Youtube Machine Learning Project: Data and Methods

Dano Gillam

April 20, 2016

It is universally accepted that the internet is the best place to hold civil discussions
Abraham Lincoln

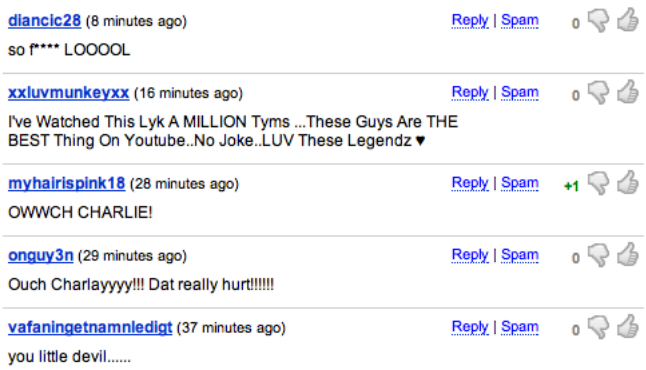
Many of the greatest writers of our day use commentary sections on Youtube videos to express their opinions on political issues. We can learn much by studying the literature of these aspiring authors. Their unique usage of adjectives may give us insight as to what the public opinion is on various topics.

1 the Data

I created a compilation of comments from on youtube videos pertaining to the presidential election. These videos were chosen from 4 News channels covering the presidential election.

1.1 Scraping Youtube comments

- I created a customizable scraper. This was the first scraper I've ever created of this magnitude. The scraper can accept single youtube video urls or youtube playlist urls. It opens a browser containing each video, scrolls down, and proceeds to click the "show more" button until it reaches the last comment. It then saves the html of the page to the local machine to await further processing. BeautifulSoup is then used to parse through



the collected html files and extract all information about the comments. It creates an array with 4 columns: $userID$, $userName$, $usersComment$, and $videoName$. I decided to break the scraping process into these two steps to limit the effects of poor internet connection while scraping. This is the dataset that will used in our analysis.

- See yt_scraper.py and youtube_playlist_save.ipynb for the implementation.
- I scraped 4 major News channels for videos pertaining to the presidential election.
  1. FOX
  2. ABC
  3. YAHOO
  4. CNN
- Over 750 youtube videos were scraped.
- 35,659 comments were collected.

1.2 Why this dataset is interesting:

Youtube is the second most popular website. (second only to Google itself) Assessing public opinion of candidates is a popular topic in statistics. Youtube commenters are a very unique subset of people and it is interesting to use their comments as a form of Polling.

2 Problems I wish to explore

2.1 I wish to discover public sentiment of presidential candidates using youtube comments.

- This is supervised problem. I want to see what words are grouped with the presidential candidates. Are they positive or negative?
- This is a classification problem where the topics are the presidential candidates.

- Background information is abundant for this problem. I knew the candidates and the issues and how they correlated.

2.2 I wish to see what topics dominate the comments of political videos.

- This is an unsupervised problem. I don't know what topics I'm looking for.
- This is a classification problem. I'm trying to classify words into numbered topics.

3 Exploration of the Dataset

3.1 Preliminary Analysis

I did various explorations of the dataset before I tried any technical approaches. See youtube_analysis.ipynb

- The online datasets used to classify words include:
  - neg-word-list.csv
  - pos-word-list.csv
  - sowpods.txt
  - stop-word-list.csv
- candidate sentiment analysis:
  - I analyzed the ratio of positive to negative words used in conjunction with a candidates name.
    1. trump: 0.539 (not including 'trump' as a positive word.)
    2. cruz: 0.399
    3. kasich: 0.637
    4. clinton: 0.538
    5. sanders: 1.078
  - The overall positive to negative word ratio was: 0.513
- User analysis:
  - Some users changed their user names over the course of their commenting: Of 6000 comments, 12 useres had multiple aliases.
  - Many words used by commenters are invented: The percentage of real words used was 85.3%

4 Technical description of my approach

I used two separate algorithms to separate the dataset into topics: Latent Dirichlet Allocation and Non-negative Matrix Factorization.

4.1 Latent Dirichlet Allocation

In natural language processing, Latent Dirichlet allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics.

The generative process is as follows. Documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. LDA assumes the following generative process for a corpus D consisting of M documents each of length N_i:

1. Choose $\theta_i \sim \text{Dir}(\alpha)$, where $i \in \{1, \ldots, M\}$ and $\text{Dir}(\alpha)$ is the [[Dirichlet distribution]] for parameter α

2. Choose $\varphi_k \sim \text{Dir}(\beta)$, where $k \in \{1, \ldots, K\}$

3. For each of the word positions i, j, where $j \in \{1, \ldots, N_i\}$, and $i \in \{1, \ldots, M\}$

: (a) Choose a topic $z_{i,j} \sim \text{Multinomial}(\theta_i)$.

: (b) Choose a word $w_{i,j} \sim \text{Multinomial}(\varphi_{z_{i,j}})$.

4.2 Non-negative Matrix Factorization

NMF is commonly used in text mining. We first create a document-term matrix with words weighted by their frequency in our set of documents. We then factor this matrix into a term-feature and a feature-document matrix. The term-feature matrix describes data clusters of related words. The feature-document matrix describes data clusters of related documents.

