Privacy Policies

Gabriel Brock

2023-09-17

Privacy Policies

Using at least five different privacy policies from well-known web sites, obtain the text for those policies and look at the length, complexity, and required reading level for those policies (the python package textstat can be of help here, although there are likely others). Write a short paper (2-4 pages) on what you found, along with any patterns that you discovered. Include a section on your methodology.

Introduction

Privacy Policies are notorious for being long and jargon-filled documents that a majority of consumers never read. Websites have embraced a form of cautious legal language to shield themselves from potential lawsuits and penalties. The more ambiguous and flexible their wording, the lower their risk exposure in case of lawsuits stemming from data breaches and unfair trade practices. The more insidious reason for policy opacity lies in a more financial gain for these companies. In the past decade, a new industry known as "data brokerage" has emerged to assist websites in gaining deeper insights into individuals like you and me who browse their platforms. These companies cross-reference and aggregate data to construct highly detailed profiles encompassing purchasing behaviors, political affiliations, sexual orientation, religious beliefs, and medical histories. The gathering and analysis of this data have become a lucrative business, which strongly motivates the collecting firms to make it as challenging as possible for you to opt out of their data collection practices.

Virtually every consumer knows that privacy policies are convoluted word jungles but how difficult is it for the average American to get through a privacy policy for one of the countries most trafficked websites? We set out to measure the length, complexity, and required reading level for these policies.

Methods

Readability Indexes Used

Flesch-Kincaid Grade Level Index

The Flesch–Kincaid readability tests are readability tests designed to indicate how difficult a passage in English is to understand. The "Flesch–Kincaid Grade Level Formula" presents a score as a U.S. grade level. It can also mean the number of years of education generally required to understand this text, relevant when the formula results in a number greater than 10.

The grade level is calculated with the following formula:

$$FK = 0.39 \times \frac{W}{St} + \times \frac{Sy}{W} - 15.59$$

The different weighting factors for words per sentence and syllables per word in each scoring system mean that the two schemes are not directly comparable and cannot be converted. The grade level formula emphasizes sentence length over word length.

Limitations

As readability formulas were developed for school books, they demonstrate weaknesses compared to directly testing usability with typical readers. They neglect between-reader differences and effects of content, layout and retrieval aids

FORCAST Formula

Unlike most other formulas, the FORCAST formula uses only a vocabulary element, making it useful for texts without complete sentences. FORCAST is a formula focused on functional literacy. UNESCO defines functional literacy as:

> The ability to identify, understand, interpret, create, communicate and compute.'

The FORCAST result is a US education grade. The grade level is calculated with the following formula:

$$FORCAST = 20 - \frac{W^{1Sy} \times \frac{150}{W}}{10}$$

Linsear Write Index

Linsear Write is a readability metric for English text, purportedly developed for the United States Air Force to help them calculate the readability of their technical manuals. Linsear Write encourages writers to use stronger verbs and avoid passive phrases. It is one of many such readability metrics, but is specifically designed to calculate the United States grade level of a text sample based on sentence length and the number of words used that have three or more syllables.

The grade level is calculated with the following formula:

$$LW_{raw} = \frac{100 - \frac{100 \times W_{<3Sy}}{W} + (3 \times \frac{100 W_{3Sy}}{W})}{\frac{10 \times St}{W}}$$

$$LW(LW_{raw} \leq 20) = \frac{LW_{raw} - 2}{2}$$

$$LW(LW_{raw} > 20) = \frac{LW_{raw}}{2}$$

Simple Measure of Gobbledygook (SMOG)

Simple Measure of Gobbledygook (SMOG) is a readability framework that measures how many years of education the average person needs to have to understand a text. It is best for texts of 30 sentences or more. The SMOG result is a US education grade.

