DS 501 Case Study Assignment 3 (120 Points)

Due: 05:00 p.m. on 04/06/2022

Case Study Objective: Textual Analysis of Movie Reviews

- Study movie reviews from **the polarity dataset v2.0** comes from the http://www.cs.cornell.edu/people/pabo/movie-review-data/ which contains written reviews of movies divided into 1000 positive and 1000 negative reviews.
- Analyze the data set, make conjectures, support or refute those conjectures with data, and tell a story about the data!

automated data mining survey
responses co ter transcripts
qualatativ oot cause
classificati text analysis insights
ad-hoc an s product
reviews se t vo f the
customer dashboards consume
trends ad hoc analysis early warning

Case Study Deliverables: Please compress all the below files into a zipped file and submit the zip file (**team#_case_study_3.zip**) to Canvas, where **#** is your team number.

1. **Notebook File**: Save this Jupyter Notebook (see the provided template) for Problem 1 to Problem 4 and find the notebook file in your folder (for example, "filename.ipynb"). This is the file that you need to submit. Please make sure that all the answers and plotted tables and figures are in the notebook.

2. **PDF Report**: Please (1) prepare a report (**At Least 5 pages and Not More Than 10 pages**) to report what you found in the data and (2) include figures and tables in the report, but no source code.
   - Motivation and Background: Why is this case topic interesting and or important to you? Specifically, this case topic should be from a public organization, private company, social community, etc., in any domain.
   - Data Sources: What data did you collected? Please also describe your data characteristics in detail.
   - Methodology: How did you analyze the data? Please explain your step-by-step procedure in detail.
   - Results: What did you find in the data? That it, what insights can you obtain from the data?
   - Conclusions: What conclusions can you make from the results?

3. **PPT Slides**: Please prepare PPT slides (for 30 minutes' talk) to present about the case study.
   - 25 minutes for each presentation
   - 5 minutes for questions

Required Readings: This case study will be based upon the Scikit-learn Python library. We will build upon the tutorial "**Working with Text Data**" which can be found at https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html?msclkid=4afab148a79d11ec83226e8cc4289664.

Required Python Libraries:

- NumPy (www.numpy.org) (should already be installed from Case Study Assignment 2)
- Matplotlib (www.matplotlib.org) (should already be installed from Case Study Assignment 2)
- Scikit-learn (https://scikit-learn.org/stable/) (available from Anaconda)
- Python Natural Language Processing Toolkit (www.nltk.org) or (https://spacy.io/).

Note:

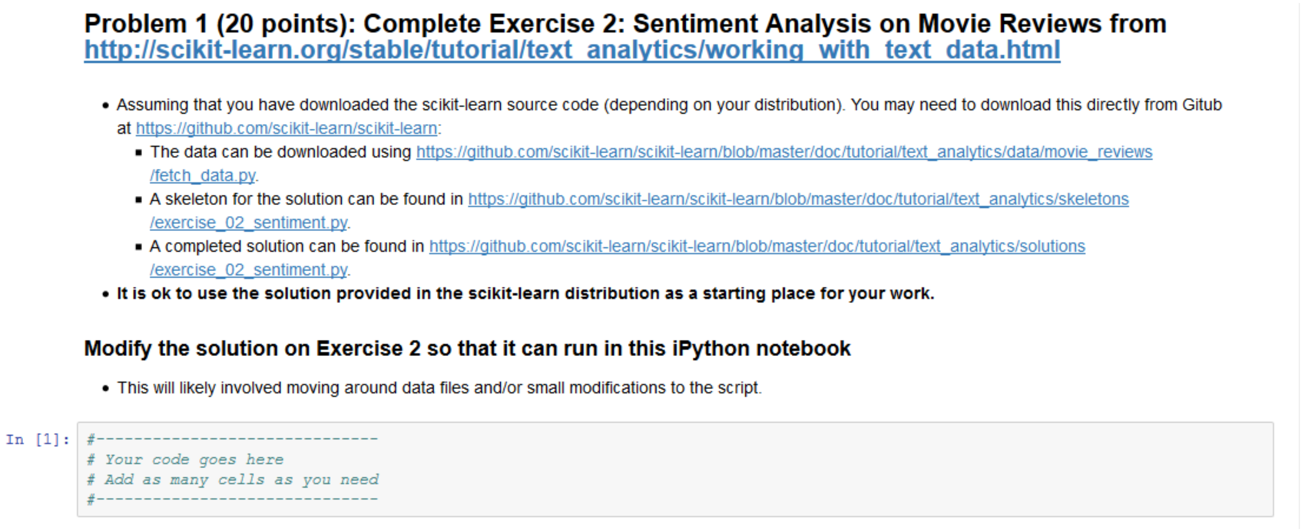
(1) *The above online website is the latest version that I could find for your reference.*

