Assigned: 03/29/2021
Due: Tues. 04/13/2021, 11:59pm

This project will cover topics of dimensionality reduction and text mining.

Submission Requirements:

This assignment is split into two main parts.

Part A:
You will be asked to walk through methods **manually**, that is not using the available functions to perform clustering analysis. You can use R, Matlab, or Python for plotting of the data when asked and for support in calculations (compute distances, averages, etc.). But, the process should be done manually with your answer presented in **typed form**.

You may then prepare your solution to Part A as a separate document using Google Docs, Word, LaTeX, with computer generated plots inserted. Note, if you prepare your solutions in Rmd, Sweave, or Python notebooks, your solution to Part A can also be included in that document using Markdown or LaTeX to enter your answers (not a separate document).

Part B:
You will again have a final document submission that includes text responses to questions, tables, R/Matlab/Python code used to calculate answers, and figures.

Formatting of submissions will follow the same approaches as those used in prior assignments. Namely: Matlab + markup, Matlab LiveScript, Rmd, Sweave, Jupyter notebook, Colab notebook. Any other packages or tools, outside those listed in the assignments, examples, or those for reproducible research should be cleared by Dr. Brown before use in your submission.

For this project you are allowed to work in **groups of up to 3**. All students must sign-up on Canvas into a group (e.g., P5-Rho, P5-Tau, etc.).

Name your main submission files as *P5_GroupName*, create a zip-file called *Project5_GroupName.zip* and submit on Canvas. For example, if I was using R, and member of GroupChi, I would submit either:

- *P5_GroupChi.Rmd*, *P5_GroupChi.pdf*, or
- *P5_GroupChi.Snw*, *P5_GroupChi.pdf*

Questions:

PART A - WRITTEN QUESTIONS

1. (1 point) List your project GroupName and all group members names.
2. Text Classification

   Consider the problem of classifying the following documents:

Itinerary	Document	Class
1	"orchid iris rose tulip clematis peony iris rose"	1
2	"violet clematis peony crocus marigold clematis peony"	2
3	"violet clematis peony daffodil lavender clematis peony"	2
4	"bluebell tulip daisy lily iris rose tulip"	3

(a) (6 points) Assume that we use a Bernoulli (i.e., binary) Naive Bayes model. Compute the following feature probabilities (use laplace smoothing):
   i. $P(X_{peony} = True \,|\, Class = 2)$
   ii. $P(X_{crocus} = True \,|\, Class = 2)$
   iii. $P(X_{peony} = True \,|\, Class = 1)$

(b) (6 points) Assume that we use a multinomial NB model instead. Compute the following probabilities: (use laplace smoothing):
   i. $P(X = peony \,|\, Class = 2)$
   ii. $P(X = crocus \,|\, Class = 2)$
   iii. $P(X = peony \,|\, Class = 1)$

(c) (12 points) Compute and show the calculation for predicting the class of a test document of "daffodil crocus daisy tulip clematis peony" using the Bernoulli Naive Bayes model.

(d) (12 points) Compute and show the calculation for predicting the class of a test document of "daffodil crocus daisy tulip clematis peony" using the multinomial Naive Bayes model.

3. Text Mining
Consider the following documents:

Doc 1: cat, cat, bat, rat, fat, cat
Doc 2: mat, pat, bat, bat, bat, rat
Doc 3: fat, rat, mat, pat, sat, cat

(a) (9 points) Construct the Term Document Matrix

(b) (9 points) Construct the TF-IDF Matrix

(c) (2 points) What is the term-document pair(s) with the highest TF-IDF value.

Part B - CODE QUESTIONS

4. College Data
The set of colleges considered are the 21 top colleges based on earning potential of undergraduates with computer science degrees: `college_data.csv` (Sources: 2014 - Payscale.com and the National Center for Education Statistics, NCES[1])

(a) (8 points) Perform principal component analysis on the college data (make sure to prepare the data in any way necessary).
   Helpful functions: R - `prcomp` or `princomp` functions, Python - `PCA` function of `sklearn.decomposition`

(b) (6 points) Plot the data in the space defined by the first two principal components (labeling each point with the school it represents).

(c) (8 points) Plot the amount of variance explained. How many principal components should be used for any further analysis to be done on the data?

5. Stock Data
Consider the 30 stocks in the Dow Jones Industrial Average[2]. The stock data consists of closing prices for every trading day from Jan 2, 2020 - Dec. 30, 2020. The data was collected from: https://finance.yahoo.com/quote/AAPL/history?p=AAPL .

[1] http://www.payscale.com/college-salary-report/best-schools-by-majors/computer-science?page=12 and https://nces.ed.gov/ipeds/datacenter/institutionlist.aspx?stepId=1

[2] http://en.wikipedia.org/wiki/Dow_Jones_Industrial_Average

(a) (8 points) Perform principal components analysis on the stock data (make sure to prepare the data in any way necessary).

(b) (6 points) Plot the data in the space defined by the first two principal components.

(c) (3 points) Describe any structure you see in the plot (perhaps in terms of the types of stocks)?

(d) (8 points) Perform PCA on the DOW Jones data from 2019, plot the data as in (b). Comment on the patterns seen for 2019 and differences from 2020.

6. Text Classification

For this question you will be considering text classification using two different Naïve Bayes models. These approaches will be discussed in class and reference the following book.

