Lexical Quantile-Based Text Complexity Measure

Maksim Eremeev and Konstantin Vorontsov (Digital Decisions LLC)

RANLP 2019 Poster Session 3

Motivation

- Build a simple Reading order technique to rank search results
- In the exploratory search, the user needs a hint which of the found documents to read first, gradually moving from simple to more complex documents.
- Reading order optimization is an alternative way to content consumption that departs from the typical ranked lists of documents based on their relevance

Main Idea

- The more specific terms document contains, and the more rare they are, the more complex the document is.
- We estimate the complexity of each term in the document and then aggregate them to get the complete document complexity score.
- We use Wikipedia as a *reference collection* of moderately complex texts in order to determine what term frequencies are abnormal.

Evaluation

- We asked assessors to label 10K pairs of Russian Wikipedia articles.
- Assessors were asked to read both articles and to choose which was more difficult to comprehend.
- Documents were chosen from math, physics, chemistry and programming areas.
- Each pair was labeled twice in order to avoid human factor mistakes.

See our paper to…

- Find another model of single term complexity and its comparison with counter-based one.
- Find experiment results on a single-topic dataset built with topic models.
- Find experiments with different reference collections.
- Find our intentions for the future work.
- Click the link for the beta-version of our academic resource where we are building an exploratory search engine.

General Model

Document length

$$W(d) = \sum_{i=1}^{n_d} w_i [c(t_i) > C_\gamma(t_i)]$$

Term complexity

Document complexity

Term weight

γ- quantile for empirical term complexity distribution in reference collection

Weight examples

$w_i = c(t_i)$ Total complexity

$w_i = \frac{c(t_i)}{n_d}$ Mean complexity

Counter-Based Model

Assumption: The less often the term occurs in the reference collection the more complex it is.

$$c(t_i) = \frac{1}{count(t_i)}$$

Thus, complexity score can be defined as reciprocal to the number of occurrences in the reference collection.

Result model:

$$W(d) = \sum_{i=1}^{n_d} w_i [\frac{1}{count(t_i)} < C_\gamma]$$

Experiment Results

We tested our model on 10K labeled Russian Wikipedia dataset. We used **accuracy** score and compared the model with SOTA readability indexes - ARI and Flesh-Kincaid test.

Flesh-Kincaid test - **57 %**
ARI - **63 %**
Total Complexity Counter-Based model - **74 %**
Mean Complexity Counter-Based model - **81 %**