Can I use **social media** data to **recommend** countries to **visit** to based on my **travel history**?

Jamar Parris

Data Science Fall 2013

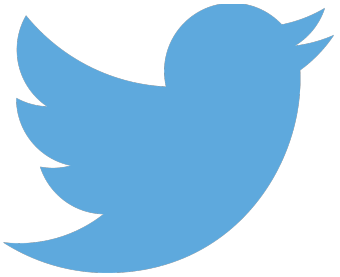
User travel data is scattered among **flights**, **trains** and **road trips**. How do I get a **complete** view of my travel history?

My History

- Since 2010
- 3K check-ins
- 1K venues
- 9 countries

Training Data

- 34K posts
- #{country}
- #hongKong
- #france



- 20K tweets
- “{country} + travel”
- “barbados travel”

Dump everything into MongoDB for later processing.

Processing Text Data

- Standardize data from Instagram and Twitter
  - Extract text from posts, comments and #hashtags
  - 500 combined posts on average per country
- NLTK
  - tokenize words (1.2M non unique tokens)
  - lemmatize words
    - Travel, travels, traveled, traveler -> travel
- SciKit Learn
  - TfidfVectorizer to convert text to sparse matrix

Initial Plan(s)

- Supervised Learning
  - Manually classify the 9 countries I visited
  - Use NaiveBayes
  - Apply categorization to list of 100 countries
- Unsupervised Learning Take 1
  - Apply KMeans to the 9 countries I visited
  - Apply the cluster function to full 100 list
- But I've only been to 9 out of 100 countries so limited training data

Final Plan

- Cluster all 100 countries instead of the 9
  - KMeans = Only algorithm I tried
- Can then see where my 9 countries are within the clusters to make recommendations based on similarity
- Decided on 5 clusters. Why?
  - Seemed best based on manually looking at the cluster applied to each country
  - Small data set of 100 countries made it easy to spot check with different n_clusters

DEMO