CS688 Final Term Project Overview

Expectation:
You will submit a written report and give a presentation after you have submitted the report.

- Similar to the weekly assignments, the written report should explain what you have learned and how the work you did addresses the questions/tasks for the project you chose. It should not simply be a copy-and-paste of code and output. The written report is worth **80** points. These points are explained on the next pages.
- Each presentation should take no more than **10** minutes. Generally, your presentation will not include all of the detail in the written report. The presentation is worth **20** points.
  -2 pts per minute overtime starting after 11 minutes.
  -5 pts for unorganized presentation or **-15 pts** for presentation with no slides.
  Between -2 to -12pts if in the facilitator's judgment your presentation is not clear or contains incorrect information.

Projects and data set to choose from:
Consider the following term project topics and datasets and select one that seems more interesting to you. **For more details see the next pages of this document.** If you believe there is more than one way to interpret the instructions, select one of the interpretations and explain what you have decided to do and why.

1. **The Text Analytics term project:**
   Dataset: By date version of Newsgroups data set provided on Blackboard
   - Use R to load the content from a large collection of text files.
   - Perform specific text analytics tasks such as classification of documents by the content analysis of the subject line.
2. **Searching and Ranking a set of given web pages term project:**
   Dataset: Your choice of 10 URLs (suggestion 2-3 sets of similar subject URLs)
   - Use R to download the web pages and extract the text content.
   - Perform preprocessing and content analysis to rank these pages for similarity.
   - Use data visualization techniques to display the results.
3. **Twitter stock market sentiment analysis term project:**
   Dataset: Your choice of 6 stocks, 3 largest gainer (loser) stocks for the day. (suggestions: http://finance.yahoo.com/; http://www.google.com/finance)
   - In your presentation include the stock charts and quotes.
   - Use R to interface with the Twitter API and perform stock market sentiment analysis.
4. **Sports Data Analytics term project:**
   Dataset: You will use data from the web and you will need to scrape it. You may also use data from the R SportsAnalytics library, or data from another R sports library.
   - Highlight interesting data about teams and players in your favorite sports league.
5. **The Web site scraping term project:**
   Dataset: A website of a movie theater of your choice.
   - Use R to scrape movie theater show times from a movie theatre's web site of your choice.

See next pages for more!

Preparing and analyzing the data:

- If you use data of your choice indicate and document clearly what you have used, especially if the data content changes over time (i.e., stock quotes, social media feeds, etc.).
- Provide code that enables us to reload your data either from the internet or from files created by an R function such as save(). If you save files with R, please provide copies in your ZIP submission and make sure the R code you submit is able to read and process the files.
- Document the steps to import the data set into R.
- Explain all preprocessing you did both before and after importing into R.
- Please keep the naming convention of the R objects if indicated.

Data Visualizations:

- All of the projects have elements (sometimes extra credit) requiring data visualizations.
- All data visualizations that you create must follow good data visualization principles. You are held accountable for ALL of the data visualization principles given in Module 6. This includes, but is not limited to:
    - All data visualizations that you create must have meaningful titles and axis labels.
    - If you use different colors, line styles, plot symbols, etc., you must also create a legend.
    - Ensure that your labels are not cut off and that you are not missing labels.
    - Ensure that all information added to the plot is clearly readable.
- Making good data visualizations usually requires you to adjust plot settings. That is, don't use the defaults!

Submitting the Project:

Consider submitting these files:

- Slide deck: PPT, PDF, or similar. Rmd is acceptable only if rendered to a presentation format (see https://rmarkdown.rstudio.com/lesson-11.html).

- Written report: DOC or PDF preferred. With facilitator permission, you may be able to submit HTML, Rmd, or similar formats. Be sure not to put all the R code in the body of the document. Highlight only key code segments in the body of the document and consider putting the full code in an appendix.
- R code: only if the code is not in an appendix of your written report
- Any data sets that allow us to *reproduce your results exactly*. (Put these in a ZIP and upload.)

The term project submission is due at the same time as Assignment 6 and the Quiz. You will schedule the presentation on the Tuesday, Wednesday, Thursday, or Friday evening after submitting the term project, Assignment 6, and the Quiz.