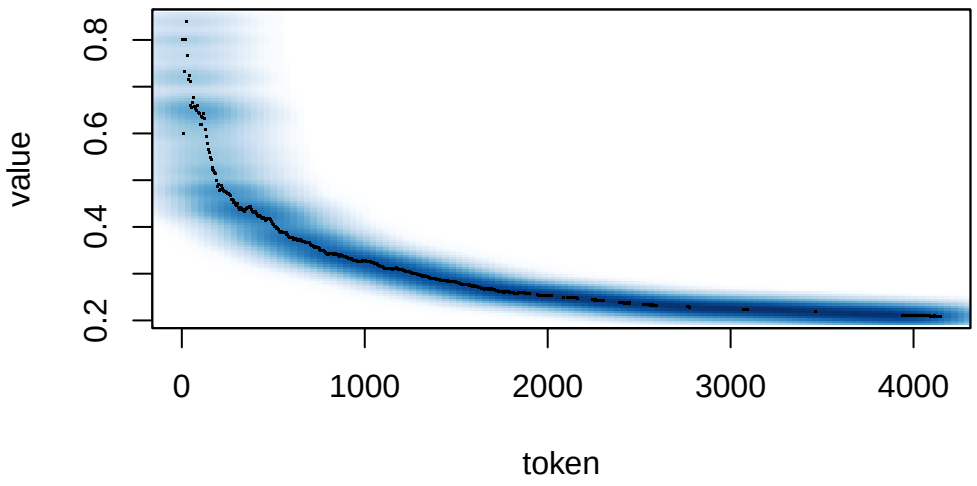
The grade level is calculated with the following formula:

$$SMOG = 1.043 \times \sqrt{W_{3Sy} \times \frac{30}{St}} + 3.1291$$

Results

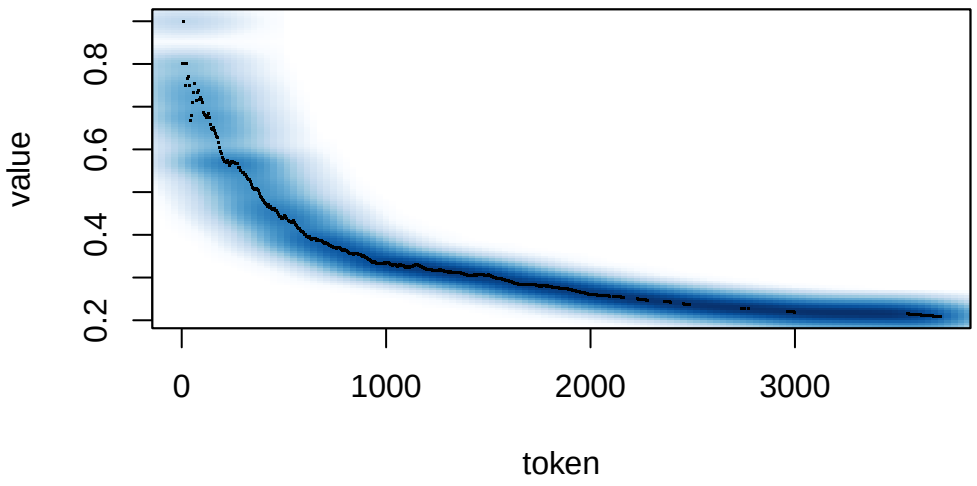
Apple

```
# Plot the classic type-token ratio with characteristics
apple_viz <- smoothScatter(apple.ttr@TTR.char, nrpoints = 500)
```



