Peer Review - Project One

Anonymous

March 16, 2016

Peer-assessment

First I'd give some general feedback, and then I will quickly give feedback on each of your sections.

General Remarks

First of all, use the latex template. It will help structure the report, which is badly needed.

The language used in the report should also be revised, for example there's a constant change between *we* and *I*, pick one and stick to it. Remember, you're writing an article. Think of it not as a first draft, neither as a short recab for a fellow student, but rather as a piece which you aim to have published, and have read by people outside this couse. This is why proper grammar, precise language and a strong structure is key, it will all help the reader to better understand and follow what you're trying to say.

When including data from your code, I strongly suggest you put these into either a table, a list or a figure, table and figure being described in the template. It allows for easy reference, while also making it clear for the reader that the data is special. Also, use references when mentioning foreign tools or reports.

Correctness of report

Other than having only computed values for one vaccine, the method explained in the report seems sound, although a bit lacking, as RMSE alone is a weak result.

Other general remarks

It does appear that the report covers all the essential parts of the requirements.

Certain parts of the report seemed to be more like filler comments than actual useful observations, keep descriptions short and to the point.

I'm missing examples and conclusions upon the examples. I don't see many choices being made, and it seems like you're just working towards a goal, without stopping and thinking of why and how to best get there.

As a fellow student, it's easy to understand what you are doing, but if the report were to be presented to a random computer science student, I am not sure they will get what it going on, there's too much you assume the reader already knows about.

Report feedback

Intro

Remember, it's no guarantee that the reader is a fellow student for this course. Try and quickly cover the motivation for why we are doing this.

Methodology

Finding queries

The section on how you found your list of queries is too long, and feels like an unnecessary summary. It can easily be summarized as taking descriptions of vaccines, and stripping them of unnecessary stop words and punctuations, after which descriptions for a certain vaccine was compared, and words extracted. It's far more important that you cover the general method of processing the data, rather than going into detail of everything you did, for example creating a file from another file, creating exactly four lists, appending elements to lists and using certain functions are hardly relevant for the general reader.

I don't see the relevance of ordering alphabetically.

Why are numbers excluded from the queries? Is it because they are very general? If so, can't certain other words may also be removed. I would like a quick reasoning as to why.

Getting frequencies

When you mention something new, like PyTrends, it's a good idea to make a note to where you can read more. For example PyTrends[1] or maybe even a reference[1], for the end of your report. This comes back to the general lack of structure.

Having tried, and failed, to run your code (spend about an hour working on it), I strongly recommend you give it a thorough rework as well, as I was unable to verify your methods. When you have two functions, which are doing exactly the same thing, but for different vaccines, they should be merged into a single, general method. Your code, as supplied in the archive file, cannot run by itself. There's dependencies on files which are missing (*stop_words_modified.txt*, *all_data_end.xls*). Furthermore, having no comments and much of the code inactive was no help.

I mention this, as you are including a certain code fragment directly in the report. Which opts me to try and confirm your statements in this section.

[1] https://github.com/GeneralMills/pytrends

Putting data into one document

Probably the trickiest part of the project, was to make sure the weekly and monthly data could coexist, and your method of 4.5 weeks allows this to happen. However, looking at the code, I recommend not hard-coding values for weekly csv files, as they may change. Rather, you should, at runtime, find a way to determine if a file is either on a monthly or weekly basis, this can be easily achieved, for example by using regular expressions.

Method used for prediction

When mentioning useful functions, even if their names may give a clear indication of what they do, I suggest adding a small description to them. Remember to once again make a reference to sklearn.

I would like some more information on how you got your results. All you mention is using the default parameter, OLS method and KFold. What happens in the folds, or maybe an example of what a *predict*() call returns may be interesting, as a plot maybe?

Findings and Conclusion

Having these in a table would make it look nicer. It's a shame you only got the result for a single vaccine, especially with all the extra code you wrote. Giving the RMSE is okay, but I suggest you compare these to the ones in the handout. Are you worse or better than them? And what can you conclude from these values? I would like more Findings if possible, is RMSE enough? As is, I would like to know what you achieved in the project. What was the point? Remember, the reader is not necessarily a fellow student.

In recap

Use the latex template and tighten up both language and content.
Use citation, if your work is dependent on the work of others, they should get credit.
Give more examples (plots, different score values etc), or make the code executable (include missing files), it's impossible to evaluate your work based purely on the RMSE of a single vaccine.
Include the other vaccine as well.
Rework your code. Don't rely on you to do anything manually. Make your code fully automatic, from getting the csv files all the way to your have a RMSE value, it should be scalable and require minimum effort to adjust for other vaccines. Add comments!.

References

[1] Author, Website, etc