DATA SCI 8410: Data Mining & Information Retrieval

Tozammel Hossain

Data Science & Analytics
University of Missouri

Basic Info

- **3 Credit Hours**
- **Prerequisites**
  - Data Sci 7020: Stat & Math Foundations for Data Science
    - Or instructor's consent
- **Relevant Computing Skills**
  - Python Programming
    - Using Python packages to learn data mining
  - Familiarity with DSA computing environment + Git
    - Please check module 0
  - If you need help on Python, DSA environment and Git, please send me an email

Learning Objectives

- **Understand the basic concepts, principles, and methods in data mining and information retrieval**
- **Obtain hands-on experience using data mining and information retrieval toolkits**
- **Recall important pattern discovery concepts, methods, and applications**
- **Apply various document analytics methods including traditional (e.g. clustering and classification) and advanced (e.g., topic modeling and sentiment analysis) to various data sets**

Course structure

- **We aim to employ a "flipped classroom"/ mostly asynchronous model**
  - A module(video lectures, slides, coding practices) will be posted online **each sat at 6:00 am** and you are expected to digest this information before class (on-campus)/office hours (online)
  - Assignment is due by the **following Sat 11:59 PM**
  - Class times/Office hours will be used to reinforce these concepts and for exploratory discussions
- **Your active participation is important for the success of this model!**

Tools/Services

- **Canvas**
  - Announcements + grades
- **Jupyterhub**
  - Computing enviornment
- **Git**
  - Committing code
- **Communication**
  - Slack (online cohort)
  - Teams (on-campus cohort)
  - Office hours

Office Hours

- **Only for online cohort**
  - Thu, 7:30 PM – 8:30 PM
- **On-campus**
  - One in-person session per week
    - Location: Zoom
    - Time: TBA

Module Layout

- **8 modules**
  - delivered over 8 weeks
  - 7 modules for concepts + 1 module for mini project
- **No textbook is required**
- **Each module has**
  - Readings
  - Labs (examples done within the script)
  - Practices (solutions are provided separately)
    - Grades for completion (2 points)
  - Exercises
    - Graded for accuracy (18 points)
- **Score distribution**
  - Practices (10%) + Exercises (65%) + Project (25%)
  - Project: a problem will be given
    - Need to develop individually

Road Map

- **Module 1: Introduction & Classification**
  - Overview of DM and IR methods
  - Ensemble Learning/Meta Classifier
  - Multi-label Classification
- **Module 2: Advanced Clustering**
  - Similarity measures
  - Semi-supervised Learning
  - Fuzzy Clustering
  - Biclustering
- **Module 3: Frequent pattern mining - Part I**
  - Itemset mining
    - Apriori, FPGrowth
  - Summarizing itemsets

Road Map

- **Module 4: Frequent pattern mining - Part II**
  - Sequence mining
  - Graph pattern mining
  - Pattern and rule assessment
- **Module 5: Information Retrieval models - Part I**
  - Web scraping
  - Regular expression
  - Vector-space model
  - Language models
- **Module 6: Information Retrieval models - Part II**
  - Indexing
  - Search Engines
  - Link analysis
  - Retrieval evaluation

Road Map

- **Module 7: Advanced document analytics**
  - Topic modeling
  - Sentiment analysis
- **Module 8: Recommender system**
  - Popularity-based recommendation
  - Content-based recommendation
  - Collaborative Filtering
  - Mini Project