Tweet Classification

David Sanford

Data Science Immersive
General Assembly

Wednesday April 5, 2017

Enter the Cacophony

Twitter contains an enormous amount of data, but most is unfiltered

- Keyword/user/location filtering is somewhat effective
- Many subject-related tweets will not contain keywords
- Some keywords have more general context than desired

Tweet content beyond keywords may indicated subject relevance

- Able to select around mis-spellings and abbreviations
- Captures related words and/or terminology beyond the scope of keyword searches
- Captures sets of relevant and/or iconic words

NLP and Machine Learning can attempt to identify these features

It’s a me! Mario! – And Friends

Wish to classify tweets with video game franchise names as keywords

- Producing a clean sample requires both sufficient volume and unique keywords
- Gathered $\sim 200,000$ tweets, divided evenly between keyworded and an unfiltered stream

Tweets are best suited to “bag-of-words” style models

- Mis-spellings, abbreviations, lack of grammar, and emojis are common
- N-grams and other models are more computationally expensive

Tweets are messy! A significant amount of cleaning is required

RT @ForceComYT: #Overwatch - Deutsch / German Let’s Play - S03 - #Competitive Placement Match #07 - https://t.co/PVp3YzYQBf #LetsPlay	- deutsch / german let’s play - - placement match -

Modeling Results

Initial Data Set – 10K tweets from video game and unfiltered streams

- Convert words to numerical inputs using a “tf-idf” vectorizer model + “truncated svd”
  - tf-idf weighs words based on frequency in tweet and corpus
  - truncated svd selects the most important combinations of features

Binary classification performed using six models, with similar performance

Model	Accuracy	Precision	Recall	F1-Score
SVC	0.76	0.90	0.60	0.72

Performance probably requires more cleaning and curation

- Probably useful as a “signal boosting” intermediate filter
- Needs more computational power and processing for better results

Future goal: content modeling on tweets