TEXT ANALYTICS and LINGUISTICS

Finding linguistic pattern for the best writing practice using text mining and NLP in SAS environment

ROHIT BANERJEE, MS in Business Analytics, OKLAHOMA STATE UNIVERSITY

ABSTRACT

Text Analytics is the process of examining large collections of written resources to generate new information; transforming unstructured text into structured data helps us find meaningful insights from the text. It is a subgroup of Natural Language Processing (NLP). Statistical methods, rule-based modeling, and machine learning techniques are applied in text analytics allowing for the extraction of topics, keywords, semantics, and sentiments from the raw text in an effort to categorize terms.

INTRODUCTION:

The effectiveness of a technical paper is marked by the extent of its readability. This paper illustrates the derivation and interpretation of the Automated Readability Index (ARI), the Coleman–Liau Index and the lexical density. We acknowledge that a constant change is happening the way we converse, the use of languages and phrases, simultaneously we also find it difficult to comprehend written materials irrespective of the maturity of the writer. This paper examines the abstracts of the peer-reviewed journal to measure these mathematical parameters and find a likeness among them to render the idea of an exemplary writing.

DATA

My corpus includes abstracts from 13 marketing papers published in International Journal of Research in Marketing. The journals are chosen because peer-reviewed publications are scrutinized at various stages of readings and corrections to elucidate the subject matter to its readers.

Readability statistics helps us to predict the reading difficulty of a given text or texts. Common readability statistics include the following:

Total character count.
Total word count
Total sentence count
Total paragraph count
Average word length
Average sentence length
Average paragraph length

Distinct word count.

FORMULA:

To provide a better perspective of the readability statistics and how these work the formulas are provided as below,

Automated Readability Index

- ARI = (4.71 x Characters/Words) + (0.5 x Words/Sentence) – 21.43

  Character/Words = Average length of words

  Words/Sentence = Average sentence length

The Coleman-Liau Grade Level score is calculated as follows

- CLGL = (5.89 x Average word length) - (30 x (Number of sentences/ Number of words)) – 15.8

The ARI test is developed and used by the US Army to understand technical documents, whereas the Coleman-Liau grading is more suitable for 4th grade to college level texts. As evident from the previous sentence, these 2 tests have been selected to examine both perspective of an article. The ARI, along with the Coleman-Liau, relies on a factor of characters per word, instead of the usual syllables per word. The number of characters is more readily and accurately counted by computer programs than syllables.

Lexical Density

Lexical Density measures how much information there is in a text. The proportion of content words to function words in a text. The higher the proportion of content words, the greater the lexical density.

Ld = (Nlex / N) x 100

APPROACH

The abstracts are extracted from the journals, and a separate set of text files are prepared. The files are in the .txt format and named as “Obs1”,”Obs2” etc. Using SAS ‘infile statement’ all the files are obtained in the SAS environment. A SAS macro is written that calculated the above mentioned variables and the result is maintained in a SAS table called “finalstats.sas7Bdat”.

To count the number of sentences the number of periods (.) are counted. The punctuation marks like (‘.’ ‘?’ ‘!’) are considered as stop words since these marks the end of the sentence. Counting the number of stop-words are helpful in counting the number of sentence. However, in any level of technical paper there is high usage of Latin abbreviations, considering this factor, a separate table is prepared with the abbreviations that consist of periods such as “e.g.”, “i.e.” to eliminate the confusion of stop-words

```
proc sql;
/*can do total words /total sentences to give avg wds/sentence*/
create table sentencev1 as
select
upcase(compress(word,'"'))as word1
from scsug.paper
where calculated word1 contains '?' or calculated word1 contains '.' or calculated word1 contains '!';
quit;

proc sql;
/*can do total words /total sentences to give avg wds/sentence*/
create table sentencev2 as
select *
from sentencev1
where word1 not LIKE '%E.G.' and word1 not LIKE '%I.E.' and word1 not LIKE '%ET AL.' and word1 not LIKE '%CF.' and word1 not LIKE '%IBID.' and
word1 not LIKE '%Q.V.' and word1 not LIKE '%VIZ.';
quit;
```

```
    data paperstat&i;
    obs = &i;
    avg_wd_len= &total_chars/ &total_words;
    avg_sentence_len=&total_words / &total_sentences;
    ARI=4.71*avg_wd_len + 0.5*avg_sentence_len - 21.43;
    CLGL=(5.89*avg_wd_len) - (30*(&total_sentences/&total_words)) - 15.8;
    pct_distinct=&total_distinct./&total_words;


    proc append base  = scsug.finalstats data = paperstat&i;
    run;

%end;
%mend textanalytics;

%textanalytics(13)
```

Figure1: Partial representation of SAS macro

Each readability index gives an estimated grade level required to be able to read and comprehend a text without difficulty. In other words, the grade level can be considered as the number of years one has undergone a formal education. Thus, the lower the index, the easier the text is to read, and conversely, the higher the index, the more difficult the text is to read. In the United States, 12th grade is the compulsory education level, so index value above that is considered an adult level. In that way, technical literature published in scholarly journals should score at least 12. However, a good article does not need to be complex. Sometimes, lower text complexity is an indicator of lucid writing whereas higher value may demonstrate ambiguity.

