

[← BACK TO GRADES](#)

Final Project

Faculty Feedback

Score

Last published: 12/9/2019 9:32 AM PST

98 / 100 (98.00%)

Comments

Dear Ryan, David, and Diego - AWESOME!!

Projects are never easy :) because they involve working with other folks. Sometimes, teams are lucky and the members are involved and active. Other times - well - not so much. However, I thought that your presentation and your project were both excellent. It is evident that you put in the time (even though there is not a lot of time available) and you used the methods from our class to really think about a "real" dataset and topic.

To review, the beginning of text mining starts with preparing data for analysis. This preparation process can consume 80 - 90 % of the overall time. Text data is often very disorganized and dirty.

Next, some text data is labeled - such as sentiment data - and other text data is not labeled. The following methods are great for text mining (in Python) one the data is either in a corpus or cleanly in a csv file:

1) CountVectorizer - to tokenize and vectorize the data. This creates a sort-of dataframe where the words are the columns and the documents or reviews (or whatever) are the rows.

2) It is important to be able to convert between formats - such as CountVectorizer objects, matrices, and dataframes.

3) Next, if the data is not labeled, one might update the data into *transaction format* and then use Association Rule Mining.

4) Other non-labeled data discovery options for text include word clouds, frequency tables, clustering, and LDA topic modeling.

5) If the text data is labeled one might use NB, SVM, DT, RF, or any other supervised learning method to model the data.

I hope you found this term to be very informative and a great review of supervised and unsupervised methods. Finally, continue to code in Python. Both Python and R are the two key languages for data science. Other good languages to learn or to learn better are SQL, and either C++ or Java.

Thank you all!! DrG

See below for **Basis for Grades – what grades mean.**

Grade: 96/100 A

Grade explanation....

<p>Telling a cohesive, professional, clear, and well-thought out story.</p> <p>Well done!!</p> <p>Here, a 9/10 means that you told a very nice story in the project paper and also gave a solid and clear presentation. A 9/10 is a very excellent grade and means that you are right on track.</p> <p>Part of this grade is an over measure of flow, look and feel, and readability. Have you created clear separations (subsections and headings)?</p> <p>Do you use a lot of tables for clarity?</p> <p>Did you compare methods and discuss and show outcomes?</p> <p>Do you use a lot of visualization for illustration?</p> <p>Are you not placing code or code output in your paper? Code and code output is generally not appropriate in papers. You can sometimes do this IF you create a figure and properly note key elements, etc. It should never be a direct copy/paste.</p> <p>Does your paper tell a cohesive story, does it have a nice flow, do you look at, discuss, stats-analyze, and visualize (explore) the variables, etc. etc.</p>	<p>10/10</p>
<p>Introduction: (4 - 6 paragraphs)</p>	<p>19.5/20</p>

Really nice work!!

For a relative comparison, an 18.5/20 means you met all the key and core requirements and did a very nice job. Can you get 19 or 20/20? Sure – be creative, be more awesome, exceed, think about side the box.

An Introduction is about the **area or topic**, not about the datasets, variables, methods, or models.

The introduction helps the reader to understand what the topic of the data area is about, who it affects, and why it matters.

An introduction and conclusion should always tie together. *You* should always understand what information you discovered (or predicted), what it means, who it affects, why it matters, what the relevance and applications might be, and how to explain this to non-tech people.

Writing skills are very important in real-life. In fact, it is often the case that a first impression is based on writing. Writing in this class should be professional, at a high level of skill, and never about you. Avoid using “I”, “me”, “my”, “you”, “us”, “we”, etc. Write about the topic, methods, models, results, etc. Do not write about yourself.

A professional or graduate-level paper should never have a “studenty” or “undergrad” feel, such as using sentences like “My Assignment is about...”, or “I am going to talk about...”, or “In this Assignment, I will....”. At the graduate level, write Assignments as though they are technical papers.

The Intro should never contain any information about the dataset variables, where the dataset came from, the data types, or the data cleaning, prep, processing, etc. Everything about the dataset and data wrangling goes into the Analysis section under the “About the Data” subsection.

Analysis and Models - Excellent!!

Here, it is best to have explored the key models and methods from the class. At a minimum, this should include using CountVectorizer, using clustering in some way, using DT, NB, and SVM, and using LDA. It would be great, though not required, to use ARM.

A 23/25 means that you did well – very well – but you can do more. Generally, more includes more data EDA, feature generation,

24.5/25

normalization (with discussion), transformation, aggregation, detailed cleaning WITH the before and after measures, an easy-to-follow flow – so the viewer can quickly see (so tables) what you did – etc...

The Analysis section always contains two or more subsections.

The first subsection in Analysis is “**About the Data**” which contains all the information about the dataset, the variables, the types, visual and statistical EDA, the cleaning and prep, checking for an dealing with missing values, checking for and dealing with incorrect values, checking for an dealing with outliers, feature generation, normalization (if needed), etc.

In this subsection, you will also “explore” the data. This means that you write about each variable, visualize each variable, and talk about what the variable represents. Tables are great for this as well.

The second and remaining subsections of Analysis are the model(s).

In some cases, there may only be one model.

A model is any method used to analyze the data.

Each Assignment specifies which models to use. You can always use more, but never fewer. If a model or method is not specified, start simple and go as deep as you can.

Always include model details and parameter values when applicable.

***** Have Visualizations throughout the assignment. Use color.**

There are 1000s of vis options so while bars, pies, and box-plots are great, have many others as well.

