Read Me

* **000\_Scraping\_Diabetes\_Reviews:** This file scraps reviews and ratings for 6 drugs from Drugs.com
* **01\_PreProcessing:** It takes in the raw, scraped data for reviews for 6 drugs. It cleans the text. Then it divides each review into single words and pair-words. It also creates a sentiment score for each review (row). It then outputs a matrix dataframe ‘data’ with columns corresponding to a word or pair-word in a review. Each binary observation is a review and shows whether a word appears (1) or not (0).
* **11\_DescriptiveStats:** This takes in the cleaned dataframe from ‘01\_PreProcessing’ and performs descriptive statistics on the whole dataset.
* **21\_Meaning\_Analysis:** This file takes in the cleaned dataframe and performs TfidfVectorizer to convert the text into a matrix of TF-IDF features. It then takes the sum of each word’s term frequency value in a review (‘sum\_tf’) and appends it to ‘data’ from ‘01\_PreProcessing’. This sum is used as one dependent variable. The file further divides ‘data’ into 6 different datasets for each drug.
* **22\_Rating \_Analysis:** This file takes in the cleaned dataframe and appends the rating column to ‘data’ from ‘01\_PreProcessing’. The file further divides ‘data’ into 6 different datasets for each drug.
* **23\_Sentiment\_Analysis:** This file takes in the cleaned dataframe and performs ‘SentimentIntensityAnalyzer’ on each review. It makes a column ‘sentiment’ which takes in the polarity scores for each review. This column is then appended to ‘data’ from ‘01\_PreProcessing’. The file further divides ‘data’ into 6 different datasets for each drug.
* **311\_LASSO\_Meaning:** This file performs LASSO with the dependent variable as ‘sum\_tf’ and independent variables as the matrix of all words and word-pairs from ‘data’. The first LASSO is run with alpha = 0.1 to store the initial words. These words are then visualized for all possible Alphas between 0 and 1 with intervals of 0.01. Afterwards, based on the visualization the optimal Alpha is found using LassoCV with cv = 10. Then, the words from the optimal Alpha are visualized. These final words are divided and visualized into the most and least meaningful words. The files ‘312\_LASSO\_Meaning\_drug\_1’ to ‘317\_LASSO\_Meaning\_Drug\_3’ perform the same operations for each drug.
* **411\_LASSO\_Rating:** This file performs LASSO with the dependent variable as ‘Rating’ and independent variables as the matrix of all words and word-pairs from ‘data’. The first LASSO is run with alpha = 0.1 to store the initial words. These words are then visualized for all possible Alphas between 0 and 1 with intervals of 0.01. Afterwards, based on the visualization the optimal Alpha is found using LassoCV with cv = 10. Then, the words from the optimal Alpha are visualized. These final words are divided and visualized into the most positive and negative words. The files ‘412\_LASSO\_Rating\_Drug\_1’ to ‘417\_LASSO\_Rating\_Drug\_6’ perform the same operations for each drug.
* **511\_LASSO\_Sentiment:** This file performs LASSO with the dependent variable as ‘sentiment’ and independent variables as the matrix of all words and word-pairs from ‘data’. The first LASSO is run with alpha = 0.1 to store the initial words. These words are then visualized for all possible Alphas between 0 and 1 with intervals of 0.01. Afterwards, based on the visualization the optimal Alpha is found using LassoCV with cv = 10. Then, the words from the optimal Alpha are visualized. These final words are divided and visualized into the most positive and negative words. The files ‘512\_LASSO\_Sentiment\_Drug\_1’ to ‘517\_LASSO\_Sentiment\_Drug\_6’ perform the same operations for each drug.