

# 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?

Local repository/processes/rapidminer - rapidminer studio Free 8.1.0.0 @ L3JUN-386LH1

File Edit Process View Connections Cloud Settings Extensions Help

Views: Design Results Auto Model Hadoop Data

Para X Actions Search

**Repository**

Add Data

- Samples
- OB
- Local Repository (chen3912)
  - data (chen3912)
    - processes (chen3912)
      - Clustering (chen3912 - v1, 2/6/18 8:02 PM - 2)
      - NaiveBayes (chen3912 - v1, 2/6/18 9:14 PM - 2)
      - RapidMiner (chen3912 - v1, 2/6/18 9:14 PM - 2)**
- Cloud Repository (disconnected)

**Operators**

tail

- Data Access (4)
- Applications (4)
  - Twitter (4)
    - Search Twitter
    - Get Twitter User Statuses
    - Get Twitter User Details
    - Get Twitter Relations

No results were found.

**Process**

Process

100%

Get Twitter User Sta... iter User Sta... Append (3) Generate a variable drop text with "RT" define your "label" Set

Get Twitter User Sta... Get Twitter User Sta... Get Twitter User Sta... Get Twitter User Sta... Set Macros

Leverage the Wisdom of Crowds to get operator recommendations based on your process design

Activate Wisdom of Crowds

**Parameters**

Process

logverbosity init

logfile

Show advanced parameters

Change compatibility (8.1.0.0)

# 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

C:\Users\egg\Downloads\aaaaa.mp4 - RapidMiner Studio Free 8.1.000 © CSOM-5181411

File Edit Process View Connections Cloud Settings Extensions Help

Views: Design Results Auto Model

Find data, operators, etc. All Studio Search

75%

Process

Process:

Get Twitter User Status... (4x) → Append (2) → Generate a variable → Group test with "RT" → define your "label" → Select Attributes (2) → Normalized to Test (2) → Extract Macro (2) → Multiply → Filter Examples → Select Rows → Apply Model (2) → Performance → Process Documentation → Clustering

Recommended Operators

Operator	Usage
Retrieve	59%
Split Data	30%
Search Twitter	25%
Generate Macro	25%
Read Excel	25%

# Group Project: Web Scraping

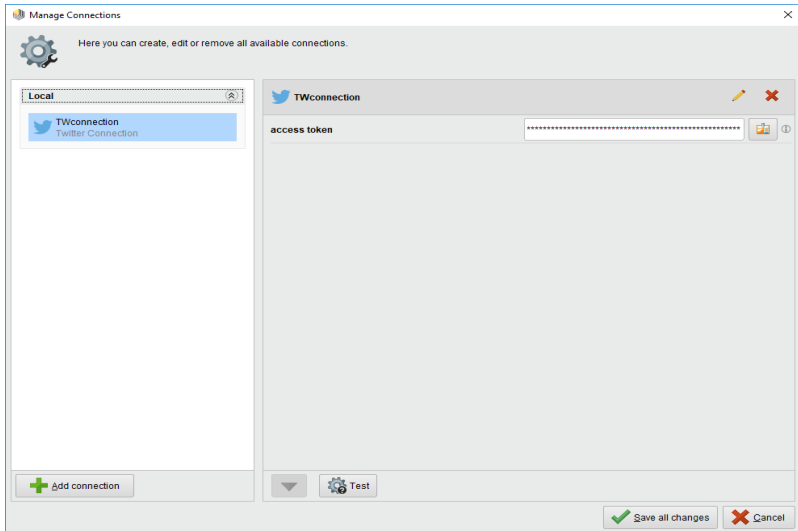
1. Data Access  $\Rightarrow$  Search Twitter  $\Rightarrow$  Scrape the twits from the user account
2. Create five similar boxes
3. Information: user id, geo latitude, longitude of the ip, language, retweet count, text, source...
4. Get 100 twits from Yahoo, GE, IBM, Amazon and Google




# Group Project: Text Analysis


1. Macros (conditions to select/drop/generate variables)
  - ▶ desired processing date
  - ▶ criteria to classify the importance of the tweets
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

 Edit Parameter List: function descriptions ✕

 Edit Parameter List: **function descriptions**  
List of functions to generate.

attribute name	function expressions	
IMPORTANT-RT	if([Retweet-Count]<eval(%{retweetcount}),"Not Importa	

 Add Entry  Remove Entry  Apply  Cancel

# Group Project: Text Analysis

Edit Expression: function expressions

Expression

1 if([Retweet-Count]<eval(%{retweetcount}),"Not Important","Important")

Info: Expression is syntactically correct.

