Peer-assessment

Project 1

Peerreview

Anonymous
18-03-2016

Indholdsfortegnelse

Introduktion 3
Comments on the introduction: 3
Comments on the Methodology 3
Find text: 3
Tokenize: 4
Queries for Google 4
Making the prediction 5
Findings 5
Conclusion 5
Overall assessment 6
Follow-ups 7
Installing packages/modules with Python: 8
Python code review 8
Missing Layman vaccine descriptions 8
Missing similes in your paper for handling vaccine descriptions 8
Error when handling data 9
Overall code-review 10

Introduktion

The peer-assessment and the code-review were done separately – some problems noticed in the inspection of the report, may be mentioned in the code-review.

Comments on the introduction:

The paper's introduction is a little vague, and forces the reader to read the paper Hansen et al.[1] to fully understand the grasp of the paper. The introduction should contain a short introduction to the method used in Hansen et al.[1], in addition to a description of those methods that were not adopt from Hansen et al.[1] – here Hansen et al.[1] used both clinical data and web data to predict vaccine uptake, while this report used web data, and clinical data as ground truth.

The introduction should also contain a short description of the work behind the report, similar to which websites the layman description were mined from and where they were located, which validation methods used and which websites used for scraping (here Google Trends).

In some cases it would be preferable to make an abstract before the introduction, that describe the paper with a very short resume, that sums up the hole paper.

Comments on the Methodology

I do like the methodology sections; Find text, Tokenize, Queries for Google and Making prediction.

But, your methodology is not detailed enough.

Find text:

A few spelling errors, and you need to do some proofreading.

There is missing a firm description on how this paper found the query terms – here how and where the paper found the vaccine descriptions. This paper state that it used the descriptions provided by the assignment job description, but the assignment job description only have one description for each vaccine, and this paper are suppose to use two for word comparison.

Reveal the method behind the extraction of the vaccine descriptions. Another approach would be to create an appendix, where this paper explains exactly how and where the query term were found. If it were needed to copy-paste the description directly from a website to a .txt file, or if it were needed to encode the .txt file beforehand to UTF-8, or used some other approach which is need for the reader to replicate the data collection, explain it in the report.

Tokenize:

A few spelling errors, and you need to do some proofreading

Instead of commenting that the solution to the tokenize problem is defined in the assignment job description, just explain how it were done and why, with your own words.

Vague descriptions and missing proofreading:

> "Now that I only use the two text provided, I can then extract all the words from on of the files, that is included in the other file."

An example of a rewrite with better understanding:

> "When basic tokenize operations were completed on both vaccine descriptions, it was possible to collect those search query terms that appeared in *both* vaccine descriptions.

The report state than when using Python string library to remove punctuation, this may have unintentionally split words, that were meant to be one. If this is the case, it would be something that should be discuss in the conclusion, rather or not if this have an impact on the final findings.

Queries for Google

A few spelling errors, and you need to do some proofreading.

It is perfect that the paper explain the week-to-month problem, that some data extracted from Google Trends are in weeks and not in month. This paper also explain the possible side effects the solution may create – which is perfect.

There are fuzzy lines that need a rewrite, such as:

> "Now that I have each word, I get the data for each word."

This line does not tell me anything, but a rewrite may be:

> "Since I extracted the search query terms, each search query word were feed to Google Trends using ect ect ect. Each word provided to Google Trends delivered a report, containing word frequencies, from some date to another date…"

There are missing the method of combining the word frequencies reports, to a single report for all word frequencies.

Making the prediction

Missing a improved and more comprehensive description of the linear model, and what the linear model and supervised learning model Lasso does to the data collected from Google Trends. Furthermore, a description of the 5-fold cross-validation.

- Some function descriptions, both for Lasso and 5-fold cross-validation

The paper perfectly state the setting used when doing the cross validation. The paper further state that the setting positive is set to true, because does not make sense to have a negative coefficient, here due to the fact that searches for one thing, does not really negatively impact searches for other things. Please explain.

Findings

This section is rather short, and does not explain anything about the findings. Diagrams showing the Google Trends data before doing a linear operation, after, cross validation would be preferable, since it gives an overview, of the operations and methods used to get the final result. A discussion of the Diagrams would also be preferable, since it shows an understanding of the Lasso function and the 5-fold cross-validation operation.

The paper state:

"…Lasso function have done a decent job on both vaccines."

But the paper does not explain with expect to what.

Conclusion

Minor spelling errors, and very short section.

Missing a discussion, that evaluates the final result (here the RMSE). Make an RSME value interpretation, that discuss a good RMSE versus a bad RMSE. Compare with the findings of Hansen et al.[1] with the findings from this paper.

The paper state that there is a few problems with the target data, but this is not entirely true – the target data is correct going from the period 2011-01 – 2015-09 as Hansen et al. The data collected from Google Trends have to be in the same period from 2011-01 to 2015-09. The data collected is from 2011-01 – 2011-12, so it is true that the algorithm will run over missing data (all zeroes) from the period 2015-09 to 2015-12, therefore may be giving a different RMSE.

Overall assessment

Bad:

It was impossible to know which vaccine descriptions that were used in the paper, and therefore unfeasible to check, if the search word queries found in this paper, did in fact consist of words represented in both vaccine descriptions.

Hansen et al. did only use reports extracted from Google Trends from the tie period 2011-01 – 2015-09, looking at the data (HPVdat.csv), it seems that data is extracted from the period 2011-01 – 2015-12. This makes it difficult to compare the results from Hansen et al., with the data from this paper.

