Project Proposal

- Team Name: More Data More Problems
    - Group Members: Arthur Jacobs, Danielle Rubin, Devan Richter, JX Xu
- Dataset of Interest: Westbury Lab USENET corpus (2005 – 2011)
    - File size – Two Variants covering 47,860 English Language news groups
        - Unprocessed
            - 36 Gb, compressed
        - Preprocessed/Cleaned
            - Over 8Gb, compressed
            - Over 7 billion words
    - Hypotheses
        - Look at frequency of words distributed through USENET over the course of 2005 to 2011. Specific area of interest includes what words occurred the most, and how the occurrence of these words changes over time (excluding common English words such as and, the, are, etc
    - Scalable Algorithm
        - As of right now, we believe that algorithms based on mapreduce would be appropriate for the dataset unless new/better-suited methods are introduced later in the course