

How difficult a book is



Streaming Methods to Evaluate Books' Readability Level

Problem Description

- We want to know the readability of a book
 - Easy reading helps learning and enjoyment^[1]
 - Choose the book that fits your reading ability
- Take a book as a stream **S**:
 - Three types of elements in the stream: **word** , .
 - Then length of the stream **S** is s

$$w_i \in \{[A \cdots Z, a \cdots z]^*, \text{comma}, \text{full stop}\}. \quad i \in [1, s]$$


[1] Fry, Edward B. “Readability, Reading Hall of Fame Book, Newark.” DE: International Reading Association (2006).

Achievements

- Good Estimation
 - 50 worldwide classic literature books has been tested
- Space Efficiency
 - Only maintains **84 bytes** as stream goes by!!
 - For an arbitrary book with less than 100 thousand of words
- Time Efficiency
 - Evaluate all 50 books costs **1 second!**

Data Set: Oxford Bookworms Series Books^[2]

These difficulty levels are given by Oxford ELT.



Readability	Books Name
Level 1	Under The Moon, Love or Money, The Coldest Place on Earth, The Monkey's Paw, The Elephant Man, The Phantom of The Opera, The Witches of Pendle, Mary Queen of Scots
Level 2	William Shakespeare, Robinson Crusoe, The Love of A King, Five Children and It, Huckleberry Finn, Alice's Adventures in Wonderland, Anne and Green Gables, Dead Man's Island
Level 3	A Christmas Carol, The Wind in the Willows, The Star Zoo, Tales of Mystery and Imagination, The Secret Garden, The Call of The Wild, Alice's Adventures in Wonderland, Kidnapped, Tooth and Claw, The Bionte Story, Frankenstein, Chemical Secret, The Prisoner of Zinda, The Piciure of Dorian Gray,
Level 4	A Tale of Two Cities, The Hound of The Baskervilles, Gulliver' Travels, The Unquiet Grave, Three Men in A Boat, Little Women, Treasure Island, Dr JEKYLL and Mr Hyde, The Thirty-nine Steps, Silas Marner, Black Beauty
Level 5	Great Expectations, David Copperfield, Far from the Madding Crowd, Wuthering Heights,
Level 6	Oliver Twist, Tess, Jane Eyre, Pride and Prejudice

Evaluation Criteria

- *Vocabulary*

- Huge vocabulary reduce the readability
 - Count distinct words amount $|\{w_i\}|$
- Difficult words reduce the readability
 - Give difficulty to each words in a book

