## Homework 5

1. Compile a list of static links (permalinks) to individual user movie reviews from one particular website. This will be your working dataset for this assignment, as well as for assignments 7 and 8, which together will make up your semester project.
   1. It does not matter if you use a crawler or if you manually collect the links, but you will need at least 100 movie review links. Note that, as of this writing, the robots.txt file of IMDB.com allows the crawling of user reviews.
   2. Each link should be to a web page that has only one user review of only one movie, e.g., the user review permalinks on the IMDB site.
   3. Choose reviews of movies that are all in the same genre, e.g., sci-fi, mystery, romance, superhero, etc.
   4. Make sure your collection includes reviews of several movies in your chosen genre and that it includes a mix of negative and positive reviews.
2. Extract noun phrase (NP) chunks from your reviews using the following procedure:
   1. In Python, use BeautifulSoup to grab the main review text from each link.
   2. Next run each review text through a tokenizer, and then try to NP-chunk it with a shallow parser.
   3. You probably will have too many unknown words, owing to proper names of characters, actors, and so on that are not in your working dictionary. Make sure the main names that are relevant to the movies in your collection of reviews are added to the working lexicon, and then run the NP chunker again.
3. Output all the chunks in a single list for each review, and submit that output for this assignment. Also submit a brief written summary of what you did (describe your selection of genre, your source of reviews, how many you collected, and by what means).