Content Modeling of Tweets on Twitter

David Sanford

Data Science Immersive
General Assembly

Thursday March 9, 2017

Enter the Cacophony

- Twitter contains an enormous amount of data
- But the breadth of topics makes an unfiltered stream meaningless noise

How can tweets related to a particular subject be tagged?

- Keyword/user/location filtering can reduce the stream somewhat
  - Requires meaningfully restrictive keywords
  - Useful for sample sets, but will cut away a large amount of useful data
- Many subject-related tweets will not contain keywords
- Some keywords have more general context than desired

Is there a better way?

Going Beyond Simple Tagging

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
  - Techinical terms related to a subject
  - Unique terms in various types of fiction
  - Terms more prevalent in a subject
- Captures sets of relevant and/or iconic words

NLP and Machine Learning can attempt to identify these features

Project Goals

- Identify a topic and tweet collection methodology which produces a sufficiently clean sample
- Identify the best modeling methodology
- Clean tweets
- Perform binary classification of topic-related tweets against an unfiltered stream of tweets
- Cluster tweets aggregated on keywords to identify genres within the topics

Choosing the Right Data Set

As a test of concept, a clean data set of subject-related tweets must be used

Tweets from a curated set of users may be usable

- Requires a large number of users and careful curation

A keyword search can get a larger number of tweets covering from many users

- Requires careful choice of keywords

Topic	**Good Keywords**	**Bad Keywords**
Academic Subject	_____ Studies, _____ Sciences	Business, Economics
Tabletop RPG	Dungeons & Dragons, Shadowrun	Werewolf, Call of Cthulhu
Tabletop Games	Settlers of Catan, Scrabble	Risk, Dominion
Video Games	Mario, Zelda, Tetris, Angry Birds	Civilization, Battlefield

It’s a me! Mario! – And Friends

Of the topics I considered, video games had the greatest number of unique names

- https://en.wikipedia.org/wiki/List_of_best-selling_video_games
- https://en.wikipedia.org/wiki/List_of_video_games_considered_the_best

Accepted Keywords: Zelda, Tetris, Mario, Chrono Trigger, Street Fighter, Final Fantasy, Metroid, Half-Life, Resident Evil, Metal Gear, Castlevania, Pokemon, BioShock, SoulCalibur, StarCraft, Shadow of the Colossus, Doom, Diablo, World of Warcraft, Donkey Kong, Pac-Man, Halo, Deus Ex, Space Invaders, Sonic, Counter-Strike, Grim Fandango, Portal, Mass Effect, Last of Us, Star Fox, Mega Man, EarthBound, Prince of Persia, Call of Duty, Dark Souls, Perfect Dark, Ico, The Elder Scrolls, Skyrim, Morrowind, Silent Hill, Shenmue, Grand Theft Auto, Okami, Double Dragon, Red Dead, Galaga, Tomb Raider, Fallout, Uncharted, Assassin’s Creed, Minecraft, Kingdom Hearts, Xenogears, Overwatch, Wii Sports, Wii Fit, The Sims, Terraria, Brain Age, Need for Speed, Lemmings, Madden NFL, Star Wars: Battlefront, Tom Clancy’s, Duck Hunt, Splatoon, Super Smash, Dynasty Warriors, Monster Hunter, Kirby, Fire Emblem, Animal Crossing, God of War, Tekken, Garry’s Mod, Myst, Angry Birds, Candy Crush, Fruit Ninja, Block Breaker, Doodle Jump, Space Invaders, Galaxian, Mortal Combat, Pong, Crysis

Examples of Rejected Keywords: Civilization, Battlefield, Asteriods, Fable, Journey

NLP Modeling for Tweets

Only bag-of-words style models with transformations are likely to be relevant to tweets

- Tweets often lack sentence structure
- Mis-spellings and abbreviations are common
- Many different levels and styles of grammar are on display
- Emojis and hashtags used in place of words
- Many tweets are "stubs"
- Large number of documents in corpus makes tfidf useful

Zelda's super neat but I've experienced more severe frame drops in the first 5 minutes then I'd like to
How To Spot The Difference Between Battleborn And Overwatch #Overwatch #Overwatch https://t.co/Xj8ryeq5Tz https://t.co/uRCxip2kaR
RT @ForceComYT: #Overwatch - Deutsch / German Let's Play - S03 - #Competitive Placement Match #07 - https://t.co/PVp3YzYQBf #LetsPlay

Tweet-Cleaning

Tweets are messy! Significant amounts of cleaning is required.

- Retweet references Retweet references
- Hashtags Hashtags
- User references
- Emojis
- Links Links
- Keywords Keywords
- Proper names
- Unintelligible strings Unintelligible strings

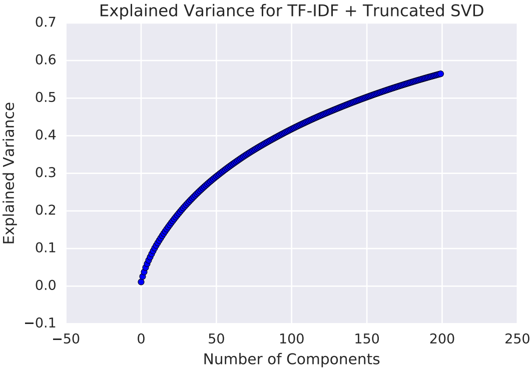
As a first pass, I treat all the above items as “stop words” and remove them.

