Airline Tweet NLP

Jai Bansal

May 18, 2016

This document conducts natural language processing (NLP) analysis on Twitter tweets about major US airlines scraped from the site during part of February 2015.  
Contributors were asked to classify positive, negative, and neutral tweets, followed by categorizing negative reasons (such as "late flight" or "rude service"). Thus, for each tweet, I have the 'correct' answer for sentiment analysis purposes.

The data can be found at the URL below. To find the dataset, search for 'Airline' on the page.  
I specifically use the 16,000 row dataset uploaded on February 12, 2015 by CrowdFlower.  
I assume the upload date is incorrect as the data includes tweets from after 2/12/2015...

<https://www.crowdflower.com/data-for-everyone/>

Note that the actual dataset only appears to contain 14,640 rows. I'm not sure where the discrepancy between actual rows and rows stated on the website comes from, but it doesn't affect my analysis.

### Word Cloud:

Below is a word cloud with the 50 most frequently used words (technically stems) in the tweet data. The larger and darker a word, the more frequently it was used.

