Knowledge Discovery and Data Mining

Mid-term project

General requirements:

- Using pure python in python notebook to solve the following tasks (Note: only use some common libs in python such as numpy, pandas…)

- Each group, only team lead submits the python notebook to Google classroom.

1. The Eclat algorithm (4 points)

Input: dataset with many transactions, minimum support.

Output: list of frequent itemset.

// Initial Call: $\mathcal{F} \leftarrow \emptyset, P \leftarrow \{\langle i, \boldsymbol{t}(i)\rangle \mid i \in \mathcal{I}, |\boldsymbol{t}(i)| \geq minsup\}$

Eclat $(P,\ minsup,\ \mathcal{F})$**:**

1 **foreach** $\langle X_a, \boldsymbol{t}(X_a)\rangle \in P$ **do**

2 $\quad \mathcal{F} \leftarrow \mathcal{F} \cup \{(X_a, sup(X_a))\}$

3 $\quad P_a \leftarrow \emptyset$

4 $\quad$ **foreach** $\langle X_b, \boldsymbol{t}(X_b)\rangle \in P$, *with* $X_b > X_a$ **do**

5 $\quad\quad X_{ab} = X_a \cup X_b$

6 $\quad\quad \boldsymbol{t}(X_{ab}) = \boldsymbol{t}(X_a) \cap \boldsymbol{t}(X_b)$

7 $\quad\quad$ **if** $sup(X_{ab}) \geq minsup$ **then**

8 $\quad\quad\quad P_a \leftarrow P_a \cup \{\langle X_{ab}, \boldsymbol{t}(X_{ab})\rangle\}$

9 $\quad$ **if** $P_a \neq \emptyset$ **then** Eclat $(P_a,\ minsup,\ \mathcal{F})$

10

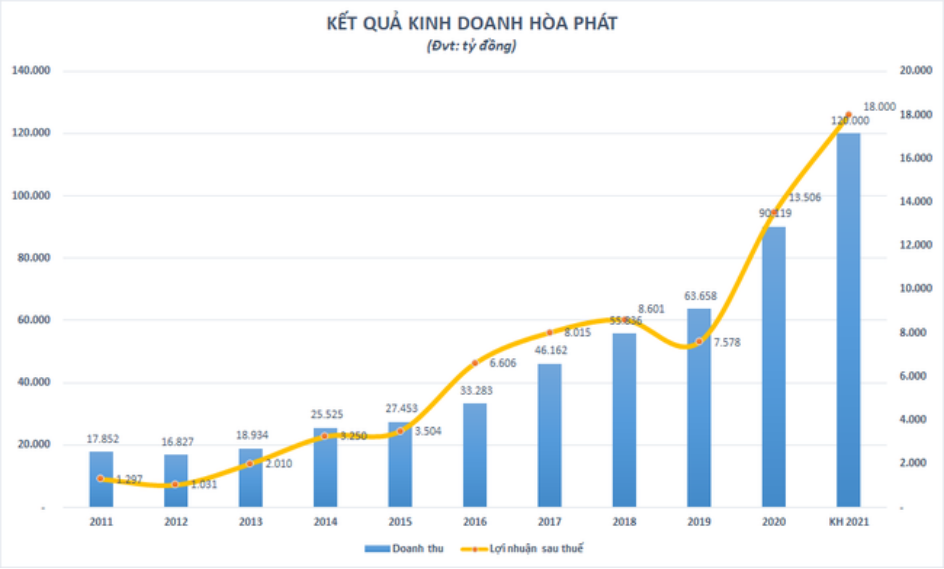
Algorithm explanation: F is result; P is 1-itemsets (whose supports are larger than minsup) with their tidsets.

2. Implement the crawling task (4 points)

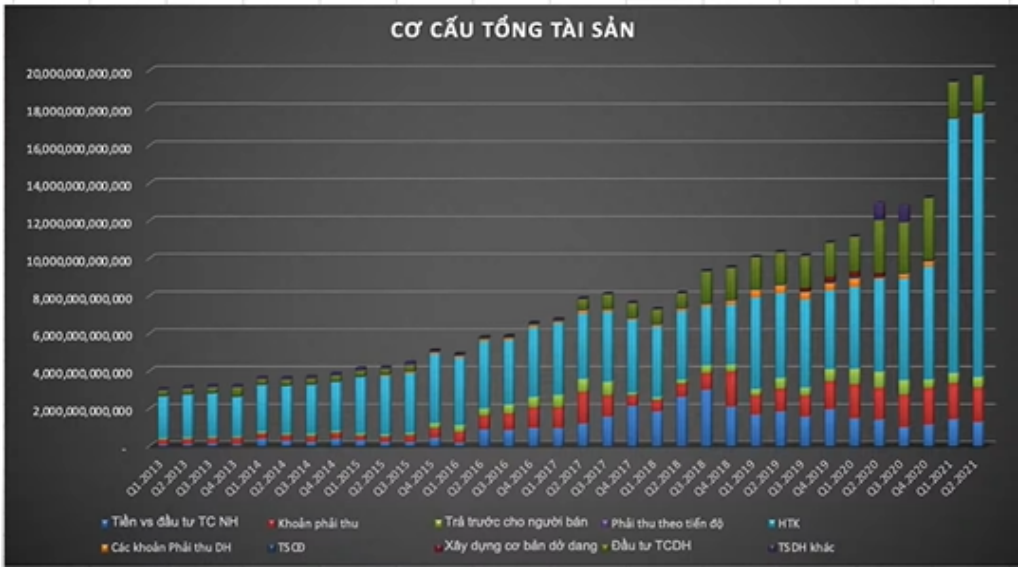
Input: stock symbol (e.g., HPG)

Process: Program goes to any stock website (such as cafef.vn, dstock.vndirect.com.vn, vietstock.vn…) to crawl the information to draw the output graph.

2.1. Revenue and profit in each year (2 points).



2.2. Structure of total assets of this company in each quarter (2 points)



3. Crawling and preprocessing text (3 points)

Input: A link of Vietnamese video (e.g., https://www.youtube.com/watch?v=6Q7mfthgMAU)

3.1. Write a python program to crawl the first 200 comments of this video (1 points).

3.2. Apply text preprocessing techniques and TF-IDF for feature extraction of the first 200 comments (2 points).