The readability index is a lower bound of the level to understand the text. If a text is rated as 12, it simply states that to comprehend the text the reader should have a minimal 12 year of formal education or above.

Apart from these, it must be acknowledged that readability index can only account for a part of several other parts such as the interest of the reader, motivation or the reading skill. Keeping that in mind, Lexical Density is also calculated for the abstracts. Lexical density describes the proportion of content words (nouns, verbs, adjectives, and also adverbs) to the total number of words. Exploring this parameter we can unfold that a text with a high proportion of content words contains more information than a text with a high proportion of function words. It will be easier to infer the coherence of a written text.

RESULTS

obs	avg_wd_len	avg_sentence_len	ARI	CLGL	pct_distinct
1	5.4375	26	17.180625	15.073028846	0.5480769231
2	5.6266666667	18.75	14.4466	15.741066667	0.6733333333
3	5.1111111111	25.875	15.580833333	13.145024155	0.5169082126
4	5.8109452736	25.125	18.502052239	17.232437811	0.6069651741
5	5.397515528	20.125	14.054798137	14.50068323	0.6211180124
6	6.2564102564	19.5	17.787692308	19.511794872	0.5961538462
7	5.961722488	26.125	19.712212919	18.166220096	0.5598086124
8	6.0625	20	17.124375	18.408125	0.5875
9	6.472392638	20.375	19.242469325	20.85	0.6257668712
10	5.4716157205	28.625	18.653810044	15.379781659	0.5720524017
11	6.8278688525	15.25	18.354262295	22.448934426	0.6803278689
12	5.3888888889	27	17.451666667	14.829444444	0.5555555556
13	5.555	25	17.23405	15.71895	0.66

Table1: SAS output

13 documents are mentioned under the "Obs" variable. The average word length in the abstracts is 5. The 'pct_distinct' variable provides the percentage of unique words used in an abstract. The average number of distinct words in these documents is 0.60. Unique words in document make the text more complete, also it increases the reader's interest towards the text. As far the statistics go, we see that the average count of words in a sentence is around 20.