Missing coherent and unyielding descriptions of the findings, function descriptions and usage.

Missing diagrams showing Google Trends data as a scatter gram, Lasso and 5-fold cross-validation.

Spelling errors and lack of proofreading.

Taken as a whole, it is a very short paper.

Missing a brief discussion of the results from the experimental results

---

Good:

The paper manage to explain the general problems, and provide a fair understanding to why search frequencies on Google Trends can predict vaccine uptakes in a near future.

This paper manage to get a final result, provided by the search queries collected from vaccine descriptions

Python scripts works (with some changes, that may be due to different Python versions)

The paper manage to bring RMSE results for both vaccines, even though the Google data period collected were from 2011-01 – 2015-12 and did not match the period from the group truth data (2011-01 – 2015-09).

The paper followed the standard provided by the assignment job description, where we were told to at least have the following sections:

- Introduction
- Methodology
- Findings
- Conclusion
- References

The paper explains which methods used for prediction and what settings used for the experiments

Follow-ups

The paper lacks acknowledgement of the related work – there are references to Hansen et al.[1], but descriptions of the Hansen et al.[1] method contra the method used in this paper is absent. Furthermore, a comparison to the results from Hansen et al.[1] would strengthen the paper and state a solid conclusion to the report.

Data collection are done using pytrends, which is a good and sound solution to the problem –if it is possible to change and modify the period that is need to scrape.

The report delivers a final result, which support the knowledge of data scraping. In addition, the report manage to calculate Lasso using the module sklearn for Python, and the 5-fold cross-validation.

The presentation of the result are inadequate, due to lack of diagrams and function representation, and basic explanation of methods used.

The project did not break project guidelines

There are no recommendation to the revision of the project description, due to the fact that it delivered at challenge, and were not hard to understand or grasp.

.

[1] N. D. Hansen, C. Lioma, and K. Molbak. Predicting vaccination uptake using web search queries. in press, 2016.

Installing packages/modules with Python:

It would be a plus, if your paper described the modules used in obtaining the data, such as information regarding:

pytrens(pip install pytrends)

numpy (pip install numpy)

more_itertools (more_itertools)

sklearn (easy-install.exe sklearn

scipy (OBS! This package does not necessarily run when one just do a pip install, I needed to download Anaconda3 at https://www.continuum.io/downloads, and change my python path to anaconda python path in windows)

Preferable information about Python version e.g. Python 2.7 or Python 3.5

Python code review

Missing Layman vaccine descriptions

Missing webpage location of the two vaccine descriptions for MFR. The paper states that the descriptions is taken from Hansen et al.[1].

> "For all vaccines I used the texts provided by the assignment text"

But this is not precise enough, and should have been described in your paper. State where you found the vaccine descriptions, plus date and time, or simply but it in the references.

Example:

> "To generate the queries for Google Trends, this paper used two different layman descriptions, for MMR-2(12)[1] and HPV-1[2], respectively"

[1] www.somethingWithTheDecription.com & www.anotherWebUrlVaccineDecription.com, visited 10/3-16 : 13:20

[2] www.somethingWithTheDecription.com & www.anotherWebUrlVaccineDecription.com, visited 10/3-16 : 13:20

Missing similes in your paper for handling vaccine descriptions

Right after running “get_data.py”, the code stops due to an invalid continuation byte. This is properly due to the fact that the descriptions I got, were not in the correct format:

UnicodeDecodeError: 'utf-8' codec can't decode byte 0xe6 in position 49: invalid continuation byte

Your paper does not state how the vaccine descriptions is formatted, and may be due to:

- Copy-pasting the descriptions directly from the websites.
- Some form of editing were performed in the descriptions, before running the python scripts.
- You did a manual text encoding of the created .txt files before running scripts.
- I am running a different version of Python – Running Python 3.5 using Anaconda extension, on Windows 10.
- Some function in the Python scripts is needed to run, but I did not find them.

In your paper, be sure to describe the following:

- Which files that were created, if any.
- Which encoding the files were changed too or already possessed
- Which functions that are needed to run, if any
- Which version of Python that were used to collect data
- Which modules were used in this project, if any.
- How to install or obtain these modules, if any.

This information will significantly improve the reads perception on the work behind you paper. Furthermore, if the reader or your professor feels that the data seems inconsistent or erroneous, and feel the need to recheck the data collection methods, a proper description of the methods behind, will significantly lower the work burden.

So recreate the data, I used the same description twice, since the two location of the second description you used for the data collection queries, were not mentioned in your report.

Did manage to find out that the .txt files needs to be encoded with UTF-8, and then the description is copied in.

Error when handling data

Tried to run “handle data.py”, when both the queries were created - using “getData.py”, but ran into an error in the function “import_csv(path)”:

“TypeError: float() argument must be a string or a number, not 'map'”

So I tried to run the script with the queries you already created, but same error were showing.

So I changed the function so that the script can as intended:

```
def import_csv(path):
    data1 = []
    with open(path, 'r') as f:
        file = f.readlines()
        f.close()
    words = file[0].split(",")[1:]
    file = file[1:]
    dates = map(lambda x: x.split(",")[0],file)
    data = np.array(list(map(lambda x: x.split(",")[1:],file)),dtype=np.float)
```

Overall code-review

The code was really good, despite the missing description of the method behind. Some code descriptions, would be preferable, since others may need to read the code at some point.

There is missing installation instructions, both Python and each module – do not know if it is preferable by the instructors or professor, but it would give a better understand of the code, and less work when replicating the data collection.

Hope this assessment will help.

Anonymous