BT5152

AY2018/2019 Semester 1 Week **8**

About TA

KEE Yuan Chuan (YC)

Software Engineer

- Graduated from NUS in 2011 with B.Eng (Hons) in Chemical Engineering
- Currently, M.Comp (Computer Science) student and a Data Engineer by day
- Primarily in area of data engineering and analytics
    - Text analytics, geospatial analytics, data infrastructure, business intelligence and visualisation
    - Working on my dissertation on deep learning applications in areas of natural language processing

Tutorial Format

- Before tutorial preparation
  - Week 8 swirl (https://github.com/weilu/BT5152/https://github.com/kylase/BT5152)
  - post-Week 8 pre-class practice code
- During tutorial
  - Key Concept Revision
  - Self-practise with swirl [20 to 30 mins]; free
  - Additional tips/demonstration not covered in swirl
    - Will be useful for assignments and project!
    - Code will be uploaded to the Github (above link) after tutorials
- After tutorial: yuanchuan@u.nus.edu and cc. Prof. Huang
  - Best is to make use of IVLE forum

W8: Text Mining

Key Concepts

Basic Text Mining

- Language
- Pre-processing
- Dictionary approach
- Text Classification


adjective
object
The quick brown fox jumps over the lazy dog
verb
subject

Relating to what you have learnt so far...

- You have learnt that feature engineering is an important process for machine learning
  - Examples of feature engineering on tabular dataset
    - Removal of columns with near zero variance
    - Creating dummy variables to represent categorical variable
- How do we do feature engineering for text dataset?
  - What constitutes as features?

Pre-processing

- Bag-of-words (BoW) representation
  - What is the disadvantage of representing text in this form?
- Normalisation, stemming, removal of specific words
  - How will this affects the BoW representation?
- N-gram tokenisation
  - Capture highly co-occurring words, e.g. “data science”, “machine learning”, names
  - What will this do to BoW?

Dictionary Approach

Allows us to understand text-based data quantitatively

- Use established list of words that have sentiment in qualitative or quantitative representation
- Simple, fast but may not be reliable due to the complexity of language
- How to we handle negation (e.g. not happy ⇒ negative)?

Text Classification

Allows us to understand text-based data

- Require manual effort to label the data
  - What issue could arise from this?
- Better outcome than dictionary method and more versatile

Things we covering today

1. Self-practice Swirl [20 to 30 mins]
    a. Word Cloud
    b. Sentiment Analysis using Naive Bayes Classifier
2. Speed up data processing of text-based data
3. Building a sentiment predictor

Take home messages

- Feature Engineering for text mining starts at pre-processing
- Text processing is highly parallelizable
- Intrinsic method to manage metadata

To think about...

For the interested

- How can we represent words and their meaning in a more precise manner?
- How do we differentiate “line” (geometry) from “LINE” (company)?
- How to we capture context?