Prediction of Legitimate Scientific Writing, and Imitation of Science Via Programmatic Characterization of Written Language.

Journal of Computational Linguistics

http://cljournal.org/submissions.html#papers

Introduction:

Non-scientific writing typically exceeds genuine scientific writing in one important criteria: in contrast to genuine science non-science ideas are often expressed using a lower complexity and more engaging writing style. However, multiple stake holders could benefit if science was more often communicated using a lower complexity expression of ideas. Knowledge would be more readily mobilized and transferred into public awareness, also, and machine readability and machine organization of factual information derived from journal articles should also occur more readily.

We believe non-science writing occupies a style niche, that academic science writing should also occupy. We show that we can blindly predict the status of writing: popular culture writing, opinionative writing, and traditional science, by using machine learning to classify the different writing types. By predicting which of the several different writing types any writing piece occupies, we are able to characterize among different writing niches.

Objectively describing the different character of the different writing styles will allow us to prescribe how, to shift academic science writing into a more accessible niche, where science can more aggressively compete with pseudo-science, and blogs, facilitating greater knowledge transference, at a moment in history when public awareness is critically at stake.

Top of Form

Bottom of Form

Methods:

We built a web-scraping and written text analysis infrastructure by extending many existing Free and Open Source (FOS) tools. The Web scrapping interface employs several common python modules, chief among those was: Google Scrape, Beautiful Soup and Selenium. The Text analysis infrastructure was based on the two substantial code bases Text-stat, which contained measures of text reading level (complexity), and NLTK (the Natural Language Processing Tool Kit), which contained measures of text subjectivity, and sentiment type.

The scraping, and analysis work we performed, rested on top of a large hierarchy of software dependencies. However, it is increasingly well understood, that, dependency heavy software stacks act as a significant impediment to investigating or reproducing any product of digital, scholary research. In order to address this problem and to enhance the reproducibility of our approach, we created the necessary web-scrapping, and analysis infrastructure inside a dedicated Docker Container.

Reproducibility is burdened by the technical task of satisfying each software dependencies necessary to recreate a digital scientific experiment. Our position, of starting with from a cloned software environment, will mitigate, the burden of duplicating our digital research environment. That is why we have used a Docker file, and associated Docker container together, as they act as self-documenting and extremely portable software environment clone.

Initially we created two different, unrelated and broad ranging lists of scientific queries. The first type of query was predominantly cultural in nature, or world view related. The second set of queries represented gains in knowledge about physical entities or physical processes in the world. We were interested in scientific, and pseudo-scientific writing.

There were two types of writing that we actively excluded from our analysis, those were websites expressed in a non-English language, and also websites, that were highly commercial in nature. These were websites advertisements, of consumer goods, and online shopping generally. These websites, utilizing wording, that significantly biases text stat metrics. Webpages of less than 400 words, were most often advertisements, and websites

Non English, websites were excluded for the simple reason that they are not amenable to Textstat, and NLTK tools, however, even if this was not the case, it is also known, that there are significant differences in per-word information entropy between different natural languages.

### Search Engine Queries:

The first two lists of queries were chosen to be belong to an overt set of exclusively scientific or cultural search terms. A third list of terms was designed to be deliberately ambiguous.

**Science Queries:**

The three lists of search engine queries were as implemented as follows: science engine queries: 'evolution', 'photo-sysnthesis' ,'Transgenic', 'GMO', 'climate change', 'cancer', 'Vaccines', 'Genetically Modified Organism', ‘differential equation’,"psycho-physics","soma”

**Cultural queries** were as follows: 'reality TV', 'prancercise philosophy', 'play dough delicious deserts', 'unicorn versus brumby', 'football soccer' , god fearing.

As discussed we also designed ambiguous queries which were equally likely return content that was either scientific, or non scientific in nature. The reason for doing this was to provide a challenge for the classification algorithm.

**Ambiguous Queries**, were as follows:

"the singularity","skynet","","killer robot","franken-science","Frankenstein,”the God Delusion","god does not play dice", "the selfish gene","political science", ", "requim for a spike"

After scraping across the two different lists were performed, the resulting queries were filtered, according to specific sets of criteria. As stated previously, we discarded from our analysis, web pages that were not written in English, since we did not have the necessary tools, to analyze them.

