This assignment is due by: 2359 4 Nov. 2015.

Sarah is stepping out from time to time (with Hasta's blessings), for a Girl's Night Out at the local winery. Sarah is part of a trivia team, and they compete for prizes and recognition at the winery. Like all teams, some members are really strong at somethings, while others are stronger at something else. To no one's surprise, movies are Sarah's strongest point. Hasta revels in Sarah's victories and commiserates in her defeats, but he understands Sarah's need to get out and spend time with her friends. Sometimes Sarah thinks that Hasta just wants her out of the house so he can play with his CouchDB.

Sarah has been thinking about the number and type of questions the quiz master has been asking about movies. She is trying to understand where the master gets his questions, so that she can focus her attention in the right way to help her team. After a particularly narrow victory, Sarah settles into the CouchDB next to Hasta, and explains her idea.

Sarah already knows how many movies have been made per year since the first movie was made (really; she knows that Hasta can tell her). She has researched the Internet Movie Data Base and has found a lengthy list of trivia facts on movies and a whole lot of other things. So, she has identified all the sources of data that her beloved Hasta can manipulate to give her the graphics she is after. These graphics will allow her to focus her training and attention on those years where with the greatest likelihood of knowing all the trivia there is. This knowledge will help Sarah's team on to victory and glory. (Hasta is trembling on the couch.)

Sarah needs two graphics. They are:

1. A histogram of trivia entries there are per movie

2. An X-Y plot with two curves. The X axis has the years from the earliest to the latest movie. Both Y curves are percentages. The first Y curve is the number of movies made that year divided by the total number of movies. The second Y curve is the percentage of movies made that year that have trivia entries.

Here are some of the files that Hasta will need[1]:

- movies.list.gz — a consolidated list of movies and TV shows. The file has some header stuff. TV show line entries start with a "(and are of no interest). All other lines may refer to a theater movie. The last field in each line is a year, the next to last field may be a formatted year. If the next to last field is a formatted year that matches the last field, then the line is a movie. Otherwise it isn't. There are also non-ascii characters in the file. Some number of tab characters (1 or more) separate the fields. (There are roughly 700K movies.)

[1]The datasets are available at:cs695-nosql:/rawdatas/MovieFiles

- trivia.list.gz — a list of bits of trivia on films. Each entry block begins with a # followed by a space, followed by the name of the movie, followed by a space, followed by the production year in parens. It looks as if each new trivia entry starts with a dash. There are other "films" in the file, so you'll need to be sure to get the correct ones.