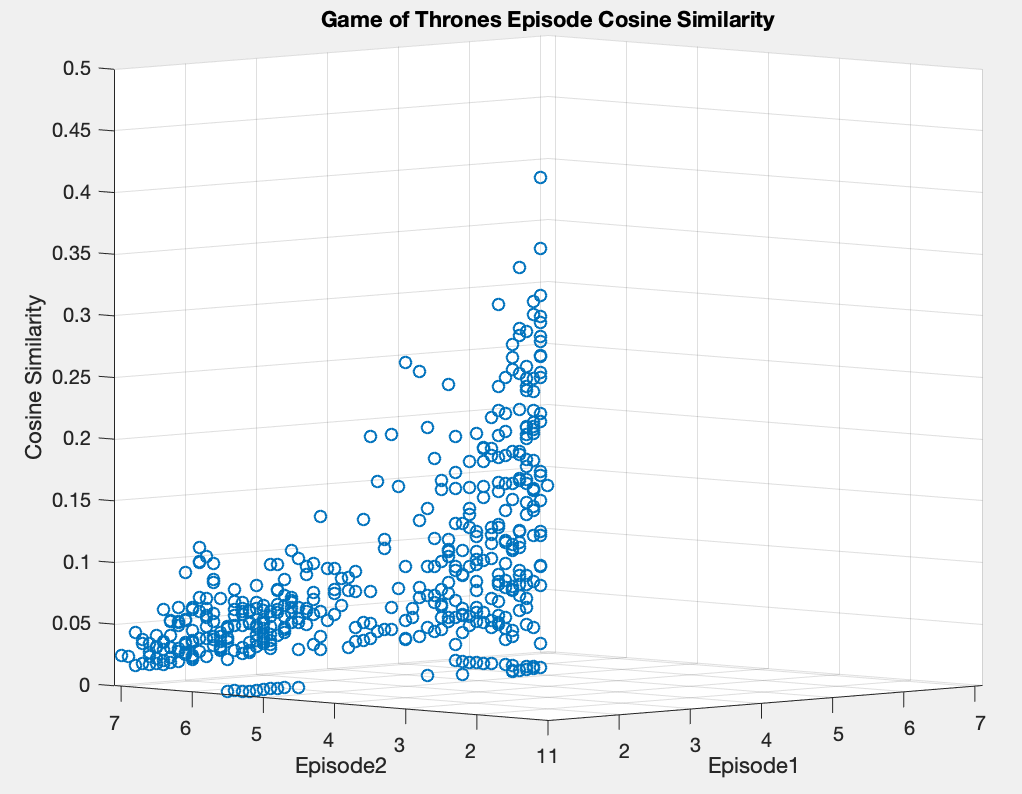
Empirical Analysis/Visualization Writeup



**Description of data:**

Depicted is a 3-dimensional scatter plot that visualizes the Cosine Similarity of all of the Game of Thrones episodes when compared to each other episode. The code for the visualization is attached in the “Cosine Similarity Plot” and the code for the calculations is attached in the “javaAnalysis” folder and is a slight deviation from the code given in HW4. The “output” file in the “javaAnalysis“ file is a CSV that allowed for easy import into MATLAB.

**Hypothesis:**

We began this assignment with the belief that episodes in the same season/with similar release dates would have a higher cosine similarity value than those that were in different seasons or were released further apart. This belief was largely due to the assumption that many of the character involved and the topics that were spoken about in each episode were likely to have a higher correlation in episodes within a tighter timeframe (in relation to the show and/or real life).

**Analysis:**

To begin, notice that the points are plotted on the X and Y axis as double values. These correspond to “season.episode” values, so 1.1 is season 1 episode 1 etc. The first and most obvious thing that is seen when examining the image is the left-skewed nature of the graph. This is simply due to the code that was used to compare the files (HW4 code). The program is a double for loop that goes through all of the files that have NOT YET been compared. This means that for each episode, the episode is only compared to those which come after it in the list of all documents. The second thing that can be seen from viewing the graph is the fact that the values down the middle are generally much larger than those on the side. This observation went to confirm our hypothesis, that episodes within closer time frames would have higher cosine similarity values than those more spread apart. The values down the middle, since the X and Y axis go from 1.1 to 7.1 (all seasons for which transcripts were available), are those which have similar values, like 1.1 and 1.2 etc. There is a clear distinction between the values in the middle of the graph and those on the edge that clearly affirms our prior beliefs.