

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA KHOA HỌC MÁY TÍNH**





**REPORT**

DATA MINING

**TOPIC: Application of the Pattern Growth Approach for detecting spam mail**

**TEACHER**: *Vo Nguyen Le Duy*

**Student :**

Luong Anh Huy 22520550

Pham Dong Hung 22520521

Phan Cong Minh 22520884

🙡🙢 Ho Chi Minh, 05/06/2024 🙠🙣

# *Index*

[**Index 2**](#_heading=h.30j0zll)

[**I. INTRODUCTION 3**](#_heading=h.1fob9te)

[**II. EXPLORATORY DATA ANALYSIS 4**](#_heading=h.tkqetd23x8ax)

[**Spam mail Dataset: 4**](#_heading=h.ivqffvy0069h)

[**Preprocessing 5**](#_heading=h.86yccdcnv1pw)

[**III. METHOD 7**](#_heading=h.3whwml4)

[**FP-Growth Algorithm 7**](#_heading=h.3tpdjkwx19un)

[**Classification 8**](#_heading=h.btwba2mgv8i4)

[**Association Rules 8**](#_heading=h.abe393ida9qx)

[**Related Formula in Association Rules 9**](#_heading=h.2siuq7ugg7a0)

[**Processing Algorithm 9**](#_heading=h.gy81f0cl8hpw)

[**IV. RESULT AND ANALYSIS 11**](#_heading=h.1pxezwc)

[**Comparision with another model 11**](#_heading=h.gtyu9eh6nrg2)

[**Advantages and disadvantages of using FP Growth & Association Rules in combination compared to other popular classification algorithms 11**](#_heading=h.sphiichzkwps)

[**V. CONCLUSION 13**](#_heading=h.3o7alnk)

[**Frequent Pattern Growth: 13**](#_heading=h.fs383me7fc0x)

[**Some challenges and limitations: 13**](#_heading=h.8g5j8wg4g5bp)

[**Suggested development directions: 13**](#_heading=h.qn9774hbhp69)

[**VI. TASKS ASSIGNMENT 14**](#_heading=h.x4vxe384ykxj)

[**REFERENCE MATERIALS 15**](#_heading=h.ihv636)

# INTRODUCTION

- E-mail is an effective way of communication as it saves a lot of time and cost. This makes it a favourite means of communication in personal as well as in professional communications. Information can then be easily spreaded within the social media networks. Because of this, websites expose to various types of unwanted and malicious risks - especially spam mails. Spam mails is the practice of frequently sending unwanted data or bulk data in a large quantity to some email accounts. It has been estimated that around 70% of all emails are spam. As the usage of web expanding, problem of spam mails are also expanding. Email spam is any email that meets the following three criteria:

1. Anonymity: the address and identity of the sender are concealed.
2. Mass mailing: the email is sent to large group of people.
3. Unsolicited: the email is not requested by recipients.

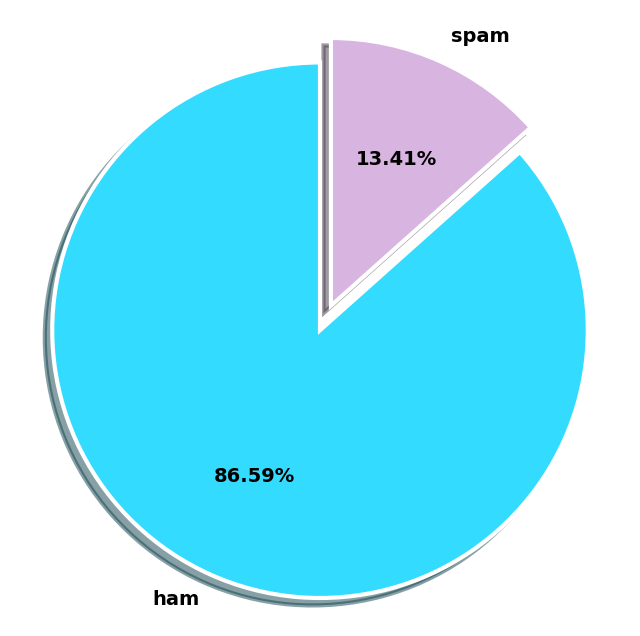
- So, if the email should be more secure and effective, appropriate email filtering is an essential process. Many methods have been proposed for classification of email messages as spam or legitimate mail and it has been found that machine learning algorithm success ratio for classification is very high. In this article, Pattern Growth Approach - one of the Data Mining algorithms is used for detecting Spam Mails.