Visualizations should have titles, proper labeling, and should have a Fig. # with a very short explanation.

Results - Excellent!

The Results will compare and discuss the methods and models used. It is great to have confusion matrices, discussions of results for different parameter values (such as linear vs polynomial SVM kernels, etc.) Another huge key to this area are visualizations and relating all results to the *topic*. What did you find? What does it mean? These

24.5/25

are technical results, so it is OK to use words like SVM or k for k means.

For a relative comparison, an 23.5/25 means you met all the key and core requirements and did a very nice job. Can you get a higher grade here? Sure – be creative, exceed, think about the story, the flow, the readability, the measures the use of tables and vis (and color and clarity), etc.

If your results are a great start, but are just way to short, this means that you did well with what you have, but did not cover a lot of things you could have. Results are generally a few pages (not paragraphs). Results have tables, comparisons, vis, discussion, exploration of attributes, measures of accuracy, etc, etc,

The Results section of the Assignment will have a subsection for results for each model (assuming that you have more than one).

Results are technical. They offer technical information about what was found in the analysis. For example, if you performed a correlation in the analysis between all pairs of numeric variables, then your results would discuss the r-value and relationship of each pair. Similarly, if you looked at measures of center and variation, the results talk about what those measures are and what they reveal. For example, if the mean is less than the median, the data is skewed, which means....

Each model we will use in this class has results and parameters associated with it.

Use a lot of vis and tables for clarity and illustration.

Results should include comparison tables –

** Always have lots of visualizations

Conclusions

5-6 paragraphs. This is directly connected to the Introduction – but – it now explains what was learned via the analysis and results. However, unlike the Results, this is non-technical and very focused on why humans care, how humans can use this information, what this all means to humans, etc.

Throughout this class, your conclusions (and your intro) should get better and better. Think about what that means. Think about where you

19.5/20

are now. (If you got an 18/20 – this means you met all the requirements and did a good job. A higher grade means you did that and a bit more – like better flow – interesting comments – more detail, better detail, more creativity, better use of visualization – etc. etc.) If you are below 18/20, you are not quite there yet. This could mean that your conclusions contain some technical matter that really belong in the Results. It could mean that your conclusion is a good start, but really could do a better and more thorough job of telling the story. Etc. Think about it.

A data analysis story is like a picture that you have painted. Everyone can paint a different picture. Everyone can choose colors, styles, flow, use of space, etc. Interestingly, some pictures represent better than others. Why? Learning to answer this is part of the learning experience. So, think about where you are, what you did well, and what you can improve. Be creative – you are the artist!

This area is not technical at all. If you are technical, such as using words like k means, kernel, linearly related, rules, sup, etc. You can lose points. When in doubt, make it more clear and less complicated. Tell the reader what happened and how it affects real-life. Keep the “results” in the Results section.

Always ask yourself if a 10-year can read and understand (and appreciate) what you did.

This area explains what was actually found in a way that would make sense to anyone.

For example, if you discovered in the analysis that association rule mining with a conf of .2 and a sup of .3 offered 10 rules, you would talk about the measures and values and rules in the *results*. In the

Conclusions, you would talk about what it all means. So you would not include the rules themselves or mention of technical measures such as conf or sup. Rather, you would say that you found (as a random example) that people who buy diapers are very likely to buy beer and that this means that a store should consider placing these items “near” to each other.

Conclusions also help readers to make decisions based on the research that you did and why it matters in real life and for them.

Do not use techy terms :)

Basis For Grades:

99 - 100: This means that your Assignment was so amazing and so perfect that **nothing can be improved**. It covered everything – cleaning, prep – analysis that makes sense – visualizations – results (that are true) – etc. There is nothing really left to improve. I generally do not give 100% grades as this does not help you to think about what you can add or do better based on where you are. Everyone can always get better – this is why we seek education.

92 - 98: This means that your Assignment is really good! You covered most of the items and areas noted and perhaps a few others not noted. You can make some improvement on pre-processing and results analysis, as well as perhaps other visualizations. Overall – you have the idea and you did well. A grade of 97 or 98 means that dazzled me with your awesomeness. This is GOOD GRADE! A 93 is an A grade that means that you did very well, met the requirements, put in the time, etc. If you want more than a 93/A, think about what more you can do – how you can exceed and base-expectations, how and where you can apply more (and better) vis options, how you can use tables, etc. etc.

89 - 91: This means that your Assignment is very good, but could be a little better. Perhaps add items such as further data cleaning and pre-processing, data normalization, better or more visualizations, and/or more robust conclusions. Many students forgot to change Section to a factor for example. Very few students summed and normalized the data to look at the percentages for each attribute.

85 - 88: This means that your Assignment is a good start and largely meets the more general and overall requirements. Here, you used R, you did some analysis, you did some cleaning, you made some graphs, and your reached some conclusions. However, there is room for improvement.



Below 85 means that the level of 85 above was not quite met and many elements were missing.

Student Submission | [Final Project](#)**Response**

Last submitted: 12/8/2019 4:12 PM PST

No response

Files | [Download all \(3\)](#)

File Name	Uploaded	Feedback
Final_Project_Timbrook_Ryan.docx	12/8/2019 4:11 PM PST	
Team5_Final_Project_code.zip	12/8/2019 4:11 PM PST	
Team5_Final_Project_data.zip	12/8/2019 4:11 PM PST	