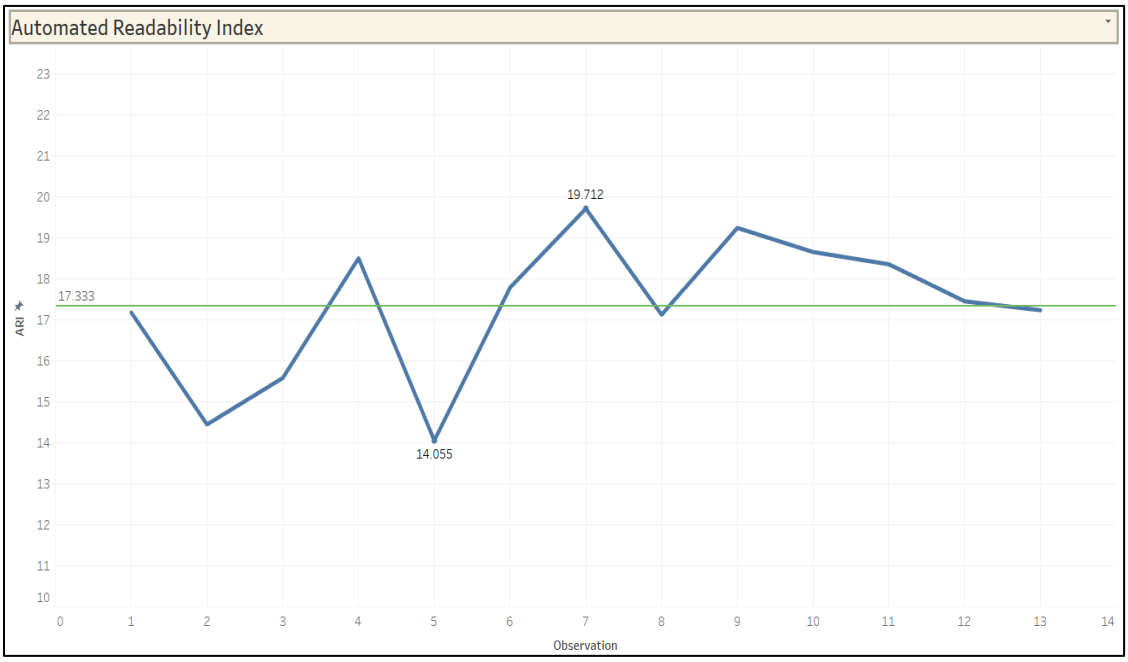


Figure2: Automated Readability Index with range and average value

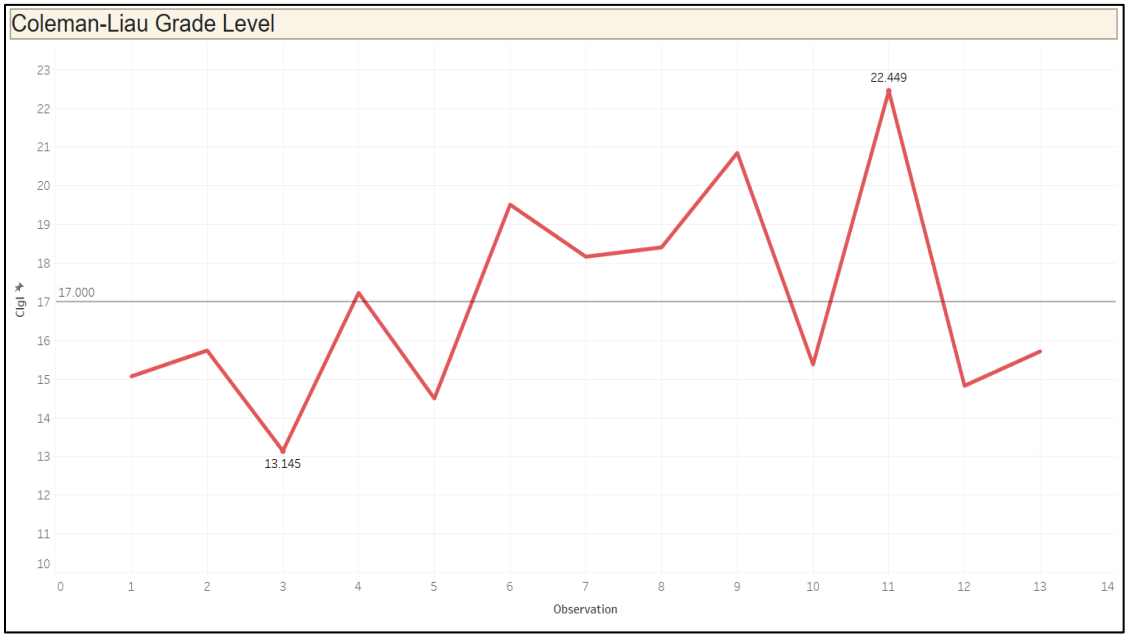
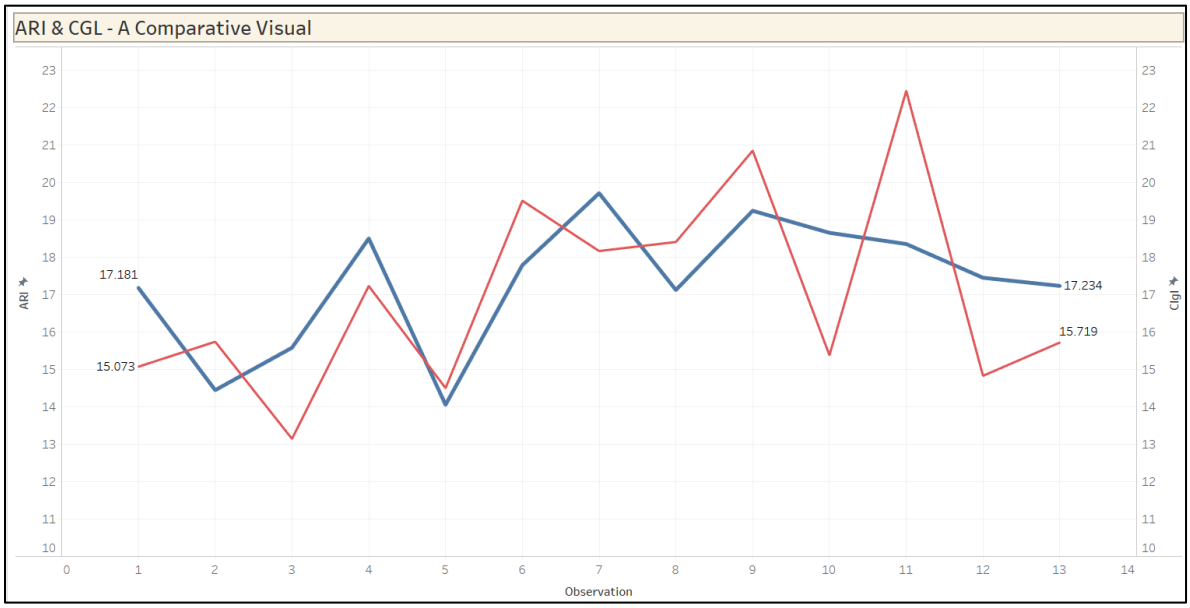


