NLP (Text) Assignment

Week #2

Submission Deadline: 12 September 2018

3 or 4 students for each group

SPAM FILTERING

- Find a spam corpus (email or sms) available in internet
  - Divide the corpus into training data and testing data
- Build a system (might be separated modules) consists of:
  - preprocessing (might be only tokenization),
  - feature extraction (might be added by feature selection – stop words elimination, TFxIDF rank, Mutual information rank),
  - Train Spam classification model using machine learning algorithms (SVM/ decision tree/naïve bayes/random forest)
  - All modules can be taken from available library (such as nltk, opennlp, sklearn, etc)
- Conduct several experiments based on the preprocessing/feature extraction/machine learning using testing data

DELIVERABLES

- Data description
  - The url,
- Describe the system architecture (modules) and the library used
  - Give an example of spam text, processed in each module
- Describe the experiment scenarios
- Write experimental results
- Write the analysis of incorrect classification result