(2) *Please save the given notebook (**.ipynb**) frequently when working in JupyterLab; otherwise, the changes that you made will be lost.*

(3) *Assignments are accepted in their assigned Canvas drop box without penalty if they are received by 05:00PM EST on the due date, or 10% of the graded score is deducted for the late submission per day. Work submitted after one week of its original due date will not be accepted.*

Grading Criteria:

I. Notebook: Points = 80



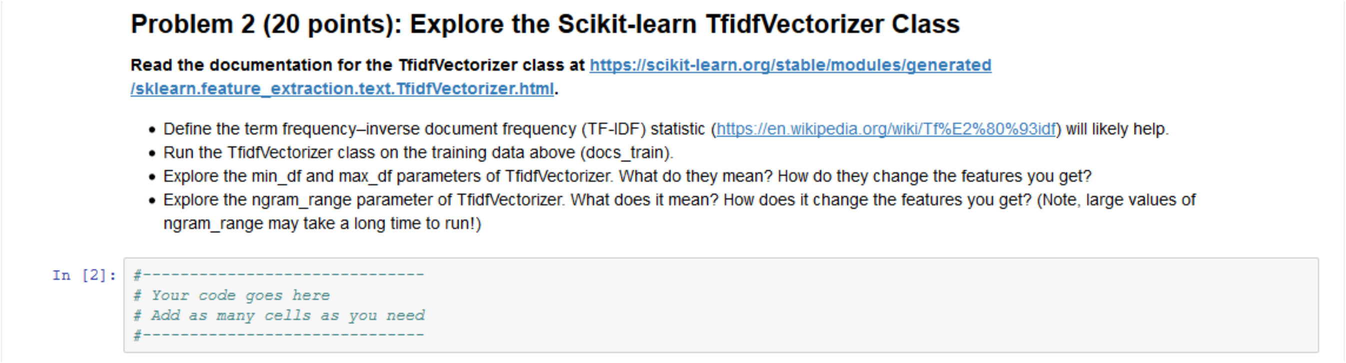
Problem 1 (20 points): Complete Exercise 2: Sentiment Analysis on Movie Reviews from http://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html

- Assuming that you have downloaded the scikit-learn source code (depending on your distribution). You may need to download this directly from Gitub at https://github.com/scikit-learn/scikit-learn:
  - The data can be downloaded using https://github.com/scikit-learn/scikit-learn/blob/master/doc/tutorial/text_analytics/data/movie_reviews/fetch_data.py.
  - A skeleton for the solution can be found in https://github.com/scikit-learn/scikit-learn/blob/master/doc/tutorial/text_analytics/skeletons/exercise_02_sentiment.py.
  - A completed solution can be found in https://github.com/scikit-learn/scikit-learn/blob/master/doc/tutorial/text_analytics/solutions/exercise_02_sentiment.py.
- **It is ok to use the solution provided in the scikit-learn distribution as a starting place for your work.**

Modify the solution on Exercise 2 so that it can run in this iPython notebook

- This will likely involved moving around data files and/or small modifications to the script.