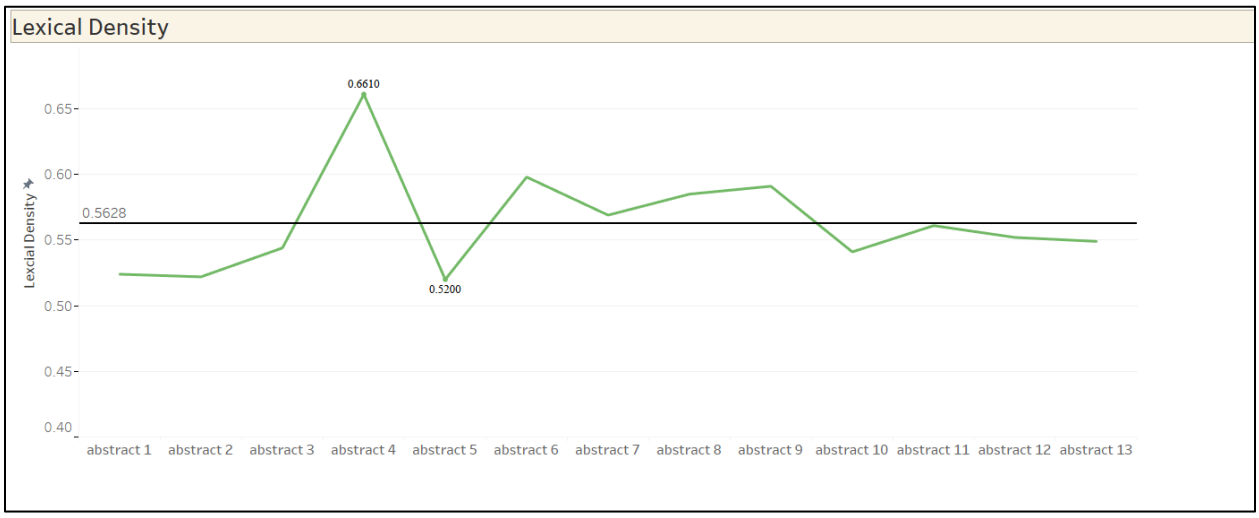
Figure3: Colemane-Liau Grade Level with range and average value

In the above figures (figure2 and figure3) it is observed that the index values indicate an average value of 17 with a range from 14 to 19(figure2) and 13 to 22 (figure3) respectively. As proposed earlier this statistics imply early undergraduate to graduate level with the average is around the first-year graduate. In other words, a college freshman should not face difficulty in reading as well as understanding this document.



Figur4: A comparative image of both of the index described in the paper

Both ARI and CGL is developed on the basis of the words instead of syllables, while ARI measures the technical articles, CGL is developed to quantify more general writings. In the above figure, a comparison is made between the two process, and it is observed both of the graphs followed an almost overlapping path.



Figur5: Lexical density with range and average value

The ARI and CGL values provided earlier, tells that the quality of the writing matches with the expectation. However, a significant part of readability index cannot be deciphered with mere calculation; other factors such as competency of the reader, grasp on the topic, motivation to read an article of such measure are also need to be considered. In the above figure (figure5) the lexical density values of the texts are represented to nullify those factors and bolster the findings inferred from the readability index.

The average lexical density of the abstracts is 0.56, with a range from 0.66 to 0.52 respectively. That indicates if a scholar wants to write a precise and comprehensible piece the lexical density should preferably above 52%.

ASSUMPTION

A few assumptions have been made to prepare the SAS code and the above proposal,

a) The written piece is in the English language.
b) Proper grammatical structure and punctuations are used in an orderly manner.
c) Above description of readability parameters or lexical density does not encourage the writer to follow the code in a mechanical way, rather it is to provide a frame of reference.
d) The documents will not contain Greek numeric.

APPLICATION

The readability statistics and lexical density are important parameters of any type of article. While its research value is obvious, we can leverage it in our schooling system to understand a student's capability and style. That not only help the student to be methodical but also help the teachers. It can be utilized in psychological analysis and also can provide a framework to researchers who write tirelessly for their first publication.

REFERENCE:

Text Analytics: the convergence of Big Data and Artificial Intelligence - Antonio Moreno, Teófilo Redondo

Lexical diversity and lexical density in speech and writing: a developmental perspective - Victoria Johansson

Dickens vs. Hemingway: Text Analysis and Readability Statistics in Base SAS® Jessica Hampton

Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) For Navy Enlisted Personnel - J. Peter Kincaid