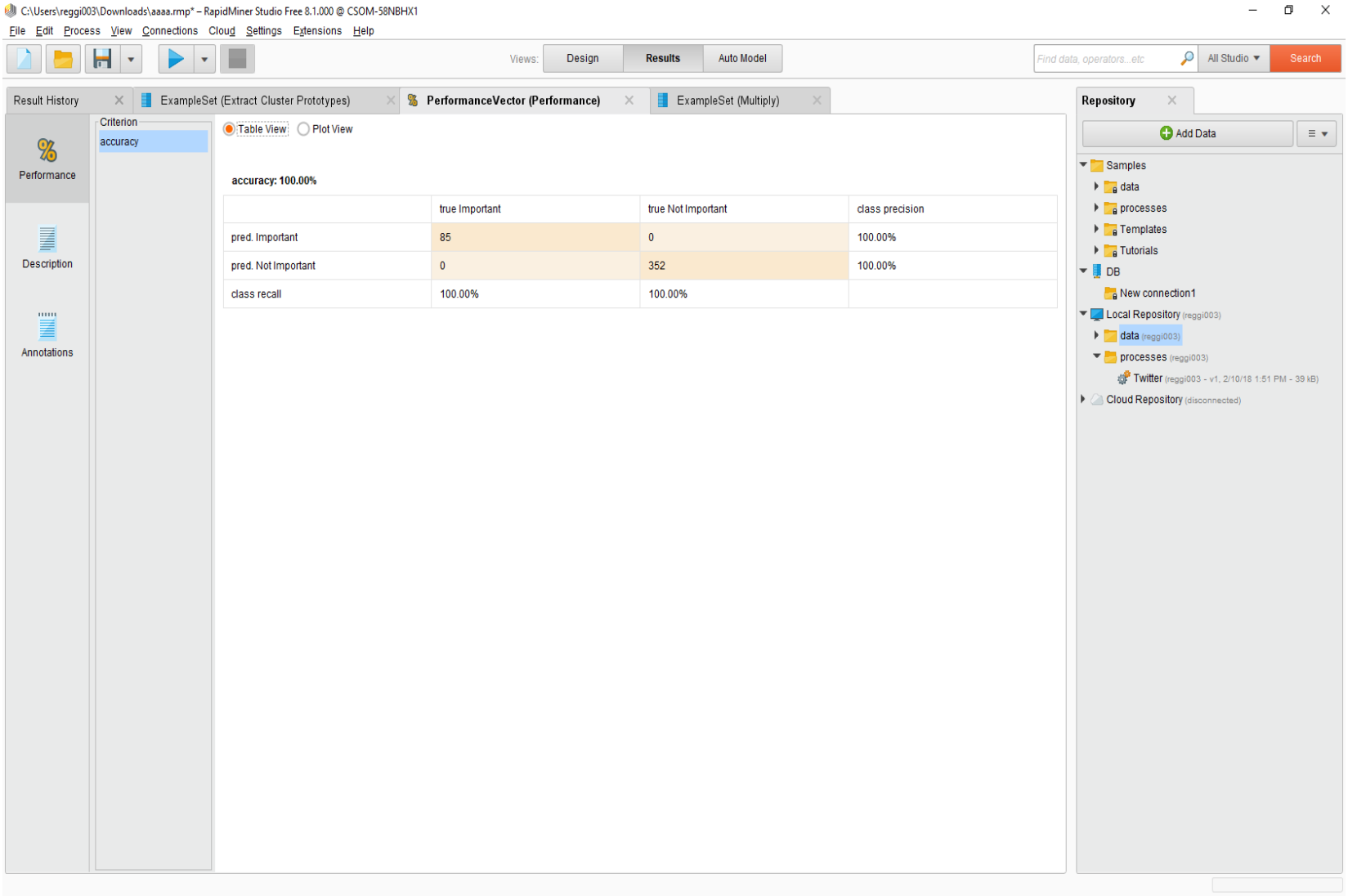
Row No.	cluster	account	appliances	appliances_...	details	email	eresponse...	eresponse...	help	know	look
1	cluster_0	0.357	0	0	0	0.062	0	0	0.194	0.145	0.103
2	cluster_1	0	0.036	0.018	0.022	0.025	0.036	0.018	0.016	0.030	0.030
3	cluster_2	0	0	0	0	0.023	0	0	0	0	0.022

Group Project: Naive Bayes

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature

1. Convert the data set into a frequency table
2. Create likelihood table by finding the probability distribution over the predictors and classes
3. Use Naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.

Group Project: Naive Bayes



Some useful material

- Exploring data with RapidMIner-Andrew Chisholm. Ebook ($\$14.95$)
- Youtube playlist on RapidMiner: Text Mining, Web Crawling, Web Scraping examples
- Tutorials on how to use neural nets and Artificial Intelligence to predict trend models
- More details on how to scrape and treat Twitter texts in RapidMIner

Any Questions?

$$Thank \quad You \ !$$