# EXPLORATORY DATA ANALYSIS

## Spam mail Dataset:

* The **Spam Mails Dataset** on Kaggle is a valuable resource for those interested in spam email detection and analysis. This dataset consists of a substantial collection of email text messages, categorized as either spam or non-spam (ham).
* Dataset Overview:
* **Total mail:** contains 5572 emails in total
* **Spam and ham:** Out of these, 747 emails (13.41%) are classified as spam and 4825 (86.59%) are classified as ham
* **Features Included:** Each email includes several key features such as the email body, which is the main content used for analysis. Other attributes like the subject line and recipient details can also be included to enhance the analysis and detection processes.
* Researchers and data scientists can utilize this dataset to develop and refine models for spam detection. By analyzing the text and identifying patterns common in spam emails, these models can significantly improve the accuracy of spam filters used in email services.
* No missing values for **label** and **message** columns
* Percentage of spam and ham:



## Preprocessing

- Before applying the FP-Growth algorithm, the dataset needs to be preprocessed to convert the raw email text into a format suitable for mining frequent patterns. The preprocessing steps include:

1. **Loading Data**: We start by loading the dataset from the provided file path. The dataset contains emails labeled as spam or ham.
2. **Column Dropping and Renaming**: We select the relevant columns 'Label' and 'Body' and rename them to 'label' and 'message' for clarity.
3. **Message Cleaning**: Each email message is converted to lowercase, and non-alphabetic characters are removed using regular expressions. Extra whitespace is also trimmed.
4. **Duplicate Removal**: Duplicate email messages are dropped to ensure each email is unique in the dataset.
5. **Stop Words Removal**: Stopwords is common words that do not contribute to distinguishing between spam and ham (e.g., "the", "and", "is") are removed using the NLTK stopwords list. Each message is divided into a list of words with out stopwords.
6. **Label Mapping**: The labels are mapped from numerical values (0 and 1) to categorical values ('ham' and 'spam') for better readability.
7. **Data splitting:**

- Splitting the dataset into train and test sets, where 30% of the data is reserved for testing.

- Sampling 10% of the training data to reduce computation time or to work with a smaller subset for quick prototyping.

- Splitting the sample data by label. The **'message'** column for rows with a spam label (1) and a ham label (0) is split into lists of words (tokens), and then converted into a list of lists, representing transactions.

- Generating unique items and creating a binary matrix for spam and ham transactions:

+ A set of unique items is created from all of the spam transactions. This set is sorted to maintain a consistent order.

+ A binary matrix is formed by going through each transaction and marking '1' if the unique item is present in that transaction, else '0.' This is stored in a DataFrame df\_te\_spam with columns representing unique items.

For further information and to access the dataset, you can visit the[Spam Email (kaggle.com)](https://www.kaggle.com/datasets/mfaisalqureshi/spam-email/data).





# METHOD

Frequent Pattern Growth Approach & Association Rules:

## FP-Growth Algorithm

**How it Works:**

- A popular method for frequent pattern mining in data mining. It works by constructing a frequent pattern tree (FP-tree) from the input dataset. The FP-tree is a compressed representation of the dataset that captures the frequency and association information of the items in the data. The key steps are:

* **Construction of the FP-Tree**:
* The algorithm begins with a scan over the transactions to identify and sort frequent items in descending order based on their frequencies. These items are then used to construct the FP-tree.
* Subsequent transactions are processed by adding their items (following the order determined previously) to this FP-tree. Shared prefixes of itemsets lead to shared paths in the FP-tree, compacting the representation.
* **Mining the FP-Tree**:
* Starting from each frequent item, the FP-tree is traversed to find all conditional pattern bases (subsets of transactions that contain the frequent item).
* A new conditional FP-tree is then constructed for each pattern base, and the process is repeated recursively to extract frequent patterns.

