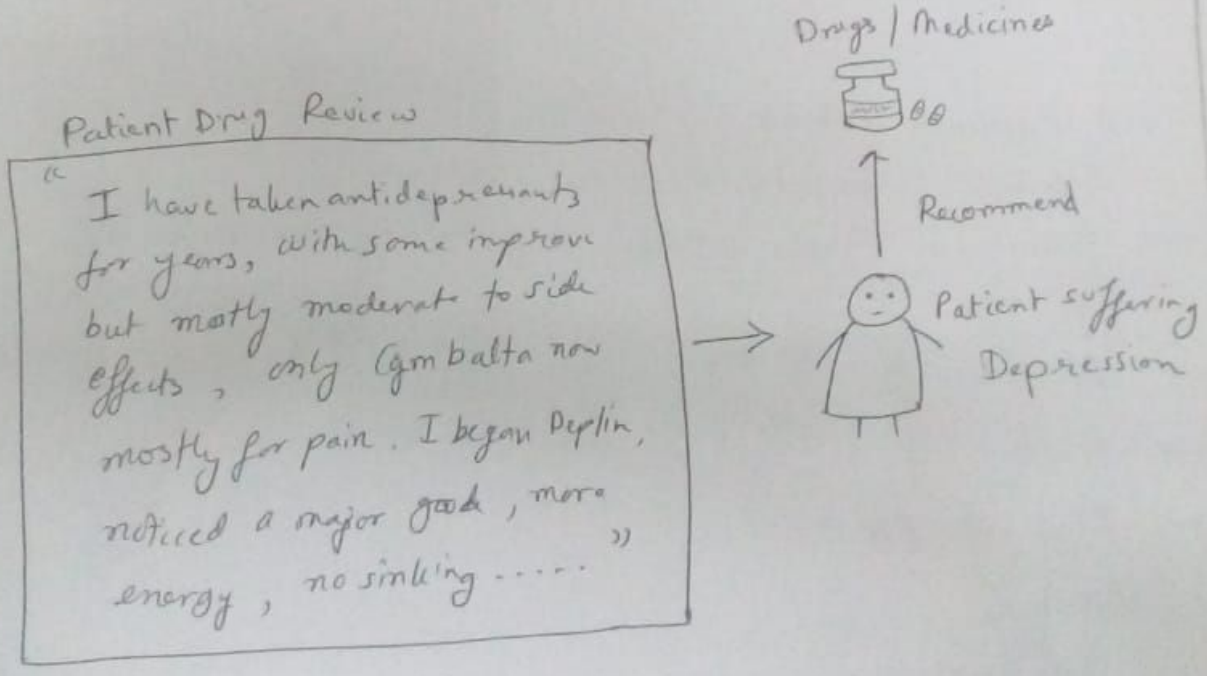
DA I

20MIC0063

Ananya Ghosh.

Project Title : Patients' Condition Classification and Drug recommendation based on drug reviews.

Abstract : Analyzing the patient reviews, we can get understanding of drug effectiveness and its side effects. We will classify the condition of patient based on his review so that we can recommend him a suitable drug



Steps of NLP Pipeline :

→ Tokenize sentences
→ clean reviews
  ↳ Remove punctuation
  ↳ Remove special characters / numbers
  ↳ Convert to lowercase
  ↳ Lemmitization

→ Create a bag of words model to vectorize
→ Apply ML algorithms Naïve Bayes & Passive Aggregasive classifier.
→ Create TFIDF model to vectorize
→ Apply ML algorithms Naive and Passive Aggressive Classifier
→ Compare.

Dataset Explaination :

https://archive.ics.uci.edu/ml/datasets/Drug Review Dataset (Drugs.com)

→ Multivariate, Text
→ Attribute characteristic is integer
→ Associated Tasks are regression, classification, clustering
→ Number of instances is 215063
→ Number of attributes are 6.
→ 2018-10-04
→ Attributes
  ↳ drugName
  ↳ condition
  ↳ review
  ↳ rating
  ↳ date
  ↳ usefulCount

Libraries :

- pandas (data preprocessing)
- itertools (confusion matrix)
- string
- numpy
- sklearn . feature_extraction . text → CountVectorizer, TfidfVectorizer
- sklearn . model selection → train_test_split
- sklearn . linear_model → PassiveAggressiveClassifier
- sklearn . naive_bayes → MultinomialNB
- Sklearn → metrics
- seaborn
- matplotlib . pyplot
- wordcloud
- nltk → nltk . corpus → stopwords
  → nltk . stem → WordNetLemmatizer
  → nltk . stem → PorterStemmer
- bs4 → BeautifulSoup
- joblib

Working :

Drug reviews dataset → Data preprocessing → Train Test Split → Vectorization TFIDF → ML Model → Flask App

Flask App → Medical Condition

Flask App → Drug recommended

→ Import Libraries
→ Import tsv
→ train data
→ segregating dataframe for analyzing individual condition
→ word cloud
→ data preprocessing
  ↳ clean punctuations, special characters, numbers
  ↳ common stopwords.
→ Lemmatization
→ Creating features & target variables.
→ Bag of Words
→ Train Test Split
→ ML Model : Naive Bayes
  ↳ Confusion Matrix
→ ML model : Passive Aggressive Classifier
  ↳ Confusion matrix

→ TFIDF
→ ML model - Naive Bayes
→ ML model - TFIDF
→ TFIDF : Bigrams, Trigrams
→ Most Important Features
→ Sample Predictions.

Models :

→ Naive Bayes

→ Aggressive Passive Classifier

→ TFIDF – Bigram, Trigram

Need / Advantages of this application :

- Patients who need immediate drug recommendation can get advantage from this.
- Patients sharing their experience will help them to get known to what disease they have and what drugs they need.

Which NLP Application :

Text Extraction

NLP in healthcare

Challenges :

→ efficient data preprocessing, cleaning

→ maintaining model accuracy

→ train test effective split.

Demo :

Share review :

3rd med I tried for anxiety, depression and dizzy it is, not fuell good ...

Predicted Medical Condition :

Depression

Top 3 drug recommendation :

x | y | z