# Text Analytics Assignment #3

**Due: 4th October by 11:59 p.m. on Canvas**

Task A. The **sentiment scores** worksheet in the data file "Assignment 3 Sentiment Scores.csv" (on Canvas) provides sentiment scores (+5 to -5) of forum users on 10 car models. Each row represents a post (not shown) that can mention multiple models. Only positive and negative sentiments are noted.

From these sentiment scores, create a directed product comparison network (and use NodeXL or, even better, write your own code in Python using networkx or R). Use the principles laid out in the article "Product comparison networks" to answer this question.

Task B. Calculate both unweighted and weighted PageRank scores for each car. Note that NodeXL can't calculate weighted PageRank scores. What are the correlations between these metrics and sales figures shown below? What additional information do weighted PageRanks capture? Use a python script to calculate weighted PageRanks. Unweighted PageRanks can be calculated in NodeXL, or you can write a python script for that task as well.

| Model | Approximate # sold in the U.S.A. (2012+2013) |
|---|---|
| Audi A6 | 20k |
| Audi A8 | 12k |
| BMW 3-series | 220k |
| BMW 5-series | 60k |
| BMW 7-series | 14k |
| Jaguar XJ | 6.6k |
| Lexus ES | 135k |
| Lexus LS | 30k |
| Lexus RX | 120k |
| Mercedes S-class | 25k |

Task C. The above sentiment scores above were obtained by manually reading each post. The file "Assignment 3 Edmunds Posts.xlsx" provide a bunch of actual messages (combine the worksheets). Your task is to automate the sentiment extraction from each post. As in tasks A and B, focus on the same 10 models (note that other models may also be mentioned, but that they should be ignored).

Write one or more python or R script(s) to generate sentiment scores for the 10 models just as in the **sentiment scores** worksheet. This will be an unsupervised approach. One possibility (but

not the only one) is to take the dictionary of SentiStrength (along with the default sentiment scores) and use it as inputs in your script(s). Your script should consider lemmatization (e.g., liking and liked must be treated as the same).

Generate sentiment scores with your script(s), find weighted PageRank of each of the 10 cars and correlate with the sales figures above. How does the correlation of this automated approach compare with that of manual scoring in task B?