WS 2016 Project 1:
Due by: 07 March 2016, 23h55

This project counts towards 30% of your final grade for this course.

1 Project description

This project asks you to replicate some of the experiments reported in Hansen et al. [1]. You are advised to read the paper by Hansen et al. carefully in order to understand what this project requires.

Hansen et al. present a method for automatically predicting future vaccination uptake in Denmark, by combining clinical and web-mined data. Put simply, by considering how many people were vaccinated every month in previous years, as well as how frequently people Googled information regarding vaccinations every month in previous years, we are able to predict how many people will get vaccinated in the future on a monthly basis. This allows to reduce the reaction time of health professionals and streamline clinical resources nationwide, potentially resulting in significant financial and societal gains.

Hansen et al. experiment with 13 different vaccines using training data from the period January 2011 - September 2015. You must experiment with any two of these 13 vaccines (you are free to choose any two you wish).

You will be given the clinical data that Hansen et al. used (details below). Your task will be to (i) mine the web data yourself (details below), (ii) run the prediction algorithm (using ready packages – details below), and (iii) produce a report where you analyse your findings (details below).

The clinical data will be available on Absalon in the directory of this project. The web data that you should mine consist of queries submitted to Google and their frequencies in the period January 2011 - September 2015. Those queries should be relevant to a given vaccine. I.e. we wish to find how often people googled for information related to a vaccine, and use this information to predict how many people may likely get that vaccine. This is how you can create such queries:

- choose any two vaccines from Hansen et al.
- for each vaccine separately, visit the associated urls shown here below. These urls point to information about that vaccine, written in layman terms

  - MFR: http://www.ssi.dk/Vaccination/Boernevaccination/Vacciner%20i%20boernevaccinationsprogrammet/MFR-vaccine.aspx, https://www.sundhed.dk/borger/sygdomme-a-aa/rejsemedicin-og-vaccciner/sygdomme/boerne-og-ungdomsvaccination/mfr-vaccinen/
  - DiTeKiPol: http://www.ssi.dk/Vaccination/Boernevaccination/Vacciner%20i%20boernevaccinationsprogrammet/Di-Te-Ki-Pol-Act-Hib.aspx, https://www.sundhed.dk/borger/sygdomme-a-aa/rejsemedicin-og-vaccciner/sygdomme/boerne-og-ungdomsvaccination/difteri-stivkrampe-kighoste-polio-og-hib/
  - PCV: http://www.ssi.dk/Vaccination/Boernevaccination/Vacciner%20i%20boernevaccinationsprogrammet/Pneumokokvaccine.aspx, https://www.sundhed.dk/borger/sygdomme-a-aa/infektioner/sygdomme/bakteriesygdomme/invasiv-pneumokoksygdom-hos-boern/
  - HPV: http://www.ssi.dk/Vaccination/Boernevaccination/Vacciner%20i%20boernevaccinationsprogrammet/HPV-vaccine.aspx, https://www.sundhed.dk/borger/sygdomme-a-aa/kvindesygdomme/sygdomme/infektioner/hpv-vaccination/

- download the vaccine description
- apply basic pre-processing: remove punctuation, convert capital letters to small, remove stopwords (see attached file)
- tokenize the resulting text (i.e. split into separate words)
- treat as a query each individual term that appears in at least two different descriptions of that vaccine

At the end of this process, you should have a list of queries for each vaccine. To get their frequency on Google, you should submit them to Google Trends http://www.google.com/trends, specify Denmark as location, and the time period you wish to search. You might want to make a new Google account for these experiments. Google Trends will return output of this format: `2011-01-02 - 2011-01-08,91`, i.e. the time period followed by a comma and the frequency of your query. To get the Google Trends output, you might want to build a web scraper. Again you can use any library you wish, e.g., the selenium library: http://www.seleniumhq.org/, which has APIs for many programming languages: Python, Java, C# and more.

The above query frequencies should be used as input features for your prediction. Even though Hansen et al. use both clinical and web data for predictions, you are only asked to use web data (the above query frequencies) for prediction. You can use the clinical data as ground truth. To make the predictions you can use any prediction package you wish. For instance, from the scikit-learn library http://scikit-learn.org/stable/supervised_learning.html#supervised-learning you can use any of the supervised learning models, e.g. Ordinary Least Squares or Lasso. You should use 5-fold cross-

validation when optimizing parameters and report as result the average root-mean-squared-error of the five folds:

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(\hat{p}_t - p_t)^2} \quad (1)$$

where $\hat{p}_t$ is the prediction at time t and p_t is the real observation at time t.

You must submit a report where you describe:

- which vaccinations you chose,
- how you found query terms,
- how you mined query frequencies from the web,
- what method you used for prediction,
- what settings you used for the experiments,
- the experimental results and
- a brief discussion of these results.

You should include as an attachment the queries you submitted to Google Trends and a compressed folder with your code. Your report should include the sections specified in the template file (introduction, methodology, findings, and conclusions). You can use any tools you want, including your own programs, commercial or public domain tools like Unix commands.

2 Submission

2.1 What to submit

You must submit **a single tar.gz file** that contains:

1. your report in pdf (not the latex sources), formatted according to the template,
2. your queries, and
3. all code that you used.

Everything in your submission must be anonymous (i.e., do not write your name in the report or your code). There are no length restrictions for the report, however it must contain **all the sections in the template**, plus any more sections you wish to add.

2.2 How to submit

- You must upload your submission to Absalon **by 07 March 2016, 23h55, at the latest**.
- If you are unable to submit via Absalon for some reason, you must send your submission by e-mail to c.lioma@di.ku.dk **with cc to** oswin.krause@di.ku.dk and nhansen@di.ku.dk **by 07 March 2016, 23h55 at the latest**.
- Submissions received after the deadline without prior approval for e.g. medical reasons or similar, by C. Lioma, will not take part in the peer-assessment. This results in an immediate -20% reduction of your final portfolio grade (15% for the peer-assessment you will not make + 5% for your missing amendment list - see the *Portfolio Guidelines* for details).

References

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