Functions

Search

Logical

Comparison

Text information

Text transformation

Mathematical functions

Statistical functions

Trigonometric functions

Rounding functions

Conversion functions

Date calculation

Bitwise operations

Inputs

Search

Regular Attributes

Special Attributes

Basic Constants

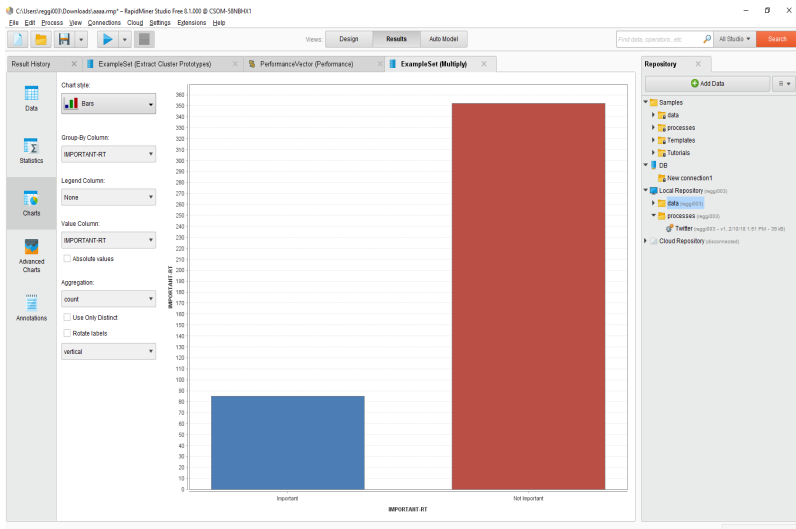
Date Function Constants

Macros

✓ Apply

✗ Cancel

# 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

C:\Users\regg007\Downloads\aaaaa.mpg - RapidMiner Studio Free 8.1.000 © CSOM-SINB4K1

File Edit Process View Connections Cloud Settings Extensions Help

Views: Design Results Auto Model

Find data, operators, etc. All Studio Search

Result History ExampleSet (Extract Cluster Prototypes) PerformanceVector (Performance) ExampleSet (Multiply)

ExampleSet (3 examples, 1 special attribute, 23 regular attributes) Filter (3 / 3 examples): all

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	0.184	0.145	0.103
2	cluster_1	0	0.036	0.018	0.022	0.025	0.038	0.018	0.016	0.030	0.030
3	cluster_2	0	0	0	0	0.023	0	0	0	0	0.022

Repository

Add Data

- Samples
  - data
  - processes
  - Templates
  - Tutorials
- DB
  - New connection1
- Local Repository (regg003)
  - data (regg003)
  - processes (regg003)
    - Twitter (regg003 - v1.2/15/18 1:51 PM - 39 KB)
- Cloud Repository (disconnected)

# 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

C:\Users\egg00\Downloads\aaaaa.mp4 - RapidMiner Studio Free 8.1.000 © CSOM-SINBAKIT

File Edit Process View Connections Cloud Settings Extensions Help

Views: Design Results Auto Model

Find data, operators, etc. All Studio Search

Result History ExampleSet (Extract Cluster Prototypes) PerformanceVector (Performance) ExampleSet (Multiply)

Criterion: accuracy

Table View Plot View

accuracy: 100.00%

	true Important	true Not Important	class precision
pred. Important	85	0	100.00%
pred. Not Important	0	352	100.00%
class recall	100.00%	100.00%	

Repository

Add Data

- Samples
  - data
  - processes
  - Templates
  - Tutorials
- DB
  - New connection1
- Local Repository (egg003)
  - data (egg003)
  - processes (egg003)
    - Twitter (egg003 - v1.2/15/18 1:51 PM - 30 KB)
- Cloud Repository (disconnected)

# 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 You !*