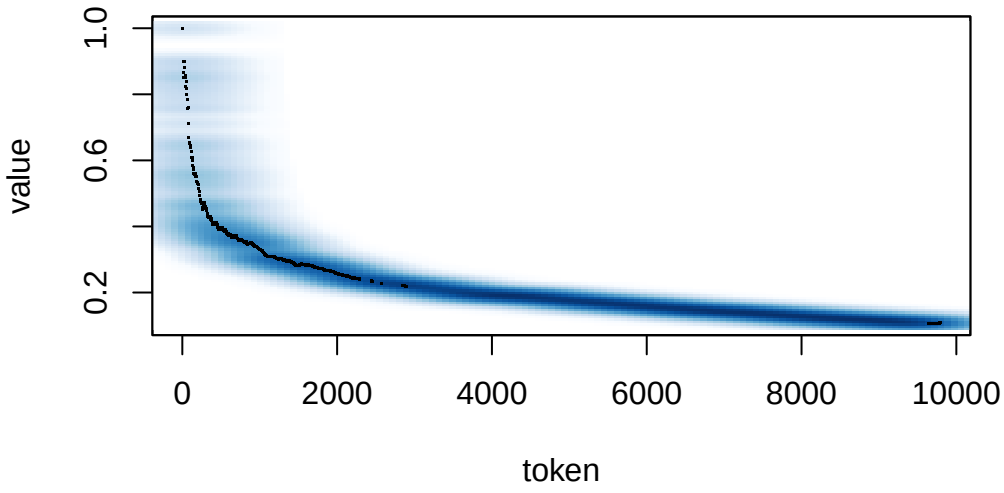
Amazon

```
# Plot the classic type-token ratio with characteristics
amazon_viz <- smoothScatter(amazon.ttr@TTR.char, nrpoints = 500)
```



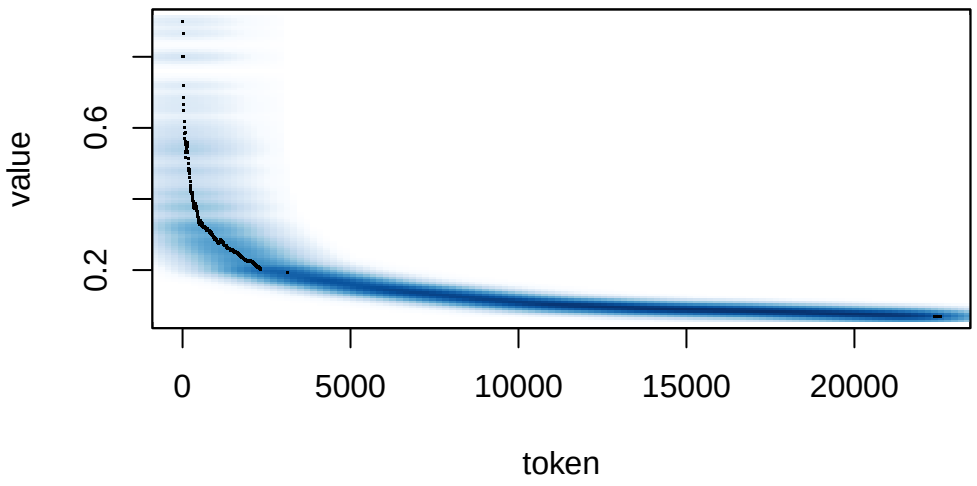
Fox News

```
# Plot the classic type-token ratio with characteristics
fox_viz <- smoothScatter(fox.ttr@TTR.char, nrpoints = 500)
```



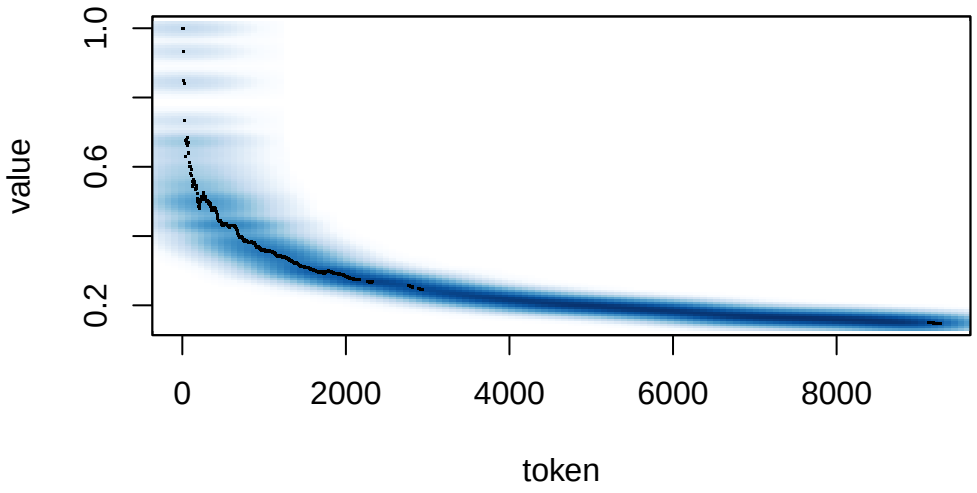
Meta

```
# Plot the classic type-token ratio with characteristics
meta_viz <- smoothScatter(meta.ttr@TTR.char, nrpoints = 500)
```



