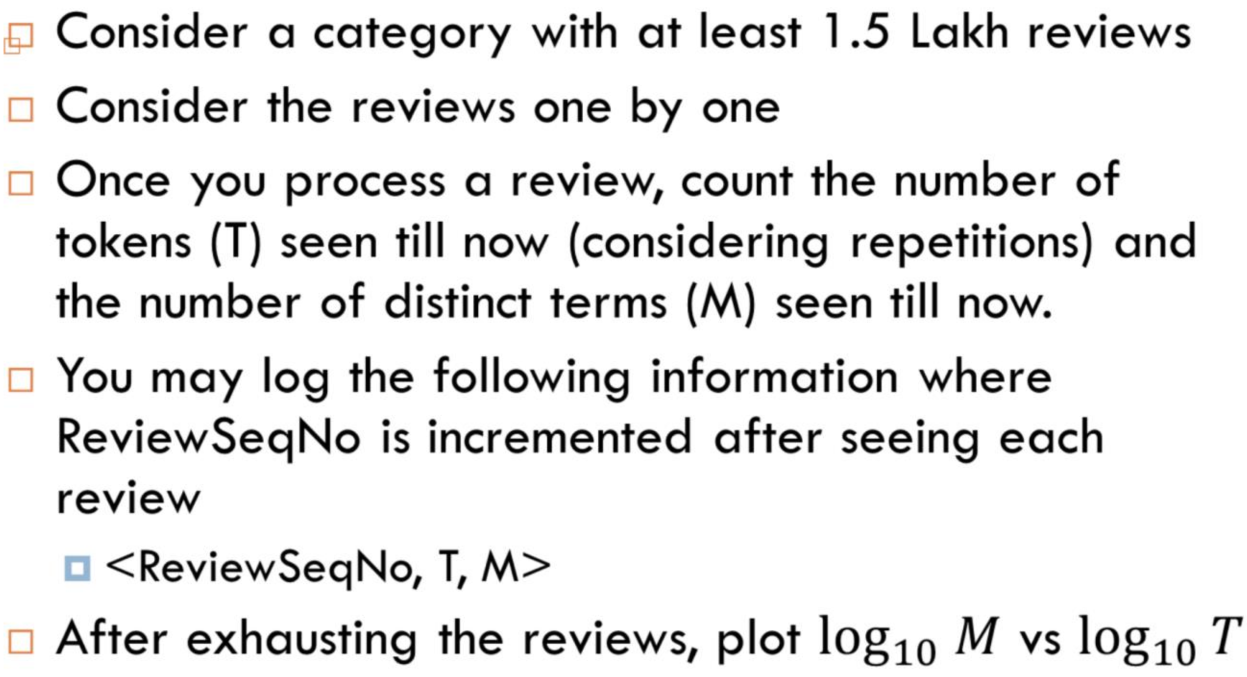
- The source of data for this assignment will be: http://jmcauley.ucsd.edu/data/amazon/
- Download reviews for any three product types.
- Consider any time period (Say date A to Date B)
- Consider the number of reviews for different products belonging to those categories

Task 1:

Identify top words in the reviews. For each category, mention the top 20 words (based on occurrence count).

Task 2:



□ Consider a category with at least 1.5 Lakh reviews

□ Consider the reviews one by one

□ Once you process a review, count the number of tokens (T) seen till now (considering repetitions) and the number of distinct terms (M) seen till now.

□ You may log the following information where ReviewSeqNo is incremented after seeing each review

  ▣ <ReviewSeqNo, T, M>

□ After exhausting the reviews, plot $\log_{10} M$ vs $\log_{10} T$