Using Data Science Skills Now: Text Readability Analysis

towardsdatascience.com/using-data-science-skills-now-text-readability-analysis-c4c4641f5875

November 9, 2020

When marketing effectiveness analytics are being developed, the content reading level is often overlooked. Should this be a part of the analysis or your model? If you decide it is, how would you easily tag your different documents or creatives with the current reading level? Today I will review how to use python to do reading level analysis and create a report. We will focus on English as the primary language. This same technique can be used to determine your other content's readability, such as Medium Blogs.

Why readability it important

Whether you want to educate your customers on the benefits of your new service or a medical practice sending home after-care instructions, you need your audience to read and understand the content. Too difficult, and your reader gives up on you. Too easy, and the reader feels that you are talking down to them. Not only do you want your reader to be comfortably engaged with your text, but it is also a key consideration in SEO.

Who is your reader?

It is important to understand and research your typical reader. If you need to focus on one area to start, consider the level of education. Estimate the number of years of education your readers have, then subtract between three and five grade years to estimate the average reading level.

What to do if you aren't entirely sure of your audience's education level? Some industry rules of thumb can help.

Written text that scores in the 4th to 6th-grade level is commonly considered 'easy to read.' If you are trying to explain complex concepts simply for wide audiences, this is your range. Think general blogs and educational material.

Written text scoring Grades 7th to 9th is considered of average difficulty. These readers expect some more complex vocabulary. They expect that they may need to re-read a section of text to understand it fully. How-to articles may fall into this range.

Any text rated 10th grade and above is considered very difficult. This should be reserved for white papers, technical writing, or literature when you are sure that your audience is 'up' for it and/or expects it. Your reader is devoting quite a bit of mental effort to read and absorb your content. Do you know when you read a book that exhausts you? This is that range.

What are the limitations?

Beware of headings, lists, and other 'non-sentences.' These can throw your scores off. Also, depending on which metric you use, the grade level can vary. These scores act as general starting points for analyzing your text. You can identify which content needs further review.

The Code

Many commercial products will scan your content and provide readability metrics. When these products are not available to you or have hundreds or thousands of documents, you will need to create a script to automate the process.

I have a script here that does common automated tasks. It reads in all text documents in a folder and then scores them. For more information on how to scrape texts across many folders and data sources, please see this related article.

Using Data Science Skills Now: Text Scraping

Have a tedious document searching task? Automate it with python in 5 steps.

towardsdatascience.com

There are several packages you can use in your analysis, including textstat and readability. I will use textstat in this example. It is straightforward to use.

```
import textstat  # https://pypi.org/project/textstat/
import os
import glob       # using to read in many files
import docx2txt   # because I'm reading in word docs# where is the folder with your content?
folderPath = '<path to your top folder>\\'# I want to search in all of the folders, so I set
recursive=True
docs=[]
docs = glob.glob(folderPath + '/**/*.txt',recursive=True)
docs.extend(glob.glob(folderPath + '/**/*.docx',recursive=True))
#... keep adding whatever types you needprint(docs)# the language is English by default so
no need to set the language

# Loop through my docs
for doc in docs:
    if os.path.isfile(doc):
        text = docx2txt.process(os.path.join(folderPath,doc))

            print('Document:                    ' + doc)
        print('Flesch Reading Ease:         ' + str(textstat.flesch_reading_ease(text)))
        print('Smog Index:                  ' + str(textstat.smog_index(text)))
        print('Flesch Kincaid Grade:        ' + str(textstat.flesch_kincaid_grade(text)))
        print('Coleman Liau Index:          ' + str(textstat.coleman_liau_index(text)))
        print('Automated Readability Index:  ' + str(textstat.automated_readability_index(text)))
        print('Dale Chall Readability Score: ' + str(textstat.dale_chall_readability_score(text)))
        print('Difficult Words:             ' + str(textstat.difficult_words(text)))
        print('Linsear Write Formula:       ' + str(textstat.linsear_write_formula(text)))
        print('Gunning Fog:                 ' + str(textstat.gunning_fog(text)))
        print('Text Standard:               ' + str(textstat.text_standard(text)))

print('**********************************************************************************')"""Flesc
Kincaid Grade Level = This is a grade formula in that a score of 9.3 means that a ninth grader
would be able to read the document""""""Gunning Fog = This is a grade formula in that a
score of 9.3 means that a ninth grader would be able to read the document."""""SMOG - for
30 sentences or more  =This is a grade formula in that a score of 9.3 means that a ninth
grader would be able to read the document."""""Automated Readability Index = Returns the
ARI (Automated Readability Index) which outputs a number that approximates the grade level
needed to comprehend the text."""""" Coleman Liau Index = Returns the grade level of the
text using the Coleman-Liau Formula."""""Linsear = Returns the grade level using the Linsear
Write Formula."""""" Dale Chall = Different from other tests, since it uses a lookup table of the
most commonly used 3000 English words. Thus it returns the grade level using the New Dale-
Chall Formula."""
```

The results of my four documents. The first is this blog. The second is this blog, with the links removed. The third is a USAToday article, and the fourth is the USAToday article with the header and photos removed.