In [1]:
```
#-----------------------------
# Your code goes here
# Add as many cells as you need
#-----------------------------
```

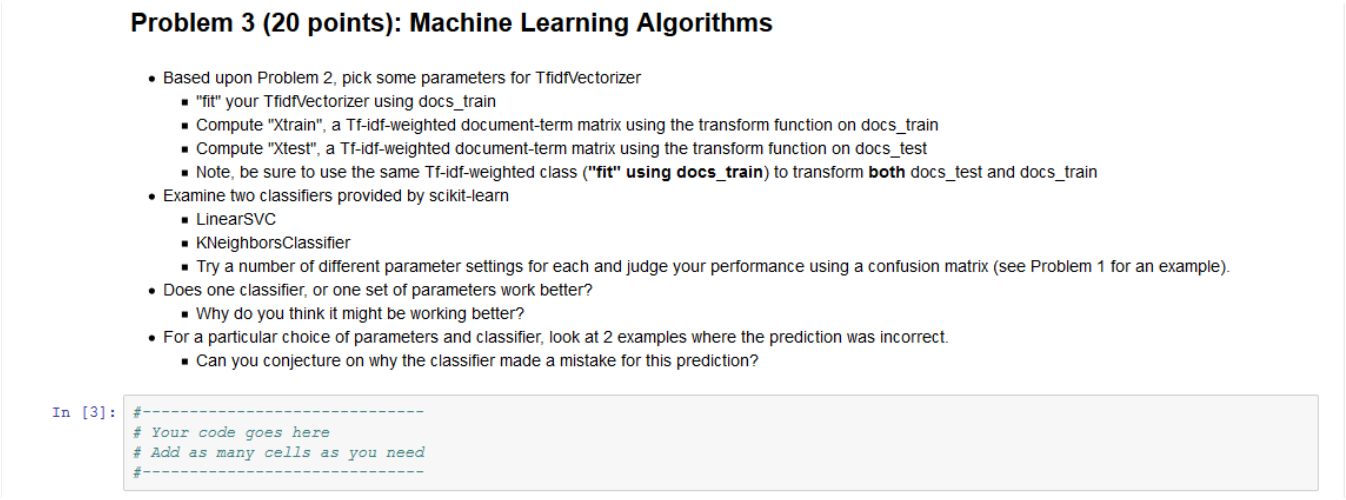


Problem 2 (20 points): Explore the Scikit-learn TfidfVectorizer Class

Read the documentation for the TfidfVectorizer class at https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html.

- Define the term frequency–inverse document frequency (TF-IDF) statistic (https://en.wikipedia.org/wiki/Tf%E2%80%93idf) will likely help.
- Run the TfidfVectorizer class on the training data above (docs_train).
- Explore the min_df and max_df parameters of TfidfVectorizer. What do they mean? How do they change the features you get?
- Explore the ngram_range parameter of TfidfVectorizer. What does it mean? How does it change the features you get? (Note, large values of ngram_range may take a long time to run!)

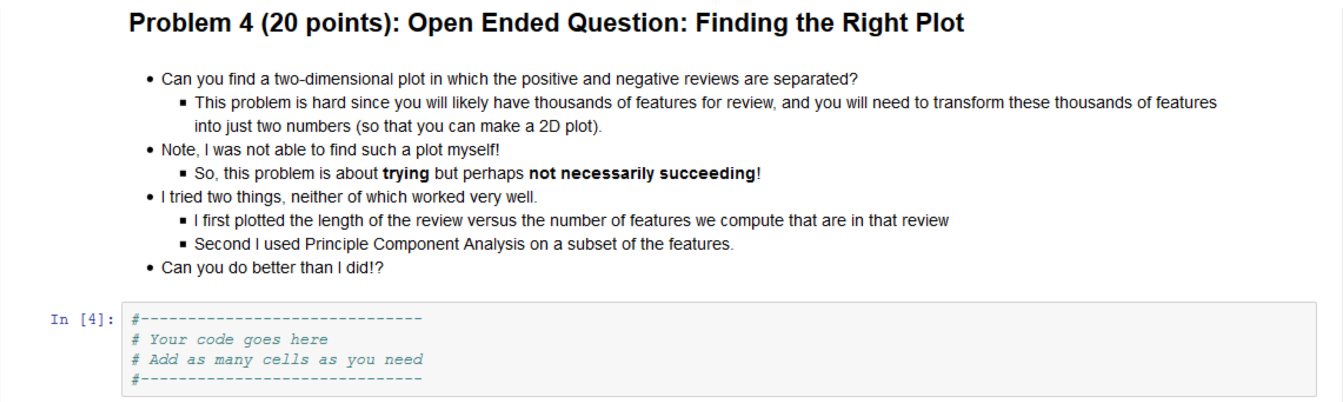
In [2]:
```
#-----------------------------
# Your code goes here
# Add as many cells as you need
#-----------------------------
```



