DATA MINING GROUP

ADVISOR: PEERAPON VATEEKUL, PH.D.

DEPT. OF COMPUTER ENGINEERING

FACULTY OF ENGINEERING, CHULALONGKORN UNIVERSITY
Advisor

- Peerapon Vateekul, Ph.D.
  - Email: peerapon.v@chula.ac.th
Graduate Members

- **Natchanon Phachongkitphiphat (Nat)**
  Master of Engineering: Computer Engineering (CM)
  Expected Graduation: 2015

- **Suthipong Daengduang (Top)**
  Master of Engineering: Computer Science (CS)

- **Teerawit Choeikiwong (Beam)**
  Master of Science: Software Engineering (SE)
  Expected Graduation: 2015

- **Tanawat Limungkura (Mum)**
  Master of Engineering: Computer Science (CS)
Undergraduate Members

- B.Eng (CP) Expected Graduation: 2015
  - Nuttapon Pattanavijit (Nut)
  - Vibhavee Trairattanapa (Vee)
  - Kankawin Kowsrihawat (K)
  - Panida Nimnual (Chompoo)
  - Piyawat Lertvittayakumjorn (Tui)
Projects
Large-scaled Hierarchical Multi-label Classification

Objective: Propose new methods for addressing hierarchical multi-label classification problem which outperform the existing standard algorithm in terms of time and accuracy.

Challenges: Large datasets, Multi-label, Tree/DAG class structure, Full/partial depth labeling

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Train docs</th>
<th>#Test docs</th>
<th>#Leaf classes</th>
<th>#Features</th>
<th>Max Depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wiki Medium</td>
<td>456,886</td>
<td>81,262</td>
<td>36,504</td>
<td>346,299</td>
<td>12</td>
</tr>
<tr>
<td>Wiki Large</td>
<td>2,365,436</td>
<td>452,167</td>
<td>325,056</td>
<td>1,617,899</td>
<td>729</td>
</tr>
</tbody>
</table>

Proposed method:
Hierarchical Multi-Label Classification for Complex Hierarchical Structure

- **Challenges:** Complex Hierarchical Structure, Directed Acyclic Graph (DAG) and Non-Mandatory Leaf Node (NMLN)

**Dataset:** PubMed Medical Document Database 12,628,968 Document

**Class Hierarchy:** Medical Subject Heading 26,831 Words

**Hierarchical Multi-Label**

- o belong to f then o belong to B,C
Software Defect Prediction
Using Unbiased Support Vector Machine

- **What?** This research aims to achieve high defect prediction performance in terms of PD, PF, F1, and G-mean.

- **Why?** Prediction performance was limited due to the imbalanced issue since the number of defected modules is very small comparing to that of non-defected modules.

- **How?** By employing unbiased Support Vector Machine called “R-SVM” to tackle the imbalanced issue. In R-SVM, the separation hyperplane is adjusted to reduce a bias from the majority class (non-defect).
Gist Extraction System on Thai Supreme Court Cases

- **Goal:** Prototype Information Extraction and Automatic Summarization System for Thai Supreme Court Judgment
- **Benefit:** Help on matters of law summarization and useful judgment selection for legal research
- **Create as Web-based Application**
Twitter Data Dispersion Analysis

- **Objective:** Apply suitable data mining techniques for analyzing dispersion characteristics of Twitter data. Also, develop web application processing and reporting results for user.

- **Focus on:** Political Domain – Influence User, Fake news
Data Quality Management for Hydrological Data

- **Objective**: Devise a system to capture and notify inconsistency in hydrological data. Moreover, research new methods to enhance detecting flawed cumulative rain level data.

- **Common Tasks**: missing value imputation, outlier detection, data consistency check, and homogenization