Report

Task 1

**Method Description: Explain your text cleaning and pre-processing steps, as well as your approach for constructing the distributional semantic representations.**

For the pre-processing steps (function “process\_document”), I have removed all numbers, abbreviations, internal hyphens words, special characters, and punctuation from the reviews, then convert all of them into lowercase. After that, I removed the stop words including “u” and “p”. Finally, I used lemmatization, which only removes affixes if the result is in its dictionary to maintain the meaning of a word. To get the top 50 most frequently occurred words, I merged all the words in the reviews in a list and then do the pre-processing steps (use function “process\_reviews\_str”). For constructing the distributional semantic representations, I have conducted a term-context matrix and used latent semantic indexing (LSI) as the singular value decomposition (SVD) for constructing a low-dimensional dense representation.

**Result Analysis: Analyse and discuss the obtained clustering results.**

Without using SVD and normalization, the feature dimensionality d is 4584, the obtained mean of probabilities after testing 10 times was at 57.6% and the standard deviation was at 0.023323807579381173, then I improved my model by using SVD and normalizing the term-context matrix M, the new feature dimensionality d is 100, then I obtained the mean of probabilities at 93.8% and the standard deviation at 0.02599999999999998. The novelty improvement here is the normalization, which highly improves the performance and training stability of the model, without the normalization, the performance is still around 57.6%, and after normalization, the performance reached 93.8%.