Twitter Sentiment Analysis using Super Bowl 50 tweets

—

Who are “They” going for?

- We all hate people who hop on bandwagons!
- Imagine if we could quickly determine which team your bandwagoning friend supported before major sporting events
- Using Super Bowl tweets and the binary classification predictive techniques learned in class we can build a predictive model to achieve this goal

The Dataset

The Data

~2,000,000 tweets streamed using the StreamR module in R from wednesday to the sunday preceding the event in JSON format.

Parsed the JSON file using python to collect ~800,000 tweets on the sunday of the event

32 variables

Interesting observations:

- Most popular hashtags:

#sb50:233423; #superbowl:109966; #keepounding:60295,#broncos:53510

- Most popular relevant mention:

Panthers:73094,broncos:45305,cam:39593,peyton:24599

- Number of times each teams twitter were addressed: @panthers:39020, @broncos:22959

Steps ahead

or problems ahead...

- Choosing Labels: Probably will randomly sample ~3000 tweets in order to manually assign label value
- Feature engineering
- Get familiar with the NLTK package, Naive Bayes, SVMs and maybe more ML techniques
- Get familiar with Regular Expressions to process tweet text

The end

Questions?