Problem 3 (20 points): Machine Learning Algorithms

- Based upon Problem 2, pick some parameters for TfidfVectorizer
  - "fit" your TfidfVectorizer using docs_train
  - Compute "Xtrain", a Tf-idf-weighted document-term matrix using the transform function on docs_train
  - Compute "Xtest", a Tf-idf-weighted document-term matrix using the transform function on docs_test
  - Note, be sure to use the same Tf-idf-weighted class (**"fit" using docs_train**) to transform **both** docs_test and docs_train
- Examine two classifiers provided by scikit-learn
  - LinearSVC
  - KNeighborsClassifier
  - Try a number of different parameter settings for each and judge your performance using a confusion matrix (see Problem 1 for an example).
- Does one classifier, or one set of parameters work better?
  - Why do you think it might be working better?
- For a particular choice of parameters and classifier, look at 2 examples where the prediction was incorrect.
  - Can you conjecture on why the classifier made a mistake for this prediction?

In [3]:
```
#-----------------------------
# Your code goes here
# Add as many cells as you need
#-----------------------------
```



Problem 4 (20 points): Open Ended Question: Finding the Right Plot

- Can you find a two-dimensional plot in which the positive and negative reviews are separated?
  - This problem is hard since you will likely have thousands of features for review, and you will need to transform these thousands of features into just two numbers (so that you can make a 2D plot).
- Note, I was not able to find such a plot myself!
  - So, this problem is about **trying** but perhaps **not necessarily succeeding**!
- I tried two things, neither of which worked very well.
  - I first plotted the length of the review versus the number of features we compute that are in that review
  - Second I used Principle Component Analysis on a subset of the features.
- Can you do better than I did!?

```
In [4]: #-----------------------------
        # Your code goes here
        # Add as many cells as you need
        #-----------------------------
```

II. Report (At Least 5 Pages and Not More Than 10 Pages): Points = 20

(1) Motivation and Background: Why is this case topic interesting and or important to you? Specifically, this case topic should be from a public organization, private company, social community, etc., in any domain. **Points: 5**

(2) Data Sources: What data did you collected? Please also describe your data characteristics in detail. **Points: 2**

(3) Methodology: How did you analyze the data? Please explain your step-by-step procedure in detail. **Points: 5**

(4) Results: What did you find in the data? That it, what insights can you obtain from the data? **Points: 5**

(5) Conclusions: What conclusions can you make from the results? **Points: 3**

III. Slides (25 Min of Presentation; 5 Min of Q&A): Points = 20

(1) Motivation and Background: Why is this case topic interesting and or important to you? Specifically, this case topic should be from a public organization, private company, social community, etc., in any domain. **Points: 2**

(2) Data Sources: What data did you collected? Please also describe your data characteristics in detail. **Points: 2**

(3) Methodology: How did you analyze the data? Please explain your step-by-step procedure in detail. **Points: 2**

(4) Results: What did you find in the data? That it, what insights can you obtain from the data? **Points: 2**

(5) Conclusions: What conclusions can you make from the results? **Points: 2**

(6) Overall Quality of Your Presentation: **Points: 10**

- Storytelling, i.e., how all the above items fit together as a story?
- Description of topic motivation/background, data collection, methodology, results, and conclusions
- Explanation of Question of Interest from the audience