Readability

Thursday, May 30, 2019 1:19 AM

Details (wanted results denoted as r1, r2, etc.)**:**

1. Import packages, stop word list and common word list.
2. Raw Processing:
   a. Document cleaning -- leave Chinese characters, numbers and punctuations only.
   b. Segmenting sentences according to the following puntuations: **？ ！ 。**
   c. [**Extra Step**] Purifying the sentences -- leave Chinese characters only**.**
      i. **Only** for computing r1 and r2 because accurate word segmentation needs punctuations, etc.
   d. Returning:
      i. Segmented sentences;
      ii. **Cleaned** document size **(r1)**;
      iii. Average sentence length **(r2)**.
3. Generating frequency distribution:
   a. Raw-process the document.
   b. For each sentence, do:
      i. Word segmentation;
      ii. For each word, do:
         1) Check if the word is a stop word
         2) If not, update the freq_dist with the key-value pair: {word: frequency}.
   c. Returning:
      i. Freq_dist;
      ii. Cleaned document size;
      iii. Average sentence length.
4. Determine if a word is a 'complex' word:
   a. Use frequency threshold 1 and 3.
5. Computing the readability:
   a. Grade and Semester:

Grade **(r3)**	年级=17.52547988+0.00242523×课文长度+ 0.04414527×平均句长−18.33435443∗常用字比率
Semester **(r4)**	学期=34.53858379+0.00491625×课文长度+ 0.08996394×平均句长−36.73710603∗常用字比率

   Also returning the percentage of common words **(r5)**.
   b. Fog Value:

Formula	$\mathbf{Fog}(\text{document}) = 0.4\left[\left(\frac{\text{words}}{\text{sentences}}\right) + 100\left(\frac{\text{complex words}}{\text{words}}\right)\right]$ **Modified Fog** = $0.4\,(\,word\_per\_sent + 100 \times \%\,complex)$ **(r5-r8,** *two choices of thresholds and two choices of fog measures***)**
Complex Words Definition	The frequency of the word in the full frequency distribution is less than or equal to a threshold (1 or 3).
% Complex	• The original version: Direct definition: $\left(\frac{\text{complex words}}{\text{words}}\right)$ • The enhanced version from the paper:

$$Percent\ of\ complex\ words\ = 100 \times \frac{\sum_{j=1}^{J} n_j w_j}{\sum_{i=1}^{I} n_i}, \quad \text{(E.2)}$$

$$\text{where } w_j = \frac{\log(\frac{N}{df_j})}{\log(N)}, \quad \text{(E.3)}$$

with N the total number of documents in the population and dfj the number of documents with the word j appearing at least once. The term log(N/dfj) comes from one of the most common term-weighting schemes in the information retrieval literature and is used by Loughran and McDonald (2011) to adjust the relative importance of tonal words. We scale log(N/dfj) by log(N) to make the weight fall in the range [0,1].