Team Mount Lincoln

Team Members: Nicholas Cottrell, Jenny Kao, Molly Pagden, Upasana Rangaraju, Isaac Smith

* Dataset: <https://www.kaggle.com/datafiniti/consumer-reviews-of-amazon-products#1429_1.csv>
* Environment: Use AWS EMR and other AWS services
* Programming Language: Use Spark with Python/R/SQL/Scala
* Requirement phase (Due April 12th):
  + Connect - Slack Channel: <https://app.slack.com/client/TTTHGVCTS/G011CV7J18B/details/members>
  + Setup - Github Team Page: <https://github.com/MSBX5420/Team-Mount-Lincoln>
* Project Goal and Scope:
  + Our team is doing analysis on Amazon reviews
  + We would like to determine what the “hot” words are for each product category.
    - We will first filter by product category
    - From there, we can do word count analysis to see what the most common words are for each product category.
    - After that, we can use sentiment analysis in R (or Python) to determine what are the most “positive sentiment” products vs. the most “negative sentiment” products.
    - In addition, we can rank products by which seem to have the most positive or popular reviews
    - We then can input in Tableau to visually see differences between different product categories
    - We will also compare the reviews/review ratings of the product for the first 90 day of the product launch and after the 90 days of the product launch. ( comparison about different categories, clothings vs electronic items)
    - We will also try to predict the rating of the product with the customer’s review of the product.
  + Performance Requirements:
  + Our dataset is huge, almost 400 MB, so we will need several nodes to enhance performance and speed of processing
* Glossary: TBD
* Helpful link:

<https://aws.amazon.com/blogs/machine-learning/detect-sentiment-from-customer-reviews-using-amazon-comprehend/>