**Input and Output of FP-Growth**

**-** Before classifying emails as spam or ham, we apply the FP-Growth algorithm to mine frequent patterns from the preprocessed email data. This involves encoding the email messages into a binary format suitable for the FP-Growth algorithm and then extracting frequent itemsets.



**-** The input to the FP-Growth algorithm consists of a binary matrix represented for spam mail and ham mail.

**-** The output of the FP-Growth algorithm is a set of frequent itemsets, which are groups of words that commonly appear together in the spam and ham mail. These itemsets are characterized by their support, which indicates how frequently they appear in the dataset

## Classification

- To classify emails as spam or ham, we use the frequent itemsets mined from the FP-Growth algorithm to create association rules. These rules are then applied to new email messages to determine their classification

### Association Rules

**How It Works:**

- Find all frequent itemsets: Using algorithms like FP-Growth, all itemsets that satisfy the minimum support threshold are identified.

- Generate rules from the frequent itemsets: For each frequent itemset, all possible non-empty subsets are generated. For a frequent itemset and a non-empty subset of , an implication rule of the form is generated, provided the rule satisfies the minimum confidence threshold.

**-** After extracting frequent itemsets using FP-Growth, we generate association rules to understand the relationships between items (words) in the dataset. These rules help in classifying email messages by identifying patterns that are strongly associated with spam or ham.

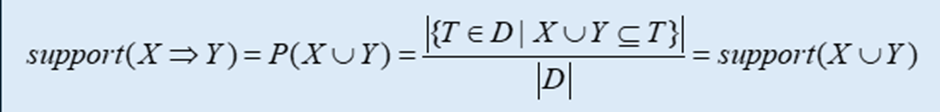
**-** The Input of the Association Rules:

* Frequent Itemsets: Derived from the FP-Growth algorithm, consisting of groups of words and their support values.
* Binary Matrix: The original binary matrix used to compute the frequent itemsets, necessary for calculating additional metrics like confidence

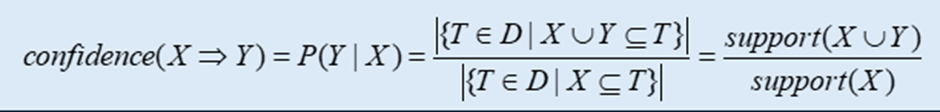
- The Output of the Association Rules: Consisting of antecedents, consequents, support, and confidence values, potentially including additional metrics. These rules help in understanding the relationships between items in the dataset and are used for classification purposes.

### Related Formula in Association Rules

- **Support**: frequency (probability) of the entire rule with respect to database:



-**Confidence**: indicates the strength of implication in the rule:



## Processing Algorithm

**Data Representation:**

The data is represented in the form of transactions and corresponding items.

Each message is split into individual words, and each transaction contains a set of words appearing in each message.

**Finding Frequent Patterns:**

FPGrowth algorithm is used to find frequent patterns in the data, for both the spam and ham groups.

This helps identify sets of words that frequently appear together in the messages.

**Generating Association Rules:**

Based on the frequent patterns found, association rules are created to describe the relationships between words.

These rules consist of a set of conditions (antecedents) and a consequent, representing the relationship between words in messages.

**Classifying New Messages:**

Based on the created rules, each message in the test set is classified by comparing the words in the message with the conditions in the association rules.

A message is classified as spam if it contains the conditions of spam association rules with higher confidence compared to ham association rules, and vice versa.

# RESULT AND ANALYSIS



## Comparision with other models

| **Method** | **Accuracy** | **Precison** | **Recall** | **F1\_score** | **CEL** |
| --- | --- | --- | --- | --- | --- |
| **FP-Growth & AR** | 0.906 | 0.78 | 0.82 | 0.79 | 0.6967 |
| **Linear Regression** | 0.972 | 0.98 | 0.89 | 0.93 | 0.0913 |
| **MultinomialNaive Bayes** | 0.964 | 0.90 | 0.93 | 0.92 | 0.1526 |

## Advantages and disadvantages of using FP Growth & Association Rules in combination compared to other popular classification algorithms

- **Logistic Regression** and **Multinomial Naive Bayes** are popular algorithms for classification, and they work well on both discrete and continuous data.

