Topic Modeling Tweets about the Rio Olympics

By Adam Levin

Motivation

- There are literally millions of tweets every day about the Olympics
- Let's use NLP and Topic Modeling to summarize what folks are tweeting about
- Maybe it will aid our enjoyment of the Olympics

Collecting Tweets*

- Source A:
  - Scraping Twitter's Advanced Search
    - Pros: We can focus on tweets that are about each sport.
    - Cons: It's slow
- Source B:
  - Using Twitter's streaming API
    - Pros: Lot's of tweets (Several Hundred per Minute)
    - Cons: Lot's of junk tweets

*only English tweets considered

Cleaning Tweets

- An Example:
  - Original Text:
    - u'RT @qz: Uh-oh, looks like we jinxed the Olympic water polo pool https://t.co/MMe3rqR8tK #olympicpool'

Cleaning Tweets

- An Example:
  - Encoded in ASCII:
    - 'RT @qz: Uh-oh, looks like we jinxed the Olympic water polo pool https://t.co/MMe3rqR8tK #olympicpool'

Cleaning Tweets

- An Example:
  - Made Lowercase:
    - 'rt @qz: uh-oh, looks like we jinxed the olympic water polo pool https://t.co/mme3rqr8tk #olympicpool'

Cleaning Tweets

- An Example:
  - Without link-like words or ‘rt’ or mentions:
    - 'uh-oh, looks like we jinxed the olympic water polo pool olympicpool'

Cleaning Tweets

- An Example:
  - Lemmatized:
    - 'uh-oh, look like we jinx the olympic water polo pool olympicpool'

Cleaning Tweets

- An Example
  - Bag of words:
    - ['polo', 'like', 'water', 'look', 'pool', 'olympicpool', 'olympic', 'uh-oh', 'we', 'the', 'jinx’]

Challenges

- Diverse topics tweeted about in conjunction with big event like Olympics
- Junk tweets: advertisers seem to tweet out the same tweet over and over but with slightly different spelling to gain attention
- How to determine the number of topics?
- Which stop words to use?

Challenges (continued)

	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10
Topic 1	usa	like	just	dont	mens	im	lol	win	know	play
Topic 2	usa	womens	day	mens	love	group	brittneygriner	year	topps	play
Topic 3	win	gold	usa	nigeria	medal	croatia	au	mens	men	news
Topic 4	usa	mens	play	brazil	coach	good	womens	beat	sexist	win
Topic 5	usa	womens	china	look	like	play	mens	star	channel	woman
Topic 6	usa	womens	cruise	ship	canada	stay	housing	woman	opt	im
Topic 7	usa	mens	samsung	nike	cnet	vr	shoe	thing	boomer	announces
Topic 8	usa	mens	great	bra	teamusa	nba	think	australia	time	just
Topic 9	player	nba	usa	best	carmelo	anthony	mens	make	ticket	ebay
Topic 10	usa	mens	live	vs	stream	spain	argentina	tv	australia	france
Topic 11	usa	serbia	mens	score	france	vs	highlight	reaction	nba	men

Modeling Pipeline

- Each sport independent modeling
- Count Vectorizer → Latent Dirichlet Allocation to get topics
- Term-Frequency Inverse Document Frequency on documents containing all tweets in each topic to get Keywords

Results

- Selected the top 2 or 3 most coherent topics for each sport.
- Selected a sample tweet.
- Made app

Future Work

- Make it so user can flip through multiple real tweets for each topic.
- Markov chain tweet generator for each topic.

Thank You!