Natural Language Processing Lab Week #2

Jason S. Chang 張俊盛 jason@nlplab.cc

TAs： Kevin Tuan 段凱文 kevintuan@nlplab.cc

Course Website:

Date:

Outline

- Term Frequencies, Document Frequencies
- Word Counts or Ranks
- This week's assignment
- Dataset
  - British National Corpora
  - Lang-8

Cross-Corpora Analysis

- Compute word counts

- What kind of counts (frequency)?
- **DF (Document Frequencies)** reflects ”a broad range of academic texts”(usage of referring to academic activities, organizing scientific discourse and building the rhetoric of academic texts)
  - can be used to eliminate unimportant words from analysis
  - can also be used to assign weights to boost / scale down the strength of words based on relative importance

Cross-Corpora Analysis

- What kind of counts (frequency)?
- **TF (Term Frequencies)** reflects subject matter (content not rhetoric)
    - how frequently a term occurs in a document

Word Counts or Ranks

- Ranks may be a better measure (Deane 2005)
    - Word frequencies tend to follow highly skewed distributions and have a large number of rare events (Zipf 1935, 1949)
    - Ranks seem like a good alternative to use in the place of word count

Task for this Week

- Count words/phrases (in BNC/clang-8)

- Compute words/phrases ranks (in BNC/clang-8)

- Compute differences of words/phrases ranks (in BNC/clang-8)

- Examine info for words/phrases in a paragraph (optional)

```
word         rank ratio
========= ==========
question
answer
area
active
research
```

Dataset

- **Lang-8 Corpus of Learner English**

This corpus contains English learners texts extracted from Lang-8.

Lang-8 is a free language-exchange social network.

Learner will write a journal in any language, then the native speaker of that language will correct it.

Dataset

- **British National Corpus**

  Contains 100 million words of text from a wide range of genres (e.g. spoken, fiction, magazines, newspapers, and academic).