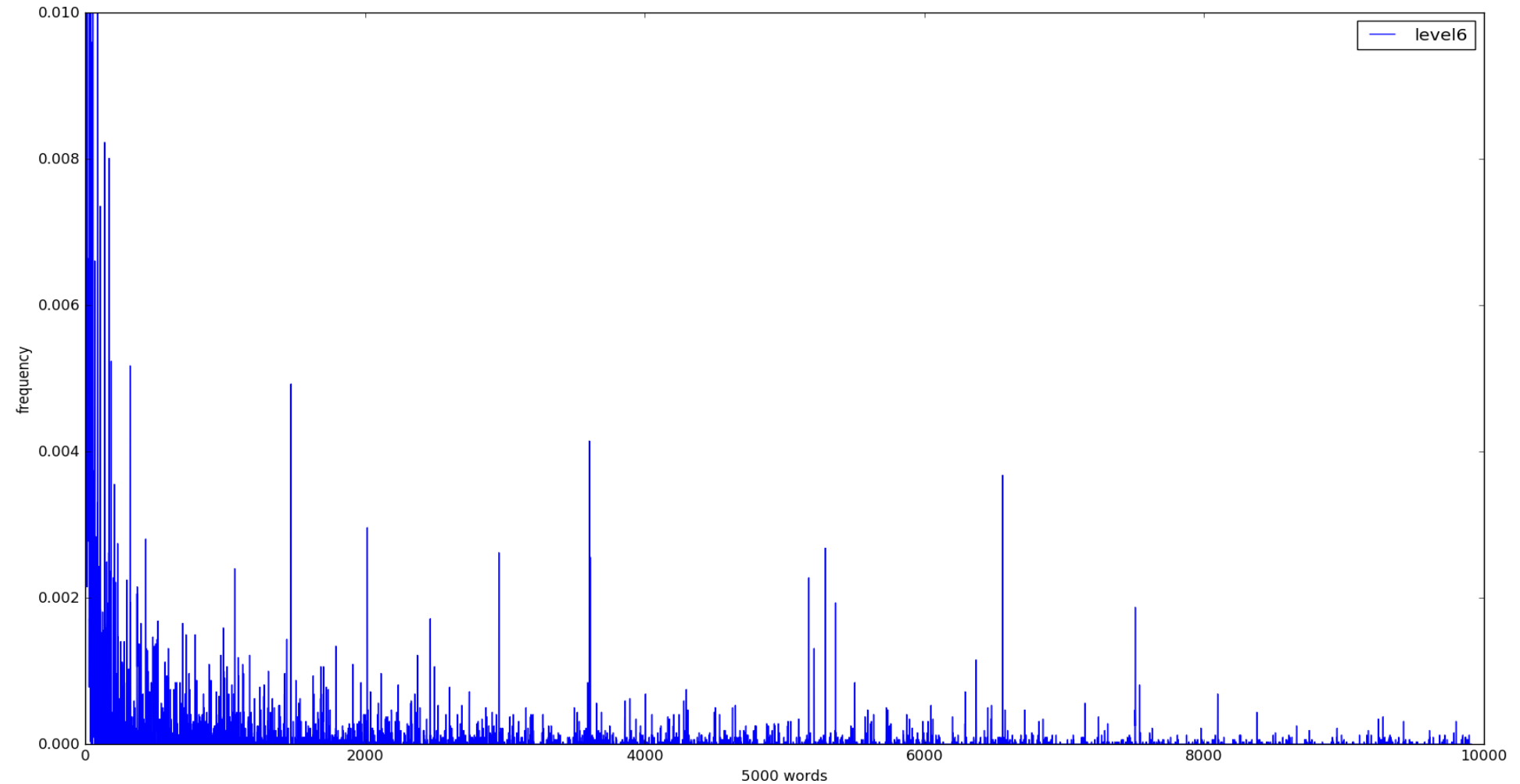
- Grammar

- Long sentences reduce the readability
 - Estimate the average sentence length.

Streaming Algorithm

- We use a value **R** to represent the readability of a book
 - **R** is the weighted sum of following three values
- What do we maintain?
 - Variance: 3 sketches (about 12 bytes)
 - Hyperloglog: 16 sketches (about 64 bytes)
 - Average Length: 2 sketches (about 8 bytes)

Vocabulary-popularity



Vocabulary-popularity

- $Var[x] = \frac{1}{n} \sum (x - E[x])^2$
- $Var[x] = E[x^2] - E^2[x]$
- Time: $O(s)$
- Space: $O(1)$

Vocabulary-amount

- Count the cardinality of multiset $\{w_1, w_2, \dots, w_s\}$
- Min-sketch: $\min\{h(w_1), h(w_2), \dots, h(w_s)\}$
- For n uniformly random variables within $[0, 1]$, the minimum one of them is about $\frac{1}{n+1}$
- i.e. $E[\min_i x_i] = \frac{1}{n+1}$