CS688 Final Term Project Option 1

The Text Analytics term project:
Dataset: By date version of Newsgroups data set provided on Blackboard

If you haven't already done so, download this dataset and save it on your computer. As you already know, the data is separated into one test and one train folder, each containing 20 sub-folders on different subjects. Choose any 2 subjects to analyze. You could choose, for example, sci.electronics and talk.religion.misc. Continuing this example, consider "talk.religion.misc" as **positive** event and "sci.electronics" as **negative** event when you create the confusion matrix and calculate the F Score.

a) (5 points) For each subject select documents for training from the train folder and documents for testing from the test folder. You must select at least 200 documents for testing, and you must select more documents for training than you select for testing. That is, do not select the same number of documents for both training and testing as you did in Module 3.

b) (5 points) Obtain 4 corpora – train and test data for each of the two subjects you've selected. You must create Corpus objects. Other methods such as tidytext (Module 5) are not permitted here.

c) (15 points) Subject line preprocessing. Implement the following preprocessing in each of the 4 corpora separately. This step will be challenging (you haven't done anything like this in this class!):

- Identify the fields "From", "Organization", and "Subject" (use grep() in a for loop) and save them as R objects. Please keep the same order and naming convention (i.e. Subject1.Train, Subject1.Test, Subject2.Train, Subject2.Test).
- What is the number of unique (you can use the unique() function) email addresses in each of the 4 corpora?

d) (5 points) Note that the R object "Subject" is an R list. Form 4 corpora (use the VectorSource() function) from the 4 "Subject" lists and combine them. Pay attention to the order in which you combine so you can reverse this later.

e) (10 points) Implement any other pre-processing you believe is necessary. Then create the "Subject" Document-Term Matrix. At a minimum, you must eliminate 1-letter-long words and ensure all words included in the DTM appear at least 5 times. You can use more strict criteria if you can defend it.

f) (28 points) Classify and display results using knn().

- Split the "Subject" Document-Term Matrix into a train DTM and a test DTM.
- Make a vector of length 400 which indicates the correct classification of the documents in the train partition. Use appropriate abbreviations for the tags.
- You may, if you wish, choose to run knn() multiple times in an attempt to improve classifier performance (e.g., get a better answer).
- For purposes of this project, classifier performance is measured by calculating both the F-score and the percentage classified correctly.
- Display these classification results as a R dataframe using methods similar to those shown in Module 3.

- Obtain the confusion matrix and explain the difference between the False Positives and the False Negatives.

g) (12 points) What percentage of documents in the test partition that have been predicted correctly? Do you think it is easier or harder to predict using the "Subject" DTM (compared to using a DTM based on the full text of the e-mail like you did in Assignment 3)? Why do you think so?

h) (For up to 10 points extra credit) Create ONE appropriate data visualization that compares classification effectiveness for the two groups in the test data set. This could be a visualization of the confusion matrix or some other plot.

CS688 Final Term Project Option 2

Searching and Ranking a set of given web pages term project:
Dataset: Your choice of 10 URLs (suggestion 2-3 sets of similar subject URLs)

Perform these web analytics tasks by a download of 10 web pages. It may be easier to get 10 web pages from the same web site. Scrape only the text content (remove HTML, CSS, and other non-text code), perform preprocessing and content analysis to rank these pages for similarity.

a) (16 points) Create an R code that will:
   - List the names of 10 web pages of your choice that you would like to rank by the similarity of their content.
   - Retrieve these web pages from the web.

b) (16 points) Create an R code that will:
   - Extract the text content of all the 10 web pages.
   - Please specify which libraries, and functions you used to accomplish this.

c) (16 points) Create a single corpus from the text content of all the 10 web pages. You must create a corpus. Other methods such as tidytext (Module 5) are not permitted here.
   - Save the corpus with the name "Web.Data.Corpus" as R objects.
   - Implement preprocessing (clearly indicate what you have used)
   - Create the Document Term Matrix

d) (16 points) Implement hierarchical clustering of your web pages.
   - Specify which library and clustering algorithm you are using.
   - Specify which distance measure you have used.
   - Create a dendogram plot of your clustering.
   - Interpret the dendogram. What does it mean to you? Does it make sense? Why or why not?

e) (16 points) Visually display the 75 most frequent words in your Document Term Matrix using the wordcloud library.

f) (For up to 10 points extra credit) For the five (5) most frequent words in your DTM, create ONE appropriate data visualization that shows how frequently these five words appear (or perhaps don't appear) in each of the 10 web pages. Most likely you will use a categorical *x* axis with 10 levels (one for each web page) and a place the frequency count of the words on the *y* axis.

CS688 Final Term Project Option 3

Twitter stock market sentiment analysis term project:
Dataset: Your choice of 6 stocks, 3 largest gainer (loser) stocks for the day. (suggestions: http://finance.yahoo.com/; http://www.google.com/finance)

Use R to interface with the Twitter API and implement sentiment analysis on two (2) different sets of three (3) stocks. For the first set select the 3 largest gainer stocks for that day, and for the second set select the 3 largest loser stocks for that day. You decide how to identify the largest gainers and the largest losers and need to explain what you chose as part of your project report.