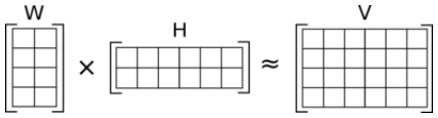


Figure 1: NMF

5 Implementation

*Warning. Due to the nature of the experiment, the following section has many explicit words. If this is offensive to you, please feel free to skip to the next section titled "Summary of Results" where I provide a summary of my findings.

5.1 Results

Using sklearn LDA and NMF implementations I discovered topics.

```
Topics in NMF model:
Topic #0:
trump vote stump supporters win way hitler love hate voting dump debate mr sucks wins hope
    ↪ support man won better
Topic #1:
people just don vote cruz america know want make think ted country good right obama say man
    ↪ going great shit
Topic #2:
fuck abc bitch fucking shit obama cnn shut ass racist piece cruz pope dumb guy yeah cunt
    ↪ rubio islam fox
Topic #3:
2016 prison cruz trump carson rand paul rubio sanders ted 2020 hilary maga ben latinos
    ↪ america 2024 dt make kasich
Topic #4:
hillary clinton prison vote win choice best voting jail liar republican better wins won way
    ↪ campaign corrupt sanders party democratic
Topic #5:
lol joke funny guy nice idiot xd comment scared true good did shit dude didn got little yeah
    ↪  hahahaha wtf
Topic #6:
president obama best choice united states running sanders worst great mr usa good vice
    ↪ america want going man hilary history
Topic #7:
donald hate trump duck racist love dump god world dick sucks man loves fucking war mexicans
    ↪ wait bitch drumpf ted
Topic #8:
like looks sounds don just look sound guy doesn hitler rubio act does dog carson real really
    ↪  feel cruz comment
Topic #9:
bernie sanders love vote win supporters feelthebern abc candidate socialist won new hilary
    ↪ wins voting video york media vs clinton

Fitting LDA models with tf features, n_samples=2000 and n_features=1000...
done in 15.399s.

Topics in LDA model:
Topic #0:
president sanders people think stupid war didn candidate states just believe government usa
    ↪ bitch feel state fact united business political
Topic #1:
people don want world just need make know like money think americans doesn america good
    ↪ fucking wall things shit change
Topic #2:
said ass yes islam paul family damn really rand listen ha wasn lady using lies jesus act
    ↪ kind attack exactly
Topic #3:
best muslims big office did funny democrats million born trumps religion watch human words
    ↪ fake maybe point romney day die
Topic #4:
hillary cruz clinton media ted news trump rubio abc shit like voting lying fox establishment
    ↪  bush joke wow debate cnn
Topic #5:
like just people white guy did hate don got god know look hell racist tell way talk oh idiot
    ↪  house
Topic #6:
trump 2016 say does old sure new lot carson supporters says race talking time just far hope
    ↪ john ben real
Topic #7:
bernie man fuck lol really stop good better china middle europe home goes class paid mexico
    ↪ poor away pope corporate
Topic #8:
obama country president going time years american isn black needs work job america truth
    ↪ question thing right mr countries like
Topic #9:
```

```
trump vote donald america great love win support people right isis women doing video party
    ↪ make gop won president woman
```

5.2 Summary of Results

NMF seemed better at finding groups than LDA. LDA seemed better at grouping sentences together in almost sentence-like ways.

5.2.1 NMF

- Topic #0 and #7 relate to Trump. He is grouped with words such as hitler, racist, vote, love, hate, support, mexicans, and war.
- Topic #4 relates to Hillary Clinton. She is grouped with words such as prison, vote, win, choice, best, jail, liar, republican, and corrupt.
- Topic #9 relates to Bernie Sanders. He is grouped with words such as love, vote, supporters, feelthebern, socialist, media, and win.
- Topic #2 groups all the explicit words together.
- Topic #3 groups all of the candidates together with the words 2016 and prison.
- In topic #6, Obama is grouped with "president united states"

5.2.2 LDA

- LDA groupings almost seem like sentences
- Topic #9 starts with "trump vote donald america great" which shows that trumps slogan "make America Great again" was picked up.
- Topic #8 groups Obama with "country american president black"
- In topic #7, Bernie was grouped with "china, europe, mexico, and pope"

5.2.3 Conclusion

We can see from the comments that none of the candidates for the presidency this year have very many positive comments made about them. Of course, Youtube tends to be biased towards haters. People with negative opinions are more likely to comment than those with positive opinions.

6 Rescources

6.1 My code

data_extract.py (my first attempt at scraping. classifies words as well.)
youtube_analysis.ipynb (preliminary analysis. imports data_extract.py)
yt_scraper.py (a better scraper class)
youtube_playlist_save.ipynb (implements yt_scraper)
lda_youtube.py (runs NMF and LDA on my youtube dataset. using sklearn)

6.2 Datasets

neg-word-list.csv
pos-word-list.csv
sowpods.txt
stop-word-list.csv
youtube_comments1.csv (The main dataset)(my collected comments)