Project 1 Report

CIS 678

Matthew Lueder

Overview

- Goal: Compute Flesch Index (FI) for a variety of texts, then visualize and discuss the results.
- The Flesch Index (FI) is a metric used to indicate how difficult it is to read a text written in English.
- An assortment of 30 different texts were analyzed.
  - Includes novels, children's books, newspaper articles, speeches from various politicians, textbooks, government documents, religious texts, and some of my own writing.
  - About 18 MB
- Results are visualized in an ordered table with color representing average grade level required for comprehension.
- Program implemented using Python and PySpark. The use of Spark allows for multiple machines to work on the problem in parallel, greatly reducing the time required to compute the scores.
  - 25 machines in EOS lab at GVSU were used in cluster

Calculating the Flesch Index

1. Text in each document split into words using whitespace as a delimiter.
2. Quotations removed from each word
3. Hyphenated words split into two
4. Words not starting with a character A-Z filtered out
5. Number of sentences counted by looking for periods, question marks, or exclamation points at the end of words
6. Syllables counted (http://eayd.in/?p=232)
   1. Words with length <= 3 have one syllable
   2. Remove last two characters if the word doesn’t end in “ted” or “tes” or “ses” or “ied” or “ies” but ends in “ed”/”es”.
   3. Remove trailing ‘e’ if word doesn’t end in “le”
   4. Look for triplets and pairs of consecutive vowels, add to total syllables for each found
   5. Count remaining vowels
   6. Add a syllable if word starts with “mc”
   7. Add syllable if word ends with “y” that is not surrounded by vowel
   8. Add syllable if “y” is not in the last word, but surrounded by consonants
   9. If word begins with “tri-” or “bi-” followed by a consonant add a syllable
   10. If ends with “-ian”, should be counted as two syllables, except for “-tian” and “-cian”
   11. If begins with “co-” and is followed by a vowel, check if it exists in the double syllable dictionary, if not, check if in single dictionary and act accordingly.
   12. If starts with “pre-” and is followed by a vowel, check if exists in the double syllable dictionary, if not, check if in single dictionary and act accordingly
   13. Check for “-n’t” and cross match with dictionary to add syllable.
   14. Handle exceptional words
7. Count total number of words
8. Use previously determined statistics in Flesch reading-ease equation

$$206.835 - 1.015\left(\frac{total\ words}{total\ sentences}\right) - 84.6\,\frac{total\ syllables}{total\ words}$$

Results 1/2

Title	Author	Flesch Index
The Cat In The Hat	Dr. Seuss	109.2024
Goosebumps Say Cheese And Die	R.L. Stine	88.53809
The Hobbit	J.R.R. Tolken	82.04141
Adventures Of Huckleberry Finn	Mark Twain	79.12457
To Kill A Mockingbird	Harper Lee	78.20287
Trump Victory Speech	Donald Trump	77.16501
Harry Potter And The Sorcerers Stone	J.K. Rowling	76.97009
Quran	Various	76.1321
Gettysburg Address	Abraham Lincoln	74.32404
The Mysterious Affair At Styles	Agatha Christe	72.56798
The Great Gatsby	F. Scott Fitzgerald	71.02288
Hilary Concession Speech	Hilary Clinton	70.06011
King James Bible	Various	69.56787
Kitchen Confidential	Anthony Bourdain	62.99046
Finnegans Wake	James Joyce	62.67159
Obama Farewell Address	Barack Obama	61.91003
The Art Of War	Sun Tzu	61.37606
Moby Dick	Herman Melville	61.18379
Walden	Henry David Thoreau	59.05214
NY Times Article	David Leonhardt	57.24051
Godel Escher Bach	Douglas Hofstadter	56.46773
The Selfish Gene	Richard Dawkins	55.22462
Code Switching in Multilingual Chat	Carmen Loercher	51.715093
Bernie Sanders Speech	Berine Sanders	49.00663
Introduction To Machine Learning	Ethem Alpaydın	47.48751
Selection Of My Writtings For School	Matthew Lueder	35.43157
On Liberty	John Stuart Mill	35.17151
On The Orgin Of Species	Charles Darwin	34.53636
The Critique Of Pure Reason	Immanuel Kant	26.86031
US Constitution	Various	21.64909

Score	School Level	Notes
100.00-90.00	5th grade	Very easy to read. Easily understood by an average 11-year-old student.
90.0–80.0	6th grade	Easy to read. Conversational English for consumers.
80.0–70.0	7th grade	Fairly easy to read.
70.0–60.0	8th & 9th grade	Plain English. Easily understood by 13- to 15-year-old students.
60.0–50.0	10th to 12th grade	Fairly difficult to read.
50.0–30.0	College	Difficult to read.
30.0–0.0	College Graduate	Very difficult to read. Best understood by university graduates.

