P40 Project Report

Business Objective: To analyze the side effects and effectiveness of medicines from online drug reviews

Data Collection: Features such as medicine name, condition, reviews, ratings, date and useful count were collected from the internet and taken into consideration from the sources mentioned below:

- UCI Machine learning repository
- Drugs.com (web scraped using BeautifulSoup)

The final dataset consisted of an **214262** Rows and 6 columns

	Medicine Name	Condition	Reviews	Ratings	Date	Useful Count
0	Mirtazapine	Depression	"I've tried a few antidepressants over the years (citalopram, fluoxetine, amitriptyline), but no...	10.0	28-Feb-12	22
1	Mesalamine	Crohn's Disease, Maintenance	"My son has Crohn's disease and has done very well on the Asacol. He has no complaints and show...	8.0	17-May-09	17
2	Bactrim	Urinary Tract Infection	"Quick reduction of symptoms"	9.0	29-Sep-17	3
3	Contrave	Weight Loss	"Contrave combines drugs that were used for alcohol, smoking, and opioid cessation. People lose ...	9.0	05-Mar-17	35
4	Cyclafem 1 / 35	Birth Control	"I have been on this birth control for one cycle. After reading some of the reviews on this type...	9.0	22-Oct-15	4

Exploratory Data Analysis: (using pandas, numpy, matplotlib and seaborn)**.** Following were some of our findings:

- There is an upward trend of medicines being reviewed online
- Topmost reviewed condition is for a birth control medicine called levenorgestrel
- There were a greater number of positive reviews than negative reviews in the dataset

Cleaning the reviews:

- Imputation: Missing ratings were imputed using average imputation method
- Sentiment Analysis: The sentiment on the reviews were classified into positive and negative based on the ratings. If the rating on a scale of 10 is greater than or equal to 5; the review is considered as positive else, it is considered negative
- Text pre-processing (using NLTK and spacy libraries): Unstructured text reviews were pre-processed by converting to lowercase, correcting misspelled words; removal of whitespaces, stop words and unwanted html tags and other special characters
- Lemmatization: To avoid repetition, we lemmatize the words which means consider words with similar meanings as one word
- Tokenization: Cleaned reviews are then broken down into tokens/words
- After the reviews were cleaned, negative and positive bi-gram and tri-gram visualizations were created to find relevant and meaningful words

Building the Model:

Side effects: The topic modelling technique has been used to extract the side effects from the cleaned reviews. Latent Dirichlet Allocation (LDA) method has been applied to negative reviews to extract the side effects for each medicine.

The cleaned reviews are structured to form a document. The aim of LDA is to find topics a document belongs to. Each of the topics contains words with a probability score given to it using the genism library. From this we choose the most relevant topic.

Effectiveness: How effective is the medicine is calculated as a percentage.

Effectiveness %=(No. of positive reviews/Total number of reviews) *100

Evaluation: We choose the topics based on two measures.; a good topic should have low perplexity and high coherence score.
Perplexity: -6.562929451880235

Coherence Score: 0.8520706871504424

Deployment: We have used streamlit for creating the user interface/web application. The code is run using anaconda prompt in order to display the 'medicine effects analyzer' app. Simply select a medicine name from the drop-down list as shown below. This will display a list of side effects and 2 wordclouds. The biggest word in the wordcloud indicates that it is the most common side effect of that medicine. The effectiveness of a medicine is also displayed for each condition as a percentage.