Zelda’s super neat but I’ve experienced more severe frame drops in the first 5 minutes then I’d like to	’s super neat but i’ve experienced more severe frame drops in the first 5 minutes then i’d like to
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 -

NLP Processing – Warped Tweets

Initial Data Set – 10K tweets from video game and unfiltered streams for both training and validation sets

- Training set used to train tf-idf vectorized model
  - min_df= 0.001, max_df=0.5
  - Stop words left in
  - 1172 words kept
- Tf-idf vectors passed through truncated svd
  - 200 Components kept
  - Explains 56% of total variance



(770, 850) empty tweet vectors after tfidf for (training, validation) sets (∼4%)

Binary Classification

- Six models chosen with default parameters
- Cross-validation training performed with 5 folds
- 30% of training samples set aside for testing

Estimator	Train Accuracy	Test Accuracy
K-Nearest Neighbors	0.783333	0.792444
Logistic Regression	0.774190	0.781444
SVC	0.822619	0.832889
Decision Tree	0.744524	0.751778
Random Forest	0.783048	0.792778
Extra Trees	0.790571	0.794000

- SVC used for other statistics on validation set

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

Evaluating Classification Performance

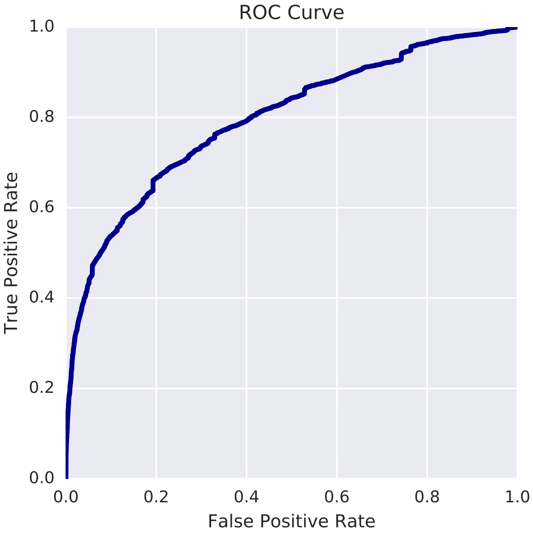
Performance is reasonably constant across models

- Large number of features even after truncated svd
- Likely reasonably linear dependence of class on features limits
- Limits probably due to remaining contamination of classes and low-uniqueness tweets as opposed to modeling error

Precision is good but recall is poor

- Precision = 0.9, indicating that the identified sample of video game related tweets should be relatively clean, though it could still be overwhelmed by a large number of irrelevant tweets in the case of a true twitter stream
- Further cuts can be performed to reduce the false positives
- Recall is only 0.6, indicating that only 60% video game related tweets in an unfiltered stream will be tagged

ROC Curve



Total AUC = 0.79

- The model can easily be optimized for a low false rate while retaining a non-negligible true positive rate
- Achieving a high true positive rate requires acceptance of a significant false positive rate
- Consistent with intuition from precision/recall

Refining Classification

Possible Improvements

- Large sample
  - Requires more processing power
- Include hashtags, and possibly usernames
- Include emojis
- Prune keywords for a cleaner topic set
- Better balance of tweets with various keywords
- Apply an a-priori cut on short tweets as “acceptable losses”

Prevalence of Keywords

Keyword	Percentage	Total
Zelda	35.056	8764
Overwatch	11.116	2779
Pokemon	7.452	1863
Minecraft	5.260	1315
Mario	5.196	1299
Halo	3.072	768
Sonic	3.052	763
Mass Effect	2.980	745
Resident Evil	2.592	648
Call of Duty	2.108	527

Some coherent method of aggregating tweets may result in significant improvements

Topic/Genre Modeling of Video Game Tweets

Once a tweet is identified as video game related, it is desireable to categorize it

Two distinct methodologies

	Clustering	LDA
Use Case	Genre classification	Topic modeling
Process	Genres are generated by clusering the tweets, then attempt to identify coherent genres by prevalence of keyword labels	Topics are identified by performing LDA on the entire corpus and identifying topics based on word prevalence
Prediction	Identify most likely genre by comparison to LSA vector	Identify most likely topic through comparison to LDA

Difficulties in Genre/Topic Modeling

Neither method produced meaningful initial results

- Clustering performed using LSA on corpus of unified tweets for each keyword
- LDA performed on original corpus of cleaned corpus
- Neither model yielded coherent categories
  - Moreover, neither model yielded consistent categories using different random seets

Too early to make conclusion on relevance of models vs. insufficient or insufficiently cleaned data

- Tweets from a large number of uses may simply not contain consistent language once keywords are removed
- Imbalance of classes probably damages genre clustering, and an iterative curation of terms may allow for meaning to be taken from topic modeling

Conclusion

Initial Binary Classification of tweets by topic was successful, with multiple models generating results with 75-80% accuracy

Many future directions may be explored

- Better cleaning and curation for improved classification
- More focused identification of types of tweets to be classified
- Refinement of genre/topic modeling to generate a sub-categorizatin procedure on tweets classified as topic related
- Application to other topics
- Testing using possibly-related keywords with hand-assigned classes (Civilization, Battlefied)