**Data Science for Industry Project 2**

**Due date: 20September 2017 (presentations on 21st)**

The US Consumer Financial Protection Bureau provides a public dataset of complaints they have received about financial products and services (<https://catalog.data.gov/dataset/consumer-complaint-database>). This project uses a subset of 20 000 of these complaints, saved as **complaints.RData** and posted on Slack and Vula. Each record/row captures a single complaint, on the following variables:

* **date\_received:** the date the complaint was received
* **product**: a broad categorisation of the type of financial product involved (5 categories: bank account or service, credit card, credit reporting, debt collection, mortgage)
* **consumer\_complaint\_narrative**: the submitted text explaining the nature of the complaint.
* **consumer\_compensated**: a binary indicator of whether the consumer was compensated (monetarily or non-monetarily) after the complaint.

The aim of this project is to build a Shiny app that will allow users to do some text mining of the complaints, using some of the sentiment analysis and topic modelling techniques we’ve learned in lectures. Specifically:

1. The user should be able to filter on the following variables (i.e. select particular levels they want to do sentiment analysis or topic modelling on):
   1. product
   2. consumer\_compensated
2. Based on the relevant complaints identified using the user’s specified conditions, the app should then do the following:
   1. Use the bing sentiment dictionary to calculate sentiment scores for the complaints.
   2. Plot the histogram of sentiment scores.
   3. Do topic modelling using LDA, The user should be able to specify how many topics they want to use (say between 2 and 5 topics).
   4. Extract and plot the word-topic probabilities (i.e. “beta”) for the 15 most popular words in each topic.
3. The app should include a text box where the user can enter a new complaint in text form. Based on the text, and the selected settings (product, consumer\_compensated, number of topics), the app should:
   1. Return the sentiment score of the new complaint
   2. Return the quantile of the sentiment score of the new complaint (i.e. the relative rank compared to other complaints)
   3. Return the topic probabilities for the new complaint (i.e. “gamma”: see the use of the “posterior” function given in the help file for the LDA function “?LDA”).

Your project consists of the following three items, with the % contribution to the final project mark:

1. Your **R Shiny app**. This should be an R script called app.R made up of a user interface object (ui) and a server function (server). All code in the app should be clearly commented (60%).
2. **Documentation** (in any format e.g. Rmd, docx, pdf) explaining how to use the app and giving a short worked example (20%).
3. A 5-10 minute **live demonstration** of the app i.e. a presentation given by you. Presentations will be done on 21st September – we can arrange convenient times and this can be done via Skype if necessary (20%).