New York Times

```
# Plot the classic type-token ratio with characteristics
nyt_viz <- smoothScatter(nyt.ttr@TTR.char, nrpoints = 500)
```



Results and Discussion

Company	WordCount	SentcCount	AvgSentc	AvgWord
Apple	3710	260	14.27	5.28
Amazon	4145	202	20.52	5.07
Fox	9793	566	17.30	5.03
Meta	22552	1267	17.80	4.95
NYT	9265	702	13.20	5.02

None of the privacy policies clocked in under 3000 words, so it would take at least 15 minutes of concentrated reading to get through the shortest of the policies, Apple, and at least 75 minutes to finish reading Meta's policy.

The sheer length of theses policies further substantiates the argument that policies are intentionally long but the another notable aspect lies in the required reading level for these policies. Most Americans read around an eight grade reading level so it is quite alarming that each of the policies had required reading levels above this benchmark. This is not simply due to unfamiliarity with legal jargon given that some of the readability indexes simply utilize token count in their estimates.

Company	FKGrade	FKAge	ForGrade	ForAge	LWGrade	LWEasy	LWHard	SMOGGrade	SMOGAge
Apple	12.7	17.70	11.70	16.70	14.26	80.51	19.49	14.55	19.55
Amazon	10.04	15.04	11.61	16.61	8.70	81.99	18.01	12.29	17.29
Fox	11.13	16.13	11	16	12.21	79.43	20.57	13.91	18.91
Meta	10.95	15.95	11.31	16.31	11.95	82.88	17.12	13.10	18.10
NYT	9.05	14.05	10.55	15.55	7.89	82.61	17.39	11.78	16.78

There was an alarmingly high trend of required education to read these documents with all policies requiring a seemingly college-bound high schooler's level of comprehension.

Required comprehension level notwithstanding there was an interesting trend of complexity among the policies. Lexical complexity as measured by Token-Type Ratio had a very drastic downward trend in all five policies. is TTR is the ratio obtained by dividing the types (the total number of different words) occurring in a text or utterance by its tokens (the total number of words). A high TTR indicates a high degree of lexical variation while a low TTR indicates the opposite. The range falls between a theoretical 0 (infinite repetition of a single type) and 1 (the complete non-repetition found in a concordance). As the above tables illustrate, as the privacy document progressed the TTR steadily declined meaning the lexical variety usually declined to around 0.2 or 20% by the end of the document. This monotony in the document also likely contributes to the unwillingness of consumers to read privacy policies.

The construction of privacy policies by companies like Apple, Amazon, Meta, Fox News, and the New York Times are likely highly calculated efforts in opacity and monotony not just to absolve the organizations from legal responsibility but to also obscure there privacy practices from most consumers through comprehension and complexity.

Methodology

Data Collection

I Google searched for the privacy policies of Amazon, Apple, Fox News, Meta, and the New York Times. I then turned each of the .html policy pages into .pdf documents and converted those documents into .txt documents from import into RStudio.

Calculations

I utilized the R package `koRpus` to calculate several metrics of text analysis including automatic language detection, hyphenation, several indices of lexical diversity (e.g., Flesch-Kincaid, Linsear Write).

I was able to use the `describe()` function to generate a list of basic summary statistics including total word count, total sentence count, average word length, and average sentence length.

I used the `readability()` function in the **`koRpus`** package to generate a series of readability index measurements including the Flesch-Kincaid Grade Level Index, FORCAST Formula, Linsear Write Index, and the Simple Measure of Gobbledygook (SMOG) all of which produced an calculated grade level recommendation.

I used the `smoothScatter()` function produced a smoothed color density representation of a scatterplot of the classic type-token ratio with characteristics for each of the privacy policies.