GROUP 12
Shiv Chandra Kumar

Peng Meng

Marc Malinowski

Chris Dayal

Important Notes:

1. We used Python libraries for most of the task except the NLTK library for finding tokens.
2. We have built a web crawler to get the Categories and Nominees information from Imdb website.
3. This code is generic for all the golden globes award functions across the years. But currently our code works only for 2013 and 2015 data as we have built crawler specifically for these years.

A small write up about the approaches which we considered and eventually end up implementing is as follows:

Golden Globes Project Write Up

Throughout the process of writing our program we encountered many approaches and obstacles. As we attempted to caveat for accuracy and efficiency, we worked together to develop and build what we perceived to be the most balanced and generic program to fulfill the project goals. We looked for trends within the tweets to design a coding approach that was sufficient to the project goals, as well as efficient. To reiterate, the

basic goals of the program were to build a program that identified the hosts, winners, awards, presenters, nominees of an award show, and matched the winners to the presenters along with the nominees. Our program was to include a interface that returned a dictionary object entailing the aforementioned information. In addition to these basic goals, we also tackled some of the fun goals such as the red carpet festivities and who was best and worst dressed. We also delved into writing code for identifying key events that transpired throughout the evening such as the various sentiments and humorous highlights.

In the beginning as we began to tackle this project, as a group we drew up various blueprints to how we felt the best approach to the project would be. Our initial questions regarding the given json file with tweets were all characterized by how were we going to get our application to recognize what information was important and relevant to the goal of the project. It is inevitable that throughout a twitter session during an award show that there would be some irrelevant tweets offering little or no information at all. Another concept that we kept in mind, was that this application was going to be generic for award shows similar to the Golden Globes, so we did not want to characterize all of the categories. Award shows such as the Golden Globes follow a certain schema of awards, such as best picture, best actor, best director, and more. Keeping these notions in mind, we felt the best approach was to use a hierarchical tree structure approach, initially filtering and matching the data to what we wanted our output to be.

GROUP 12

As we began to build our program we immediately decided to filter the plethora of tweets to find the significant information. By searching the original code and filtering them for our specified key words which would be applicable to other similar award shows, we were able to instead of searching 174,000 tweets for arbitrary information, search about 9,000 tweets that entailed some sort of information that was relevant to our purpose. After doing this, we proceeded to use the remaining of the tweets as our source for the various goals. After this initial sort, we categorized and defined the awards. We then coded for the key words that would define each category within the tweets. For example, for the award of Best Motion Picture Drama, we used the key words “Best picture, drama”. We did this for the other awards as well because we knew that these common phrases would be used in the tweets to define the winners. For the nominees, we used a webscraper to pull the data to increase the efficiency of the information retrieval as opposed to having to search through the tweets, which were already filtered, and we really did not feel that the best approach was to use search the full batch. To account for the the hosts we used the host words “host”, “hosting”, “hosts”, and “hosted”, to filter the tweets for the appropriate results. This approach of dissecting our set of tweets for key words and attributing the result to the most common value within these tweets was used for some of the fun goals such as best dressed and worst dressed.