**Project**

**Identifying side Effects and Evaluating the effectiveness of Drugs**

Before accessing the files please go through the description.

**File name:** project 4 Drug extraction

**Description:**

The python code is used to extract data from webmd.com and the library that is used is the code is selenium and beautiful soup.

Output : Drug.csv

**File name:** project 4.0 Visualisations.ipybn

**Description:**

Important visualization of the data can be seen through this python code

Inputs:

* drugsComTest\_raw.csv
* drugsComTrain\_raw.csv

**File name:** project 4.1 Sideeffect.ipybn

**Description:**

The python code extracts data from drugs.com using selenium and beautiful soup

**Output : Important(please read this section)**

* Side-effect.csv: contains the list of side effect(extracted from drugs.com).
* side-effect\_S.csv: when the individual words present in the side-effect list is saved and the duplicated words are removed with all the preprocessing required.
* side-effec\_s.csv : Maually removed words from the (side-effect\_S.csv) file to get better results
* conditional\_probability.csv : this uses (side-effec\_s.csv ) file and (Side-effect.csv) to compute how many times the words present in the (side-effec\_s.csv ) file is repeated in (Side-effect.csv) file

**File name:** **project 4.2(filtered and n grams).ipybn**

**Description:**

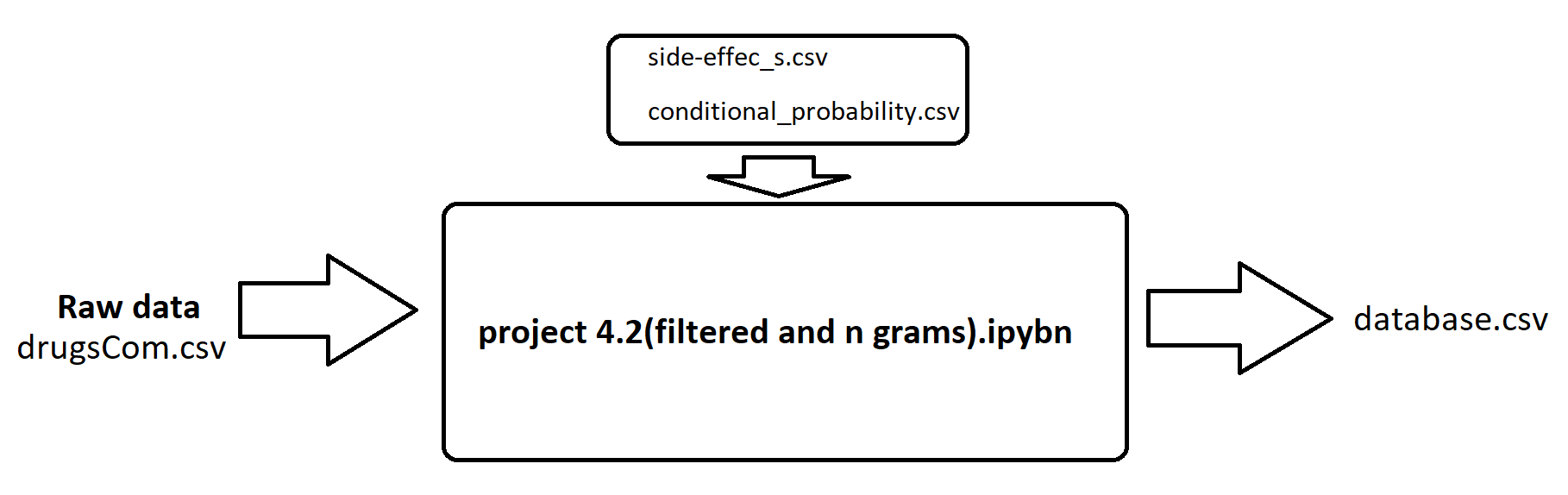
This is a core file which contains compares words from reviews to words in side- effect library and then join the words that exceeds the threshold probability (0.15).

Inputs:

* side-effec\_s.csv
* conditional\_probability.csv

output:

* database.csv: This is the final file and will be used represent all the required information. Also all the drugs in this file have number of reviews greater than 10.



**File name: Streamlit\_Deployment.py**

**Description:** The Python program is used for the purpose of deployment in streamlit

Inputs: database.csv