> Manning, C., Raghavan, P., Schutze, H. Introduction to Information Retrieval, Cambridge University Press, 2008
> http://nlp.stanford.edu/IR-book/
> In particular, look at Chapter 13 http://nlp.stanford.edu/IR-book/html/htmledition/text-classification-and-naive-bayes-1.html

You will be using data from the Presidential State of the Union Addresses available as a zip archive. The speeches are available in text files sorted by year, e.g., `a1.txt`, ..., `a231.txt`. The text files are formated such that there is one word per line and most punctuation has been removed. Note, there are still hyphens or dashes left in the text files and there may be some errors in splitting of words.

(a) (10 points) Load the addresses. You will need to create a vector listing the party affiliation of each president to match their speech, you may use the file `party.txt` to help with this classification.[3].

(b) (6 points) Remove *stopwords* from consideration for the method. The stopwords are available at `stopwords.txt`.

(c) Predict the party affiliations (Democrat / Republican) for the following speeches:

- Donald Trump, 2017
- Barack Obama, 2014
- George W. Bush, 2006
- William Clinton, 1995
- John F. Kennedy, 1962

The training set will be the remaining speeches that can be associated with the Democratic or Republican presidents (note, you will not need all the addresses, but they were included here for completeness of the data). You will need to complete the following steps:

i. (10 points) Create a term-document matrix, TD for this set of speeches. Restrict this matrix to the 3000 most frequently used words over all the speeches (not including the stopwords already removed). Show the first 10 rows and 5 columns.

ii. (15 points) For the 5 speeches listed above determine the party affiliations of the president. Calculate and report $P(C = Dems \,|\, \mathbf{X})$ and $P(C = Reb \,|\, \mathbf{X})$ under the Bernoulli model of Naïve Bayes.
Helpful functions: `sklearn.naive_bayes.BernoulliNB` in Python or `naiveBayes` or `bernoulli_naive_bayes` in R.

[3]Information on party affiliation is available at: http://en.wikipedia.org/wiki/List_of_Presidents_of_the_United_States and http://www.apples4theteacher.com/holidays/presidents-day/political-parties.html

iii. (15 points) For the 5 speeches listed above determine the party affiliations of the president. Calculate and report $P(C = Dems \,|\, \mathbf{X})$ and $P(C = Reb \,|\, \mathbf{X})$ under the Multinomial model of Naïve Bayes.
Helpful functions: `sklearn.naive_bayes.MultinomialNB` in Python or `naiveBayes` or `multinomial_naive_bayes` in R.

7. Text Classification - Bonus
Repeat the analysis from above, but implement the Bernoulli Naive Bayes and Multinomial Naive Bayes models yourself. Follow the pseudocode given in the the IR book. *Note, for this question you will not use the standard Naïve Bayes package, library or function. This means you can't use* ***sklearn.naive_bayes.BernoulliNB*** *in Python or* ***naiveBayes*** *or* ***bernoulli_naive_bayes*** *in R. And, this means you can't use* ***sklearn.naive_bayes.MultinomialNB*** *in Python or* ***naiveBayes*** *or* ***multinomial_naive_bayes*** *in R.*

   (a) (10 points (bonus)) For the 5 speeches listed above determine the party affiliations of the president. Calculate and report $P(C = Dems \,|\, \mathbf{X})$ and $P(C = Reb \,|\, \mathbf{X})$ under the Bernoulli model of Naïve Bayes. In order to avoid underflow errors, use the log probabilities as discussed in class.

   (b) (10 points (bonus)) For the 5 speeches listed above determine the party affiliations of the president. Calculate and report $P(C = Dems \,|\, \mathbf{X})$ and $P(C = Reb \,|\, \mathbf{X})$ under the Multinomial model of Naïve Bayes.

Hint: A strong suggestion for this question is to first create and test your code on a small document set, e.g., the example in the IR-book. Once you have that working correctly, then run on the SOTU addresses.

8. Congress Data
For this part of the project you will consider methods to group members of U.S. House of Representatives based on their voting records. The voting records from congress are available at Office of the Clerk, US House of Representatives[4], but not in a form that is easily digestible for analysis.

   In fact it was only in 2016, that Congress agreed to make legislative data available themselves. Govtrack.us has links to primary data sources and api's projects that collect and release the data in easier digestible forms.

   A long-standing project to document congressional roll call votes at the Inter-university Consortium for Political and Social Research (ICPSR)[5]. This data includes roll call votes from 1789 - 1990. The ICPSR formatting for storing this data has been used on other sites which are keeping up with the creating a record.

   The data you will use was downloaded in its raw state from https://voteview.com/data. Then, for each of the last 20 congresses, 50 random votes were selected for each member (ignoring the first 10 votes of each session - these votes usually have to do with electing a Speaker of the House and rules votes). The results are stored in the the files `H97_votes.csv`, `H98_votes.csv`, ..., `H116_votes.csv`.

[4] http://clerk.house.gov/legislative/legvotes.aspx
[5] https://www.icpsr.umich.edu/icpsrweb/ICPSR/series/159

(a) (4 points (bonus)) Working with data for the 116th Congress, `H116_votes.csv`, perform principal component analysis on the voting record and plot the first two principal components. Because we know the party affiliation of each member of congress, color the plot based on party (red = Republican, PartyCode=200; blue = Democrat, PartyCode = 100; green = Independent, PartyCode in 300s).

(b) (8 points (bonus)) To help answer the question on whether the voting records have always been so divided by party, perform the following analysis. Create a small multiples plot (5 x 4) showing the results of PCA (colored by party) for the last 20 Congresses.