Document: C:...readability\sampleblog.docx
Flesch Reading Ease: 56.59 (Fairly Difficult)
Smog Index: 13.8
Flesch Kincaid Grade: 11.1
Coleman Liau Index: 11.9
Automated Readability Index: 14.1
Dale Chall Readability Score: 7.71 (average 9th or 10th-grade student)
Difficult Words: 124
Linsear Write Formula: 10.833333333333334
Gunning Fog: 12.86
Text Standard: 10th and 11th grade
****
Document: C:...readability\sampleblognolinks.docx
Flesch Reading Ease: 58.52 (Fairly Difficult)
Smog Index: 12.9
Flesch Kincaid Grade: 10.3
Coleman Liau Index: 10.5
Automated Readability Index: 12.2
Dale Chall Readability Score: 7.48
Difficult Words: 101
Linsear Write Formula: 10.833333333333334
Gunning Fog: 11.95
Text Standard: 10th and 11th grade
****
Document: C:...readability\usatoday article no header photos.docx
Flesch Reading Ease: 21.47 (Very Confusing)
Smog Index: 19.8
Flesch Kincaid Grade: 24.6
Coleman Liau Index: 13.19
Automated Readability Index: 32.2
Dale Chall Readability Score: 9.49
Difficult Words: 317
Linsear Write Formula: 16.25
Gunning Fog: 27.01
Text Standard: 24th and 25th grade
****
Document: C:...\readability\usatoday article.docx
Flesch Reading Ease: 21.47 (Very Confusing)
Smog Index: 19.8
Flesch Kincaid Grade: 24.6
Coleman Liau Index: 13.19
Automated Readability Index: 32.2
Dale Chall Readability Score: 9.49
Difficult Words: 317
Linsear Write Formula: 16.25
Gunning Fog: 27.01
Text Standard: 24th and 25th grade
****

For this blog, the score was influenced by removing links. The USAToday article was NOT affected by removing the header and photos.

References and Resources

- **National Adult Literacy Survey.** The National Adult Literacy Survey (NALS) shows literacy levels for people of different ages, races/ethnicities, and health status.[3] If you know your audience's demographics, you can use data from this survey to estimate the average reading level.
- Editorial: Evidence-based Guidelines for Avoiding Poor Readability in Manuscripts
- Kaggle:

There was an interesting competition in 2019 that I applied readability scores to improve the City of LA job postings. You can see various competitors use a variety of techniques.

City of LA — Readability and Promotion Nudges

Explore and run machine learning code with Kaggle Notebooks | Using data from multiple data sources

www.kaggle.com

Conclusion

Text readability is a very interesting subject. Learning who your audience is and how they consume your content is very important. Ignore this information, and you may not be hitting the target.

Written by

Dawn Moyer

Data Enthusiast, fallible human. A data scientist with a background in both psychology and IT, public speaking in areas of data, career, and ethics.

More from Towards Data Science

A Medium publication sharing concepts, ideas, and codes.

Richard Farnworth

·2 days ago ★

4 key traits to look for when hiring a Data Scientist

Plus sample questions to use in an interview

Finding good Data Scientists can be a tricky task. Googling "Data Science Skills Gap" shows that in many countries around the world, companies are struggling to find suitable candidates for their ever growing Data Science needs.

A bad hiring choice can be very expensive, and as a result, it's important to be able to vet candidates effectively, to make sure they are a good fit for the position and are going to be effective in introducing/expanding Data Science at your company.

Rigour

Understanding the source of the data one has available, along with any Data Quality issues, is imperative to producing high-quality output. When using any dataset, the first thing a Data Scientist should do is perform an in-depth statistical analysis to delve into their raw materials and avoid making any lazy assumptions. ...

Read more · 4 min read

Kenichi Nakanishi

·2 days ago

Classifying Pet-Safe Plants with fast.ai

Training with Controlled Randomness

Creating and comparing fast.ai learners

In Part 1: Building a Database, we've scraped the web for information on plants and how toxic they are to pets, cross-referenced the fields against a second database, then finally downloaded unique images for each class through Google Images. In this part, we

will be training baseline neural networks (using the new fast.ai framework) to identify the species of plant based on a picture. We'll then assess how good the dataset we've put together is for training a neural network, and looking for ways to improve this.

> **The main goals herein will be comparing the effects of changing the number of images per class, and how we can try to compare each training run fairly by controlling randomness.** …

Read more · 12 min read

Julian Herrera

·2 days ago ★

Mastering String Methods in Python

Grow your knowledge of this programming language with this guide with applications of methods to modify, transform and format strings.

Photo by **Christina Morillo** from **Pexels**

> "Learning to write programs stretches your mind, and helps you think better, creates a way of thinking about things that I think is helpful in all domains."
> *— Bill Gates*

As the co-founder of **Microsoft** says, I invite you to continue stretching your mind in an effort to broaden your programming skills with potential applications in many domains. ...

Read more · 8 min read

Read more from Towards Data Science