Data Mining Assignment – Twitter Airline Sentiment

Created by Dave Eargle  
Last Updated 4/5/2016

Data Source: <https://www.kaggle.com/crowdflower/twitter-airline-sentiment>

Data description: “A sentiment analysis job about the problems of each major U.S. airline. Twitter data was scraped from February of 2015 and contributors were asked to first classify positive, negative, and neutral tweets, followed by categorizing negative reasons (such as "late flight" or "rude service").” The dataset has over 14,000 tweets.

# Learning Objectives

* Text mining in RapidMiner
* Association rules in RapidMiner
* Working with Loops in RapidMiner

# Before you start

* Complete all of the RapidMiner tutorials
* Complete the following chapters’ exercises “Data Mining for the Masses”, Dr. North, 1st edition. Available to purchase on Amazon or for free PDF download [here](http://docs.rapidminer.com/downloads/DataMiningForTheMasses.pdf).
  + Chapter 5: Association Rules
  + Chapter 12: Text Mining
* You will need the following datasets:
  + *Twitter-airline-rapidminer.csv.* This is a version of the dataset available on Kaggle that can be easily imported into RapidMiner. All ‘@<username>’ have been removed from the tweets, and newline characters have also been removed.
* You will need to install the ‘Text Processing’ extension into RapidMiner. The process is slightly different than the one described in the 1st edition of the textbook. Click the ‘Extensions’ menu item, then ‘Marketplace’, create an account, and search for and install ‘Text Processing.’

**Note**: this assignment borrows heavily from @Sole’s “Airline Sentiment: Part 2” script, coded in R, found on Kaggle.com (https://www.kaggle.com/solegalli/d/crowdflower/twitter-airline-sentiment/airline-sentiment-part-2). This assignment translates @Sole’s analysis into RapidMiner, and performs a different association rule analysis.

# Context

You are wondering what the most common problems are among airlines, for both the positive and negative tweets.

# Process

* What words are found most commonly for positive tweet? For negative ones?
  + Import the dataset into RapidMiner.
  + Use `Filter Examples` to remove all examples with a ‘neutral’ airline\_sentiment
  + Convert the ‘text’ field from ‘Nominal to Text’ (required for text mining process to work)
  + Use `Loop Values` to loop over all remaining values of ‘airline\_sentiment’
  + Inside the ‘Loop values’ process:
    - filter the dataset down to just the current loop\_value (review the RapidMiner tutorials if you have forgotten how to access macro values).
    - Select just the ‘Text’ attribute. That’s the tweet.
    - Use the `Process Documents from Data` operator to perform the text mining tokenization.
      * Very Important: Change the ‘Process Documents’ parameter ‘vector creation’ to ‘Binary Term Occurrence’. You will recall from Chapter 5 that the Association Rules operator requires binary attributes.
      * Also, tick the ‘keep text’ and ‘add meta information’ parameters for the `process Documents from Data` operator.
      * Inside the `Process Documents` process:
        + Perform the tokenization steps you learned from Chapter 12.
        + Hint: you can use the `Filter Tokens (by Length)` operators to remove nonsensical single-character tweets users may have commonly used.
        + Hint: use `Filter Tokens (by Content)` in combination with ‘invert selection’ to remove nonsensical tokens.
      * Take the output from your `Process Document` operator and run it through a `Numerical to Binomial` datatype conversion operator. This will convert 1’s and 0’s over to the binary data type.
      * Use the output of the `Numerical to Binomial` conversion as input for a `FP-Growth` operator.
      * Follow the rest of the steps from Chapter 5, including using a `Create Association Rules` operator, to create association rules for the tokens extracted from the tweets.
        + Important parameters for your `FP-Growth` operator:

Drop the ‘min support’ parameter to 0.02

Make sure the `max items` parameter is -1

* + - * + Important parameters for your `Create Association Rules` operator:

Drop ‘min confidence’ parameter to 0.5

* Hint: The mining can take a long time to run. While you are debugging, use Breakpoints (review the tutorials) to help cut things short. Also consider using a `Filter Example Range` operator in the first step of your process so that you can take a subset of your dataset to work on – just until you have the process down. If you take a subset of examples like this, just don’t forget to remove it before you analyze and report results!

# Questions

* How many negative tweets are there? How many positive?
* What are the top 10-or-so words extracted from *negative* tweets? What does it seem like they’re complaining about?
  + Hint: Look at your FrequentItemSets output and sort descending by ‘Support’
* What are the top 10-or-so words extracted from *positive* tweets? What does it seem like they’re complaining about?
* What are some of the interesting association rules that were extracted for *positive* tweets? Pick one, and report its ‘support’ and ‘confidence.’
* What are some of the interesting association rules that were extracted for *negative* tweets? Pick one, and report its ‘support’ and ‘confidence.’
* Based on what you saw in the negative tweets, what areas would you recommend that the airlines overall work on to improve their image?

# Deliverable

* The answers to the above questions, in .docx or .pdf format.
* A copy of the RapidMiner process(es) you used to obtain the answer(s).