Machine Learning and Deep Learning
Assignment1

Yung-Chun Chang, Ph.D.

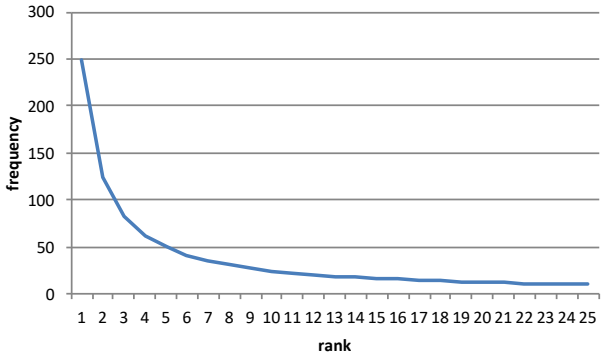
Graduate Institute of Data Science, Taipei Medical University, Taiwan.
Email: changyc@tmu.edu.tw

For Local Students

- Write a program to construct dictionary of corpus ChineseDataset_Assignment1.txt. (One row indicates one news article, there are about 8000 news documents) You have to do:
  - Preprocessing: segmentation, stopword removal & remove punctuation.
  - Sort terms by term frequency and draw a figure to prof they follow Zipf‘s law (long-tail distribution).
  - Rank terms by global TF-IDF.
  - Save the result as a txt file.

Result format

Term	G_TF-IDF
勇士	**5.83**
金塊	**4.55**
大勝	**3.88**



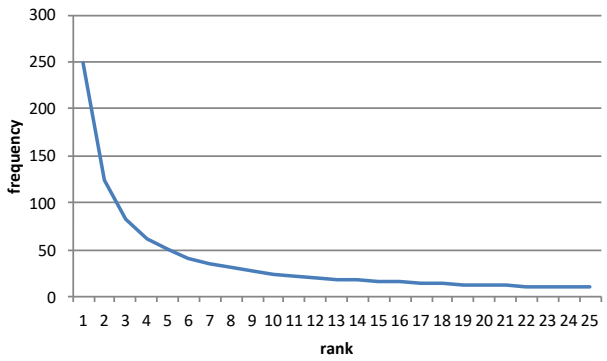
  - **Please zip and submit to My2TMU, including (1) the result and a Figure of Zip’s law, (2) the source code.**
    - 3 weeks to complete, that is **2018/10/23 12:00**.

For International Students

- Write a program to construct dictionary of corpus EnglishDataset_Assignment1.txt. (there are 25000 IMDB movie reviews) You have to do:
  - Preprocessing: tokenization, stopword removal, remove punctuation, and stemming (simple normalization if needed).
  - Sort terms by term frequency and draw a figure to prof they follow Zipf‘s law (long-tail distribution).
  - Rank terms by global TF-IDF.
  - Save the result as a txt file.

Result format

Term	G_TF-IDF
like	7.33
love	5.35
awesome	4.28



  - **Please zip and submit to My2TMU, including (1) the result and a Figure of Zip’s law, (2) the source code.**
    - 3 weeks to complete, that is **2018/10/23 12:00**.