Vocabulary-amount

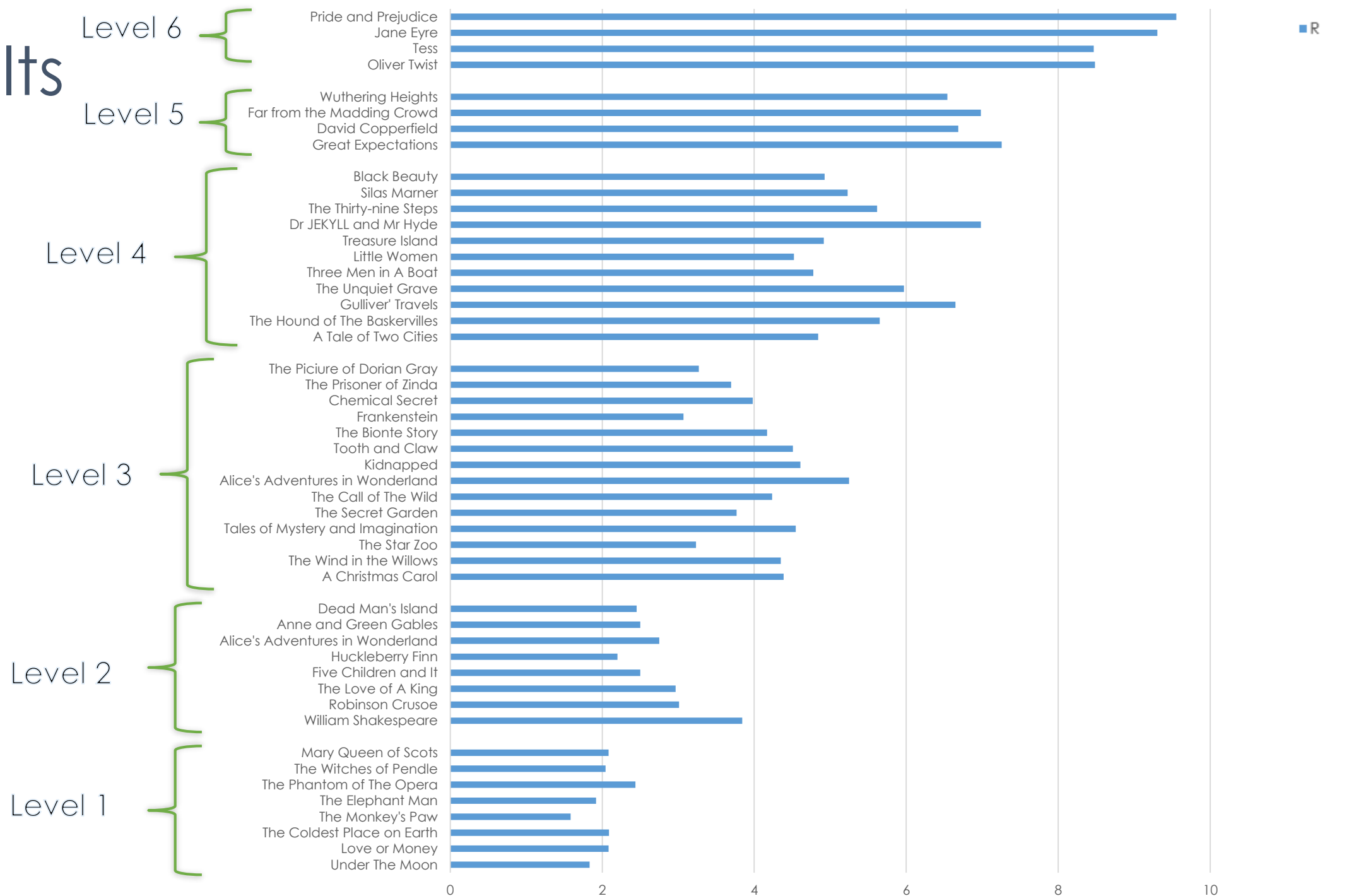
- Hyper Log Log^[4]: improves Min-sketch, using integer random variables instead of real number
- Variance is quite large
- Divide all elements into m groups
- Time: $O(s \log m)$
- Space: $O(m)$
- Accuracy: $\pm 1.04/\sqrt{m}$ (95.7%)

[4] Flajolet, Philippe, et al, “Hyperloglog: the analysis of a near-optimal cardinality estimation algorithm.” DMTCS Proceedings 1 (2008)

Grammar

- How?
- Just calculate the average length of each sentence
- Time: $O(s)$
- Space: $O(1)$

Results



Future works

- Large data set: only hundred thousand words now
- Fix Google's word-list: different word classes

811	apr	6301	catherine
812	written	6302	timely
813	talk	6303	talked
814	federal	6304	upskirts
815	hosting	6305	debug
816	rules	6306	delayed

- Sophisticated analysis on grammar

Summary

- Streaming methods
 - has great efficiency on space and time
 - Not 100% precise, but enough
- Trade-off
 - Streaming method sacrifice little accuracy to gain huge space
- Programmers also should read more books~



Thanks for listening!

- Siyi Yang
- Zhendong Liu
 - Nov 10