Aiming A lets gooooooooooo

Meeeeeeeeeeeepon

Language: Python? Jupyternotebook?

Ideally complete, Week 9, Sat 17 Oct 11am

GitLab, a group so we can have multiple projects in a group

* rtkleong10
* nanyoci
* Kailing231 (IM IN YESSSSS)

Deadline = 2nd Nov

# 3.1 Domain Specific Dataset Analysis (20 marks) - Nan, Tiffany

* “easiest”
* three (3) datasets for analysis. Each dataset should contain about 20 documents in one selected topical domain.
* Stack overflow
* Research paper in <insert domain here> (NTU OneSearch)
* Patents.google.com?? (I not sure which kind)

Patents

Patents are frhttps://patents.google.com/

Tokenization and Stemming. Tokenize all documents in each domain using a selected library (e.g., NLTK) and observe the tokens obtained. Discuss your observations from the following perspectives.

Has the tokenizer correctly recognized the domain specific tokens? Use examples to illustrate what the expected tokens are, and what are not.

Discuss how to identify the tokens that are incorrect through programs?

If you were to improve the tokenizer, what are the possible solutions.

Perform stemming and compare the token distributions before and after the stemming (you may choose any stemming algorithm implemented in any toolkit). You may compare the number of distinct tokens, and the length distribution of the tokens.

The length distribution can be compared in a plot: the x-axis is the length of a token in number of characters, and the y-axis is the number of tokens of each length. Discuss your findings.

# 3.2 Development of a <Noun – Adjective> Pair Ranker (20 marks) KL, sherna

## Get about 20 to 30 reviews for one particular product

* \*manually\* go through the reviews and identify the most meaningful 5 pairs of <noun (or noun phrase) - adjective (or adjective phrase)> pairs
  + Clothes
  + Thermal bottles
  + Food
  + Movies (e.g. rotten tomatoes, golden village, imdb)
    - e.g. Mulan, Avengers

## Design and develop a program to identify such <noun (phrase) - adjective (phrase)> pairs and rank the pairs from all the review documents.

* Q: how to use program to rank?????
* Maybe Bayes rule?!?!?!

# 3.3 Application (10 marks) KL, sherna

* Sentimental analysis