Cleaning the London Sample data

When it comes to parsing text fields and coding, I am one with little skill. However, my recent work experience has led me to gain great skills in SQL and other database languages. I feel like this knowledge is perfect for data science, as most the data explored will be in some kind of tabular format. For the London Olympic sample data, I was able to import the data into a database and utilize some basic substring analysis.

The goal of my analysis was to find if a tweet was a retweet, remove some of the ‘RT ‘ syntax, and find the user that was being retweeted at. After importing the data, this is the SQL query I used to clean the data:

WITH cleaner AS(

SELECT [created\_at]

,[text]

,[from\_user]

,CASE WHEN [text] LIKE 'RT @%' THEN 1 ELSE 0 END AS Retweet

,CASE WHEN [text] LIKE 'RT @%' THEN SUBSTRING(text,CHARINDEX(' ',text)+1,LEN(text)) ELSE TEXT END AS clean

FROM [finance].[dbo].[londonsample])

SELECT \*

,CASE WHEN [clean] LIKE '@%' THEN SUBSTRING(clean,CHARINDEX('@',clean),CHARINDEX(' ', clean)-2) ELSE NULL END AS At

FROM cleaner

The query utilizes what are called CTE’s from the ‘WITH’ command. CTE’s act, more or less, as a table variable that can contain the schema of an entire table from a select statement. I use the CTE to create temporary version of the data so I can then break apart the User name from the retweets.

This method is a good and quick for the purpose defined, however it has some problems. For instance, if a tweet contains a username that is not part of a retweet, i.e. at the beginning, then it will not pick it up. This method, while I’m sure it could, does not pick up mentioned hash tags. While this method may not do these things, there is no reason that it couldn’t, I just don’t really know how. SQL provides simple and powerful tool for string searching and can accomplish any goal required for twitter data analysis.