Golden Globes Project Sheet

Your team has been given the task of creating a system to extract and analyze information from Twitter during future Golden Globes events.

The system you design must be capable of working on an entirely new, unknown corpus of Golden Globe tweets from a different year. The ultimate goal is a system that can process data for any award show (not just the Golden Globes) as it is coming in, providing new insights and results as the ceremony progresses. However, in this iteration a system that analyzes the data immediately following the awards ceremony is an acceptable stepping stone.

Given the ultimate goal of live insights, the system may pre-process information available prior to the Golden Globes. If that pre-processing takes some time to run, it won't cause any problems. But your extraction of information from the incoming tweets during the event needs to run rapidly. For that reason, processing the corpus of tweets at a demo shouldn't take longer than a minute.

Likewise, if there is any information that is the same every year for the Golden Globes, that information does not have to be extracted automatically.

You will make an appointment to meet with your boss, Larry, and his assistant, Miriam, on February 11, 12, or 13. At the appointment you'll do an informal live demo of your system. Larry will want you to run your program in front of him, so the program should display its computed results.

Resources

- A corpus of tweets streamed during the 2015 Golden Globe Awards on the following keywords: 'gg','golden globes', 'golden globe', 'goldenglobe','goldenglobes','gg2015','gg15','goldenglobe2015','goldenglobe 15','goldenglobes2015','goldenglobes15','redcarpet','red carpet','redcarpet15','redcarpet2015','nominees','nominee','globesparty','g lobesparties'. This corpus is about 1 GB compressed. You can download it at http://bit.ly/1EFJ5pn
- A smaller corpus of tweets streamed during the 2013 Golden Globe Awards; this corpus contains stripped down tweet objects. This can be found in the Golden Globes Project section of the Resources section.
- A copy of the autograder program, which will assess how well you did on the basic tasks. This may be useful in determining efficacy as you refine your code. *There will be a note when this is available.*

* These items can be hard-coded, because they are known before the ceremony starts. However, if you hard-code the data, we expect you to invest that time in achieving more with the fun goals.

◆ These items can also be automatically extracted from the web. Depending on the method employed, this may be considered more like hard-coding, or more like the original basic task. It may even be harder. Feel free to argue either way.

Requirements

Basic Goals

- Identify:
  - Hosts*⬩
  - Winners
  - Awards*⬩
  - Presenters
  - Nominees*⬩
- Match awards to:
  - Winners
  - Presenters
  - Nominees*⬩
- Your program must include a text interface that, when run:
  - Provides an interactive menu with options the user can select to see your results; selection can occur through typing a number or letter and then pressing enter.
  - Displays the results as formatted legible text – just printing your Python objects is not acceptable.
- Your system must include a program interface that will return a dictionary object containing the basic information listed above. This will be used by the autograder to assess your system's performance on the basic goals.
- A readme file citing the libraries you used (beyond standard Python libraries), the repositories you consulted for inspiration, and describing what you did to make the system adaptable for future years or for other award ceremonies.

Fun Goals

Choose at least one:

- Red carpet – For example, determine who was best dressed, worst dressed, most discussed, most controversial, or perhaps find pictures of the best and worst dressed, etc.
- Humor – For example, what were the best jokes of the night, and who said them?
- Parties – For example, what parties were people talking about the most? Were people saying good things, or bad things?
- Sentiment – What were the most common sentiments used with respect to the winners, hosts, presenters, acts, and/or nominees?
- Acts – What were the acts, when did they happen, and/or what did people have to say about them?
- Your choice – If you have a cool idea, suggest it to the TA! Ideas that will require the application of NLP and semantic information are more likely to be approved.

A few bonus points may also be awarded for:

- Creating a graphical user interface
- Being the most accurate team(s) for any of the basic, autograded tasks. Note that hard-coding this information (and possibly, scraping) excludes you from this bonus.

* These items can be hard-coded, because they are known before the ceremony starts. However, if you hard-code the data, we expect you to invest that time in achieving more with the fun goals.

⬩ These items can also be automatically extracted from the web. Depending on the method employed, this may be considered more like hard-coding, or more like the original basic task. It may even be harder. Feel free to argue either way.