Create an R code that will:

a) (16 points) Search for the 100 tweets associated with each of the three stocks in each set.
   - You have many options for searching: cashtag, company name, tweets mentioning the company's username, and other options. Whichever option you pick, defend your choice.
   - You will combine all the gainers into one set of 300 tweets (find a way to keep track of the fact that tweets 1-100 belong to stock 1, 101-200 belong to stock 2, 201-300 belong to stock 3)
   - You will combine all the losers into another set of 300 tweets (again keep track of which tweets belong to which stock)

b) (40 points) Do the following:
   - Create 2 corpus objects (or tidy text objects) for the gainers and losers data.
     - Save the 2 objects to files and explain in your report how to reload them (in case your grader needs to reload the tweets you downloaded).
   - Use the necessary pre-processing transformations described in the lecture notes.
     - Implement the pre-processing as a function that takes a corpus and returns a pre-processed corpus.
   - If you choose to create corpus objects, create the document-term matrix for each set. Name them dtm1 and dtm2. (This task does not apply if you choose to use tidy text objects.)

c) (8 points) Find the most frequent terms from each set. Show a word cloud for each set.

d) (16 points) Using the positive and negative word lists, compute the sentiment score (as described in the lecture) for all the tweets for each gainers (losers) set. Were the tweets about the 3 largest gainer stocks for that day characterized by a positive sentiment, and the tweets about the 3 largest loser stocks for that day characterized by a negative sentiment?

e) (For up to 10 points extra credit) Create ONE appropriate data visualization that shows the stock prices and/or the change in stock prices for the stocks and day you selected for this project.

Note that a similar kind of analysis (on a large scale) published in 2010 under the title "Twitter mood predicts the stock market" (http://arxiv.org/abs/1010.3003) brought to the authors a multimillion dollar fortune.

CS688 Final Term Project Option 4

Sports Data Analytics term project:
Dataset: You will use data from the web and you will need to scrape it. You may also use data from the R SportsAnalytics library, or data from another R sports library.

Select any sport league of your choice (NBA, NFL, MLB, NHL, MLS, or any similar foreign sport league). Before selecting this project, be sure that you are able to scrape the data required throughout this project. You must accomplish the following:

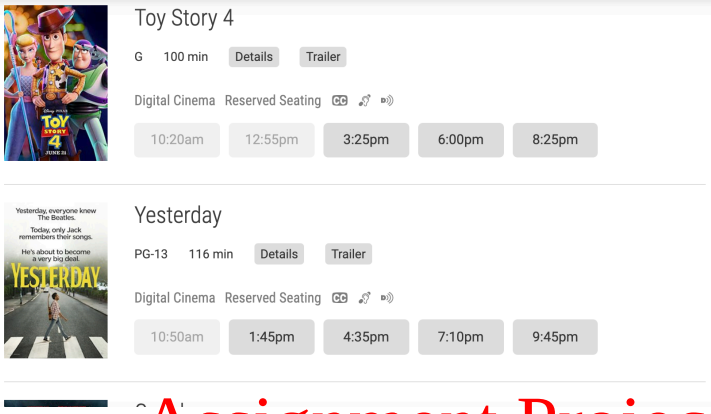
a) Select a league and a season (for example, NBA 2016-2017, or MLB 2018, or any other league and season of your choice),
b) (16 points) Load player statistics into R.
   - You must include all players from your favorite team and at least 40 additional players from one or more other teams. If it is easier, you can load all the players who played in the league. Depending on the league you choose, you may find that you will have to scrape from multiple web pages of data and combine tables together. You may have to check multiple sources to find a web site that provides pages that you are able to scrape.
   - You must collect the following data on each player:
     - Athlete name
     - Team(s) the athlete played for
     - At least three meaningful stats (such as points or goals or saves, minutes, assists, etc.)
c) (12 points) Subset the data from part a to include only players for your favorite team. Find the players who are "best" for at least three meaningful stats.
d) (16 points) Show 5 teams for the season you selected that have the most wins (you can also use ranking points in sports like hockey and soccer that give some points for wins and some other points for ties) in descending order. You will need to do some web scraping to get this information. For example, if you choose the NBA, you can follow landofbasketball.com as shown in the modules. Note many leagues are divided into conferences or divisions (AFC and NFC, Eastern Conference and Western Conference, AL and NL, etc.). If your league is divided, please ensure that you include teams from all conferences and all divisions.
e) (20 points) Create at least FIVE different data visualizations that highlight the strengths and/or weaknesses of the teams and/or the players on the teams. You must use at least THREE different chart styles (bar, line, box, scatter, etc.). Use the data from the part a and/or part b datasets. Explain why these charts are relevant.
f) (16 points) Use a mapping function (perhaps gvisGeoChart()) to display the location on a map the home locations of the last 10 champions of the league you chose in part a. Again, you will need to do some web scraping from landofbasketball.com or website appropriate for the league you have chosen.
   - You might plot fewer than 10 locations, if the same team has won the championship multiple times.

