IMC 466, Machine Learning II
Text mining

Due: Monday, Nov 11, 4pm

You may work in groups. The file `martin.txt` is pipe-delimited (|). It has roughly 1700 products at an on-line drug store that have exactly one review. You know the following

- `exposures`: number of times that someone viewed the review page for this product
- `purchases`: number of times that the product was purchased after someone viewed the review page
- `Nhelp`: number of helpful votes for the review by other customers
- `Nnohelp`: number of not helpful votes
- `verified`: equals 1 if the review was written by a verified buyer and 0 otherwise
- `price`: unit price of the product
- `valence`: number of start (1–5)
- `categoryname`: colon (:) delimited field with tags about the product
- `brandname`: name of brand
- `reviewcomments`: text of the review itself

For all models use K-fold cross validation (where $K = 5$ or 10) and report the CV error, not training error.

1. Create a variable logodds as $\log[p/(1-p)]$, where $p = \texttt{purchases}/\texttt{exposures}$. When $p \in \{0, 1\}$ add/subtract a small offset to avoid invalid computations.
2. Build your best predictive model using all variables but the text. Document your model: e.g., What variables did you create? What models did you try? Which model works best? What variables are the best predictors and how to they affect the outcome? Build two models, one predicting logodds (numerical outcome) and another predicting purchase (binary outcome). Report R^2 for the logodds model and AUC for the purchase model.
3. Repeat the previous part using only the text variables. Again, say what you did. Did you use TF-IDF? Did you stem? etc.
4. Repeat the previous part using all variables. Again, say what you did. Did you use TF-IDF? Did you stem? etc.
5. Give one table summarizing your findings.