Michael Root

Data Mining

Project Proposal

Project Title: Text Mining the Messages in Movies

Who has done this before: As far as I can tell, no one has done an in depth analysis of movie scripts. The closest thing I could find was a study that was called “Exploiting Structure and Conventions of Movie Scripts for Information Retrieval and Mining”. The main goal of this paper was an explanation of how the researchers wanted to extract the various information of things like character direction, camera control, and blocking, and annotate it. As far as who has done text mining as a whole, that is a relatively well-known endeavor, commonly described as “the process of extracting interesting and non-trivial patterns or knowledge from unstructured text documents”. My thought with this is since text mining is a well-established practice, why not apply it to movies? Since all movies stem from one thing in common, written text as a script, if those scripts can be attained and properly sifted through then someone could easily find common phrases or key words shared in big hits and big flops.

Google Scholar Paper on Text Mining:

<http://www.ntu.edu.sg/home/asahtan/papers/tm_pakdd99.pdf>

RIT Library journal containing paper on exploiting script structure:

<https://link-springer-com.ezproxy.rit.edu/chapter/10.1007%2F978-3-540-89454-4_27>

Challenges: The biggest challenge I can see, as no one I know of has done this before, will be physically making the text files containing movie scripts into pure text or CSV files for me to use. As the paper about Exploiting Scripts Structure talks about, the “pure text” files of movie scripts that are available online contain a plethora of information that is not really relevant. As of now, I’m just interested in character dialogue, so pulling out lines talking about camera panning or character entrances will just weigh down and add unnecessary data to my word counts and phrase counts. Once I can make a way to easily conform full blown scripts into pure dialogues or just manually cut out unnecessary drivel then I don’t think it will be difficult to manage. One challenge I can see though is that if there are adlibbed parts of a movie, that won’t appear in the scripts so I won’t have records of them.

Data: The only data I’ve gathered so far are freely available movie scripts from different sites that post full pages. I don’t think are any ethical ramifications, as I’m not stealing scripts for my own movies or papers, I’m just looking at what they have and presenting it in a way that looks into the smaller details of the words and phrases used. The scripts could be not 100% accurate though, as they are being obtained from different websites that could be altered in some way. Though, given that all the scripts are written out directly to the sites, for all anyone knows there could be randomly inserted or removed bits from different movies.

Measuring Data: To be honest I’m not quite there yet, I don’t know what sort of algorithms I will want to run on this data. The first thing I’m thinking about is how I want to organize it. I know I want to do individual word counts for each script. Also take “supercut” phrases, or really popular movie phrases and see when/if they occur in what film scripts. Something else I’m interested in is maybe go through the script and keep track of how many lines individual characters have. Once I have all of this info sorted I’m not sure what kind of clustering or classification I can or should use. Any advice on that would be welcome.

Other RIT books:

Mining Text Data: <http://library.books24x7.com.ezproxy.rit.edu/assetviewer.aspx?bookid=54151&chunkid=979295467&noteMenuToggle=0&leftMenuState=1>

Movie Review Mining and Summarization:

<http://dl.acm.org/citation.cfm?id=1183625>