## Text Metrics:

A list of metrics applied to downloaded corpus include: TextStat, which was used to measure word complexity (an average of several important word complexity metrics, such as the Gunning Fog measure of reading level), LZW compression-ratio, de-compression ratio, sentiment analysis, subjectivity analysis, and page rank.

Compression ratios were used to investigate the notion, that well written scientific writing, might simply be lower in information entropy, and an information theoretic analysis, can be used to both to better characterize, and corroborate the findings of other reading word complexity metrics

## Reference Texts:

Some reference texts were used, as a means of providing contrast, and context, to data points, among our web scraping derived corpus. The *Upgoer5* is a library, of scientific texts, written with the aid of a text editor, which imposes, that output documents are exclusively comprised by, only the 10,000 most commonly occurring English words.

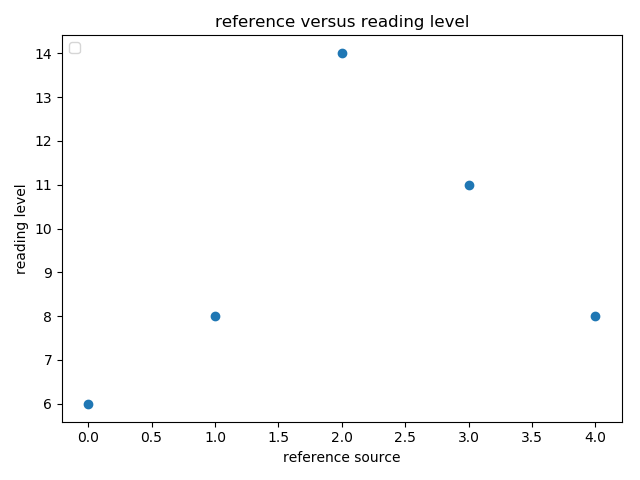
The *Post Modern Essay Generator (PMEG)*, embodies an artificial English synthesis technique. Documents that are output from the website, consist of sentences that obey the rules of written English, however there are no restraints governing the semantic conceptual references in the sentences. If any particular sentence in a PMEG document, embodies an objective meaning, it is only by chance. Output from PMEG reads as highly coded, and vague.

The reference data points in some ways provide further validation to the existing text metric tools, as we needed to verify that word readability metrics provided results that were consistant with prior assumptions about known texts. For instance, the corpus derived from upgoer editor, should require a very low reading grade level to understand. Texts, from the PMEG should require a very high reading level to understand, and cumulative entropy of such texts should be high.

Results:

# Reference Points:

Results, relative to reference points.



Elsewhere Post Modern Essay Generator

Elsewhere Post Modern Essay Generator



We created a total list all of the different queries obtained, from combining both the list of cultural queries, and scientific queries, and then applied such queries exclusively to the Wikipedia, search interface. We take the result of evaluating this pool of queries, and then plot the resulting pool of queries versus page rank. The Wikipedia, actually showed a small but consistant preference for web pages of higher complexity.

The plots below may appear to look a bit unprofessional, as not all data points have error bars. The reason for this, is pandas+seaborne allows you to plot on the same axis multiple sample data points, and single sampled data points. Only of the five mentioned search engines was made to sample beyond page rank 10, but all five sampled under page rank 10.

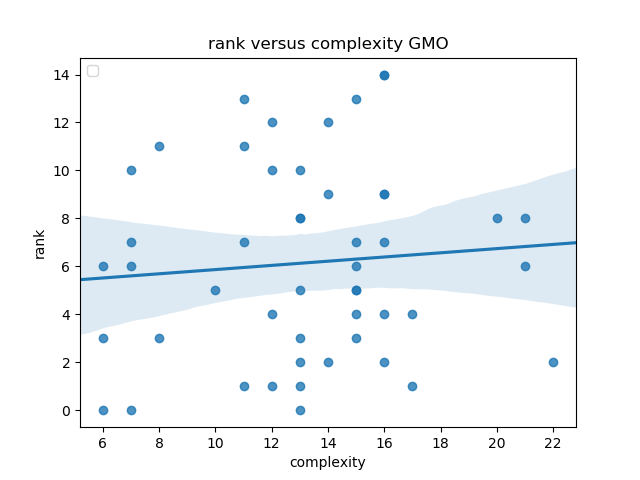
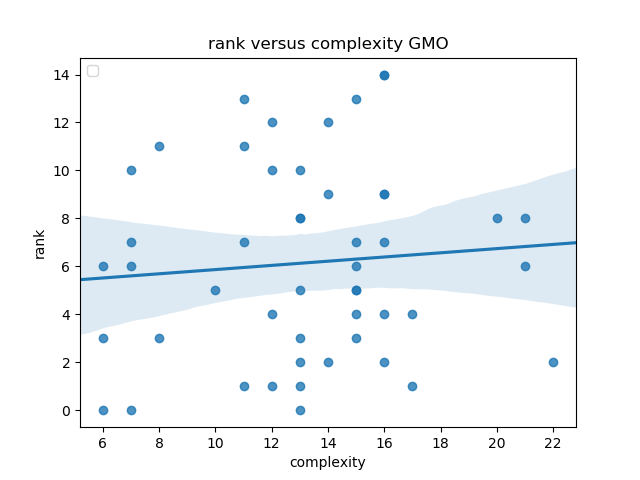