Results 2/2

Words per sentence

Goosebumps Say Cheese And Die	7.923235	Godel Escher Bach	18.65172
The Cat In The Hat	7.924623	The Art Of War	19.68668
Trump Victory Speech	9.494118	The Selfish Gene	20.26521
Gettysburg Address	12.80952	Kitchen Confidential	20.76428
The Mysterious Affair At Styles	14.01332	Bernie Sanders Speech	21.40625
Harry Potter And The Sorcerers Stone	14.64699	Selection Of My Writtings For School	22.87764
To Kill A Mockingbird	14.80092	Adventures Of Huckleberry Finn	23.49862
Ny Times Article	15.78431	Introduction To Machine Learning	23.65823
Hilary Concession Speech	15.78873	Moby Dick	24.71118
Finnegans Wake	15.86157	King James Bible	26.56723
The Hobbit	16.13456	Walden	29.23049
Quran	17.34207	On The Orgin Of Species	36.1009
Code Switching in Multilingual Chat	18.08438	On Liberty	36.26639
The Great Gatsby	18.19053	The Critique Of Pure Reason	36.75322
Obama Farewell Address	18.3397	Us Constitution	48.5871

Syllables per word

The Cat In The Hat	1.058973	Kitchen Confidential	1.451168
Adventures Of Huckleberry Finn	1.227652	The Art Of War	1.483179
The Hobbit	1.281525	Obama Farewell Address	1.493028
Goosebumps Say Cheese And Die	1.303249	Finnegans Wake	1.513758
King James Bible	1.303799	The Selfish Gene	1.54895
Quran	1.336888	Godel Escher Bach	1.553614
To Kill A Mockingbird	1.342898	Ny Times Article	1.578882
Harry Potter And The Sorcerers Stone	1.359317	On Liberty	1.594008
The Great Gatsby	1.387101	Introduction To Machine Learning	1.599697
Walden	1.396146	On The Orgin Of Species	1.603502
Gettysburg Address	1.412639	Us Constitution	1.606028
Trump Victory Speech	1.418835	Bernie Sanders Speech	1.608759
The Mysterious Affair At Styles	1.418954	Code Switching in Multilingual Chat	1.616599
Moby Dick	1.42517	The Critique Of Pure Reason	1.686409
Hilary Concession Speech	1.427297	Selection Of My Writtings For School	1.751568

Discussion 1/2

- Why did the US Constitution score the lowest in readability?
  - It has the 5th most Syllables per word.
  - However, it has by far the highest amount of words per sentence, averaging 11.84 more words per sentence then the text with the next largest amount of words. This is because the constitution makes extreme use of the semicolon. Most sentences contain lists or many clauses linked together by the semicolon. For example, the following is one sentence from the Constitution:

"The Congress shall have Power To lay and collect Taxes, Duties, Imposts and Excises, to pay the Debts and provide for the common Defence and general Welfare of the United States; but all Duties, Imposts and Excises shall be uniform throughout the United States;To borrow money on the credit of the United States; To regulate Commerce with foreign Nations, and among the several States, and with the Indian Tribes; To establish an uniform Rule of Naturalization, and uniform Laws on the subject of Bankruptcies throughout the United States; To coin Money, regulate the Value thereof, and of foreign Coin, and fix the Standard of Weights and Measures; To provide for the Punishment of counterfeiting the Securities and current Coin of the United States; To establish Post Offices and Post Roads; To promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries; To constitute Tribunals inferior to the supreme Court; To define and punish Piracies and Felonies committed on the high Seas, and Offenses against the Law of Nations; To declare War, grant Letters of Marque and Reprisal, and make Rules concerning Captures on Land and Water; To raise and support Armies, but no Appropriation of Money to that Use shall be for a longer Term than two Years; To provide and maintain a Navy; To make Rules for the Government and Regulation of the land and naval Forces; To provide for calling forth the Militia to execute the Laws of the Union, suppress Insurrections and repel Invasions; To provide for organizing, arming, and disciplining, the Militia, and for governing such Part of them as may be employed in the Service of the United States, reserving to the States respectively, the Appointment of the Officers, and the Authority of training the Militia according to the discipline prescribed by Congress; To exercise exclusive Legislation in all Cases whatsoever, over such District (not exceeding ten Miles square) as may, by Cession of particular States, and the acceptance of Congress, become the Seat of the Government of the United States, and to exercise like Authority over all Places purchased by the Consent of the Legislature of the State in which the Same shall be, for the Erection of Forts, Magazines, Arsenals, dock-Yards, and other needful Buildings; And To make all Laws which shall be necessary and proper for carrying into Execution the foregoing Powers, and all other Powers vested by this Constitution in the Government of the United States, or in any Department or Officer thereof."

Discussion 2/2

- Finnegans Wake has a reputation as one of the most difficult English novels to comprehend. So why does its Flesch Index put it in the middle of the list?
  - Finnegans Wake is difficult because of its unclear non-linear plot, stream of consciousness writing style, made-up words, and multilingual puns.
  - Example sentence: *"The great fall of the offwall entailed at such short notice the pftjschute of Finnegan, erse solid man, that the humptyhillhead of humself prumptly sends an unquiring one well to the west in quest of his tumptytumtoes: and their upturnpikepointandplace is at the knock out in the park where oranges have been laid to rust upon the green since devlinsfirst loved livvy."*
- The Flesch Index's inability to capture the difficulty of Finnegans Wake, and its overrating the difficulty of the US Constitution shows its shortcomings. However, it captured the difficulty of most of the works analyzed fairly accurately. Dr. Seuss and R.L. Stine are rated as the least difficult, while Kant, Darwin, and Mill are rated as some of the most difficult works.