National University of Singapore
School of Computing
Tutorial 8:
Text Mining 2

QUESTION 1: UPVOTED KAGGLE DATASETS (Topic Modelling)

- Kaggle dataset has become a popular place to share datasets.
- Almost every day there are new datasets uploaded.
- We need to explore what can be extracted from the information of each dataset.
- Discover the abstract "topics" based on the content of the DESCRIPTION field, using the following two methods:
    a. LDA
    b. Non-negative matrix factorization
- The dataset to be used has been provided and it includes:
  voted-kaggle-dataset.csv

QUESTION 2: SMS-SPAM-COLLECTION-DATASET

- The SMS Spam Collection is a set of SMSs tagged messages that have been collected for SMS Spam research.
- It contains one set of 5,574 English SMS messages, tagged according being **ham** (legitimate) or **spam**.
- The files contain one message per line.
- Each line is composed of two columns: v1 contains the label (ham or spam) and v2 contains the raw text.
- Use this dataset to build a prediction model that will accurately classify which texts are spam.
- The dataset to be used has been provided and it includes:
  spam.csv

QUESTION 3: Lebanese Arabic Reviews (OCLAR) Data Set

- This is a set of For Arabic sentiment classification on service reviews, including hotels, restaurants, shops, and others.
- It contains 3916 reviews in 5-rating scale. The positive class considers rating stars from 3-5 of 3465 reviews, and the negative class is represented from values of 1-2 of about 451 texts.
- The files contain one review per line.
- Each line is composed of three columns: v1 contains the service, v2 contains the text review and v3 contains the rating (5-rating scale)
- Use this dataset to build a prediction model that will accurately classify which reviews are positive (rating 3-5) and which are negative (rating 1-2).
- The dataset to be used has been provided and it includes:
  oclar.csv