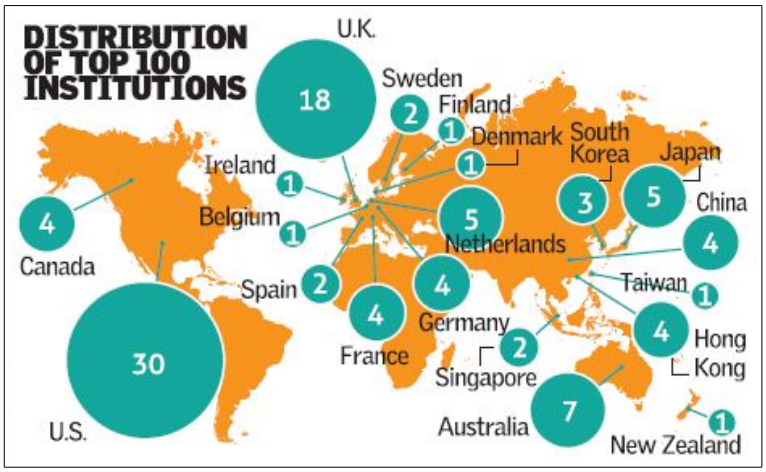
Data Mining Fall 2016

Project 3: Association Rule Mining

Assigned: 10/23/2016
Due: 11/10/2016 (via Canvas)
Points: 100

Please submit your report in **PDF format.**



We will work again with the university ranking data.

Write a report covering in detail all steps of the project. The results have to be reproducible using your report. Carefully describe every assumption and every step in your report. Also, mention any program/code/additional data that you are using for your analysis.

Submit your R code (if necessary also a description of how you used other tools) in a separate file.

Follow the CRISP-DM framework (Steps 3-5)

3. Data Preparation [30 points]

- Select and describe the features you will use. Explain how you preprocess the data. [20 points] For example:
  - Create new binary features to indicate the presence/absence of a certain fact.
  - Discretize continuous features. What method do you use? Explain why? Do you use different discretization methods for different features?
- Construct at least two transaction data sets using different subsets of a single data set. You may choose how to select the subsets (e.g., different county, different year, any other way to group universities). Show and discuss some simple statistics about the transaction sets. [10 points]

4. Modeling [40 points]

- Create sets of frequent, closed and maximal itemsets. [5 points]
- Create sets of association rules. [5 points]

- Use filtering, sorting, tables, and visualization to discuss the found patterns. What do the patterns mean and how are they useful? [20 points]
- Compare the itemsets/rules found in the different transaction data sets. Discuss what the differences mean. [10 points]

5. Evaluation [20 points]

- What findings are the most interesting? Do they support what you have found in the other two projects? [10 points]
- What recommendations do you have for different stakeholders based on your findings? [5 points]
- How useful are your results for different stakeholders? [5 points]

Exceptional Work [10 points]

Examples include:

- Using additional data.
- Comparing more subsets of the data to reveal interesting relationships.
- Exceptional postprocessing and representation of the found rules.
- Use different quality measures for rules, explain what they do and why they produce meaningful insights.