- ← BACK TO GRADES

Homework 3

Faculty Feedback

Score

Last published: 10/23/2019 4:54 PM PDT

98 / 100 (98.00%)

Comments

Ryan- Week 3 Assignment Rubric and Explanations

At this point in the class, you should feel very comfortable with reading text-type data (such as csv or corpus) in to a Data Frame. Remember that text comes in many forms, from csv files, to folders of files (corpus), to HTML/web scraping, to Twitter/JSON, etc. Text can be speeches, novels, reviews, articles, Tweets, etc.

The first step in text mining (discovering information from text) is to tokenize (break into words) and then vectorize (make the frequency dataframe). This takes text from messy into dafaframe format.

Once in a DF format, you can think about normalization methods, whether you have labeled data (such as topics or sentiment), and how you might want to model or explore.

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

Grade: 98/100 A+ Very nice work!

Grade explanation….

Telling a cohesive, professional, clear, and well-thought out story. Here, a 9/10 means that you told a very nice story. A 9/10 is a very excellent grade and means that you are right on track. Part of this grade is an over measure of flow, look and feel, and readability. Have you created clear separations (subsections and headings)? Do you use a lot of tables for clarity?	10/10 Nice!

Do you use a lot of visualization for illustration? 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. 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.		
Introduction: (2 - 3 paragraphs: 3 is better.) An 18.5 means that you are on the right track and did well – but can now start to think about how to make it better. Technically, the topic is always based on the dataset. However, for this case only, I will accept either an intro about AI and sentiment analysis or about the data. However, remember, for the future, the intro is always about the topic of the data. For example, if the data is about movie reviews, the intro is about movies, movie reviews, who completes them, who reads them, who they affect, why they matter, etc. 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. For example, if your data focuses on restaurant reviews, then this is your topic. You might first talk about how often American eat at restaurants. Then you can talk about how much money is spent eating out. Then you might navigate into how the use of reviews has become common and how such reviews affect restaurant choices. From here, you can talk about the dangers of fake reviews to both restaurant goers and restaurant owners, etc. etc… **The introduction helps the reader to understand what the <u>topic</u> of the data area is about, who it affects, and why it matters.**	19.5/20	

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 For this assignment - I will accept the use of Corpus, or .csv, or both. However, be sure you can read data from many sources into a DF. A 23/25 means that you did well – very well – but you can do more. Generally, more includes more data EDA, feature generation, 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… In text mining, visual EDA often uses word clouds to “see” the common words. These can also be used to clean the data and to prep it. 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	24.5/25

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 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. Results should 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	24.5/25

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 – such as comparing the top 5 rules for conf, sup, and lift, or comparing different kernels (and a measure of accuracy) for SVMs, or different values of k for k – means – etc. etc. … ** Always have lots of visualizations	
Conclusions – this is not an easy area. **An 18.5 means you did well, but I want you to think about how to keep improving.** Throughout this class, your conclusions (and your intro) should get better and better and should connect to each other. Think about what that means. Think about where you 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 works like k means, kernel, linearly related, rules, sup, etc. You can lose	19/20

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 old person 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. Conclusions also help readers to make decisions based on the research that you did. For example, a university group might perform protein assay studies to understand amino acid digestion systems in the gut. However, the conclusions will not talk about assays J Instead, it will talk about what the results IMPLY…such as the need for those with ulcers or GERD to take glutamine alone (not with other protein) to assist in healing…	

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 | Homework 3

Response

Last submitted: 10/22/2019 10:01 PM PDT

No response

Files | Download all (3)

File Name	Uploaded	Feedback
HW_3_Timbrook_Ryan.docx	10/22/2019 9:54 PM PDT	!
text_mine_nfl_players_list.ipynb	10/22/2019 9:54 PM PDT	?
HW3_Timbrook_Data.zip	10/22/2019 10:01 PM PDT	?