CSE 464 Data Science

TERM PROJECT – SMS Spam Classification

Due: 5-May-2019 - Submit to COADSYS before the class.

In this project, you are required to develop a Python notebook application based spam classifier from the SMS Spam Collection dataset. In order to get a full credit, you need to cover the following operations:

- Feature Engineering: Define at least 5 derived features from text message such as text length, digit num, etc.
- Exploratory Data Analysis
- Data Preparation (Data Cleaning, Handling Outliers, etc.)
- Text Vectorization (CountVectorizer, TfidfVectorizer, Word2Vec, etc)
- Model Building with at least 3 different classifiers
- Ensemble Learning with the models built (Voting Classifier)
- Parameter Tuning
- Comparative Performance Analysis (Confusion Matrix, Accuracy, F-Measure, AUC, etc.)
- Conclusions (Report best performing classifier, features, etc.)
- Present your solution and make a working demo

You are welcomed to ask any questions until deadline. Undergraduate students may form project groups. Please, email your project group members (max 3 for undergraduate) information until April 1st.

Good Luck!