- **FP Growth & Association Rules**, by nature, are not typically used for classification problems. However, they do have some characteristics that can support email classification:

* **Ability to handle large datasets**: FP-Growth can efficiently process large datasets to find frequent itemsets.
* **Generating association rules from these frequent itemsets about the co-occurrence of words in emails**: These rules can be used to infer the likelihood of an email being spam or ham.

- However, based on the performance metrics, it is evident that FP-Growth & Association Rules do not yield results as good as Logistic Regression and Multinomial Naive Bayes, due to several reasons:

* **Large Number of Rules and Difficult to Manage**: When data is large and complex, the number of frequent itemsets and association rules can become very large and difficult to control. This makes it difficult to find rules that have real value in classification task.
* **Need to tweak the parameters**: To determine which itemsets or rules to choose, support and confidence thresholds need to be established first. The choice of those is not always obvious and can significantly affect the mining results.
* **Poor performance in handling imbalanced data**: The spam email dataset often has an imbalance between classes, with spam emails usually being much fewer than non-spam emails. Frequent pattern growth combined with association rules can struggle with handling this imbalance, leading to poor classification performance.

# CONCLUSION



Although Frequent Pattern Growth and Association Rules is not a common method in the field of classification in general, they still shows relative effectiveness when it comes to classification tasks, particularly in this case of text processing and feature extraction. ​

## Some challenges and limitations:

- Lack of training on multiple datasets to capture the diversity in semantics and context of words and sentences.

- Not flexible, with lower classification accuracy when encountering new, unseen data.

## Suggested development directions:

- Implement more effective data preprocessing steps, possibly deriving some new artificial features from the existing ones to enhance data analysis.

- Train on multiple datasets to increase the model's learning capability.

- Develop a classification model for other languages, such as Vietnamese.

# 



# TASKS ASSIGNMENT

# 

| **Name** | **Role** |
| --- | --- |
| Luong Anh Huy | Slide, Code |
| Pham Dong Hung | Report, Code |
| Phan Cong Minh | Text Documents, Code |

# 

# 

# 

# 

# 

# 

# 

# 

# REFERENCE MATERIAL

**English**

1. [**Spam Email Detection||Naive Bayes (kaggle.com)**](https://www.kaggle.com/code/madhvisaxena/spam-email-detection-naive-bayes)
2. [**16. 75-79.pdf (iosrjen.org)**](https://iosrjen.org/Papers/Conf.ICCIDS-2018/Volume-8/16.%2075-79.pdf?fbclid=IwZXh0bgNhZW0CMTAAAR0nfxO_boCBy715RHR3k2AzhRl5AbSr4pt6kf5Id_y5c941z9xVOnESYRo_aem_Aearl-Q2jNdClUlPiDe_X0yag-4XDVDGpVy5S6qZhHeDS-eE35mHcnm9b0KkzMgzl960Actig-HxetCxQwe_8Rvx)
3. [**IJMECS-V5-N12-5.pdf (mecs-press.org)**](https://www.mecs-press.org/ijmecs/ijmecs-v5-n12/IJMECS-V5-N12-5.pdf?fbclid=IwZXh0bgNhZW0CMTAAAR2py2_TiohmeCgHuIzxrTirJ92R9GoFOCU2Dv4oD2hbK1YBE2EEfp9jFC0_aem_AebFui6edOo6shJ-oFRjcTOa0RV2jhY2lAA3up7l-GQZ_NWjBX-XfxYh5lv6rCnDmBQNRkSjskgLRcFgP3Fbw3hd)

**FP-Growth:**

1. [**mlxtend/mlxtend/frequent\_patterns at master · rasbt/mlxtend · GitHub**](https://github.com/rasbt/mlxtend/tree/master/mlxtend/frequent_patterns)

**Association rules:**

1. [**Apriori Algorithm Program in Python from Scratch - ja**](https://japp.io/machine-learning/apriori-algorithm-program-in-python-from-scratch/)