National University of Singapore
School of Computing
BT5151 Foundation in Data Analytics II

Tutorial 7:
Text Mining 1

QUESTION 1: TWITTER SENTIMENT ANALYSIS

- The objective of this case study is to detect hate speech in tweets.
- For simplicity, a tweet contains hate speech if it has a racist or sexist sentiment associated with it. So, the task is to classify racist or sexist tweets from other tweets.
- Formally, given a training sample of tweets and labels, where label '1' denotes the tweet is racist/sexist and label '0' denotes the tweet is not racist/sexist, your objective is to predict the labels on the test dataset.
- Full tweet texts are provided with their labels for training data.
- The mentioned user's username has been replaced with @user.
- The dataset to be used has been provided and it includes:
  twitter-train.csv, twitter-test.csv

QUESTION 2: WHAT IS COOKING?

- Picture yourself strolling through your local, open-air market. What do you see? What do you smell? What will you make for dinner tonight?
- If you are in Northern California, you will be walking past the inevitable bushels of leafy greens, spiked with dark purple kale and the bright pinks and yellows of chard.
- Across the world in South Korea, mounds of bright red kimchi greet you, while the smell of the sea draws your attention to squids squirming nearby.
- India's market is perhaps the most colourful, awash in the rich hues and aromas of dozens of spices: turmeric, star anise, poppy seeds, and garam masala as far as the eye can see.
- Some of our strongest geographic and cultural associations are tied to a region's local foods.
- You are required to predict the category of a dish's cuisine given a list of its ingredients.
- You will need to install **wordcloud** package in anaconda via Anaconda Prompt:
  conda install -c conda-forge wordcloud
- The dataset to be used has been provided and it includes:
  cuisine-train.json, cuisine-test.json, sample_submission.csv

QUESTION 3: IMDb Movie Review

- This dataset contains binary sentiment classification of 25,000 highly polar movie reviews for training, and 25,000 for testing.
- Reviews are either positive (marked as '1') or negative (marked as '0').
- The reviews are web scratched so text pre-processing is needed.
- Explore the data.
- Try different models on this dataset and find the best model for solving this problem.
- You will need to install **unidecode** package in anaconda via Anaconda Prompt:
  conda install -c anaconda unidecode
- The dataset to be used has been provided and it includes:
  movie_data.csv