- If your sport has a system of promotion and relegation, you can choose to plot locations of champions or instead plot locations of the teams that have been promoted or relegated.
- You may have to convert the team names to locations that your mapping function recognizes. For example, you might convert “Cleveland Cavaliers” to “Cleveland” or “Colorado Rockies” to “Denver”. Double check that the names you use are recognized correctly. In some cases you may have to look up and use the latitude and longitude instead of naming the place (see Module 6).

g) (For up to 10 points extra credit) Perform any one additional challenging activity related to the data set. The amount of extra credit you can earn depends on how challenging the additional task is and how correct that work is.

CS688 Final Term Project Option 5

The Web site scraping term project:
Dataset: A website of a movie theater of your choice. Start with a page that shows the showtimes for **all** movies playing at that theater, similar to what is shown below. To complete part d below, it is **required** that you be able to click on the name of a movie (for example, clicking on "Toy Story 4" should take you to a new web page focused on that movie).



Use the R URL connectivity to a website to access this page. Please take screenshots of every web page used in your project to confirm your web page scraping tasks. For more reference regarding this you can review Module 4 and/or ask your facilitator.

Write an R code that will accomplish the following tasks:

a) (20 points) Identify the show times web page from your movie theater website
   - Access the specific movie theater's show times web page from your R code.
   - Import the HTML content of the show times web page into an R object called "AllMovies". Save this content to a file and explain in your report how to reload it (in case your grader needs to reload the show times as they may change between after you submit your project but before your grader can grade it).
   - Illustrate that step with a screenshot of the show times webpage showing the movie list currently playing in the theater.
   - Describe what packages or functions you have used to accomplish this task.

b) (20 points) Scrape the "AllMovies" web page to extract the list of movies currently playing.
   - Create an R list, vector, or data frame called "MovieTitles" containing all movie titles currently showing in the theater. Most likely you will get this by scraping the "AllMovies" HTML. Using the above screenshot as an example, your "MovieTitles" would include "Toy Story 4" and "Yesterday" together with other movie titles.
   - Include relevant screenshots of the HTML tags used to accomplish this task.
   - Describe what packages or functions you used to accomplish this scraping task.

c) (20 points) Choose ONE of the movie titles and scrape the "AllMovies" web page to extract the actual times at which that movie is playing.
   - Create an R list, vector, or data frame called "MovieTimes" containing all show times for that movie. Most likely you will get this by scraping the "AllMovies" HTML. Using the above screenshot as an example with "Toy Story 4", you should get "10:20am", "12:55pm", "3:25pm", "6:00pm", and "8:25pm".

- Illustrate that step with a screenshot of the show times webpage showing the chosen movie show times.
- Describe what packages or functions you used to accomplish this scraping task.

d) (20 points) For each movie in the R object “AllMovies”, determine the page address (URL) for the pages that would come up if you were to click on the titles of movies showing in the theater.
   - Create an R list, vector, or data frame out of them called "Page_Links" and save it to a file. Depending on how the site codes the HTML, this may or may not include the leading http:// or https://, may or may not include the server name, and may or may not include query parameters (? followed by a list of arguments).
   - Create another R list, vector, or data frame which contains a shortened version of the URL stored in “Page_Links”. This should contain only the last part of the URL, for example, “https://www.cinemark.com/toy-story-4?showDate=2019-07-27” becomes “toy-story-4” or “toy-story-4?showDate=2019-07-27” (your choice whether to include the query parameter, if there even is a query parameter for the movie theatre you are browsing). You might name this object “Short_Page_Links”.
   - Use a hierarchical visualization technique (perhaps gvisOrgChart()) to show the relationship between the web pages that were scraped. This is a network map with the “AllMovies” page as the parent node, and each of the names stored in the “Short_Page_Links” object.

e) (For up to 10 points extra credit) Open each web page in the “Page_Links” R object and scrape particular text content of your choice (such as review or synopsis) from this retrieved web pages in the "Page_Links" R object. Then complete the following tasks.
   - Create a corpus or tidy text for the text you are able to scrape. (Maintain the ability to identify which “Page_Links” contained which text.)
   - Implement preprocessing (clearly indicate what you have used)
   - Save the corpus to a file and explain in your report how to reload it (in case your grader needs to reload the corpus).
   - Create a Document Term Matrix
   - For the five (5) most frequent words in your DTM, create ONE appropriate data visualization that shows how frequently these five words appear (or perhaps don't appear) in each of the web pages in the "Page_Links" R object. Most likely you will use a categorical x axis with a separate level for each web page and a place the frequency count of the words on the y axis.