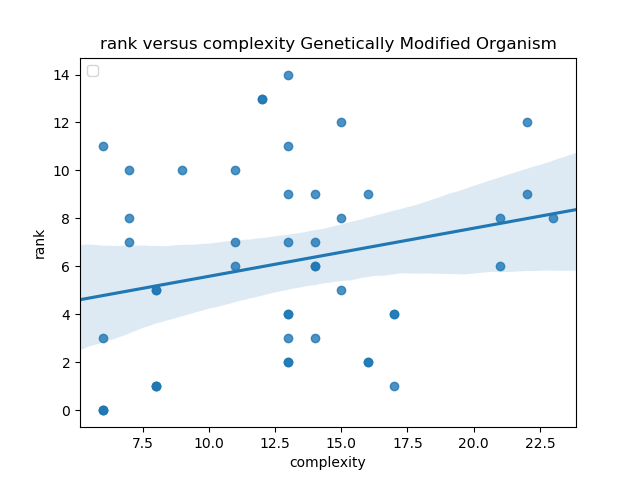
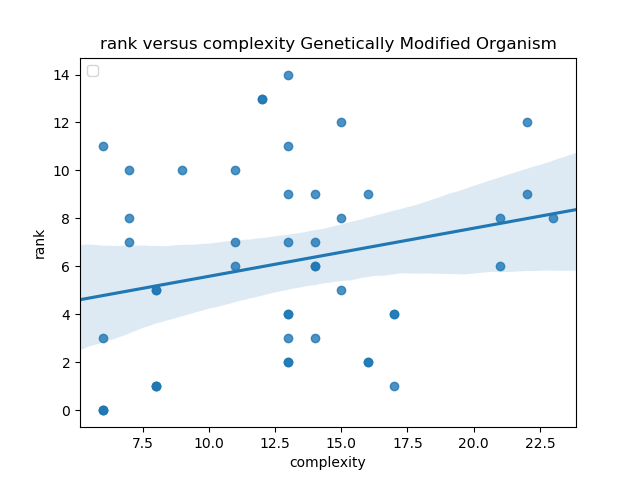
The same plot but with mean and std deviation plots when multiple samples per page rank are available. When we plotted the Wikipedia queries, where reading (text-stat standard) level is instead plotted against page rank, we again see that there is a slight trend towards increasing text complexity with decreasing page rank.



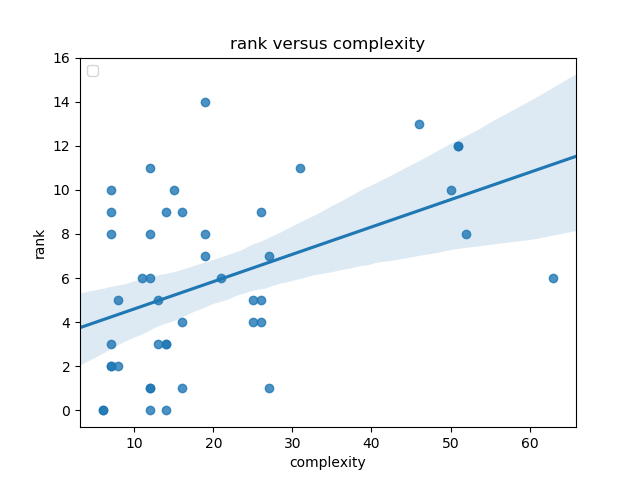
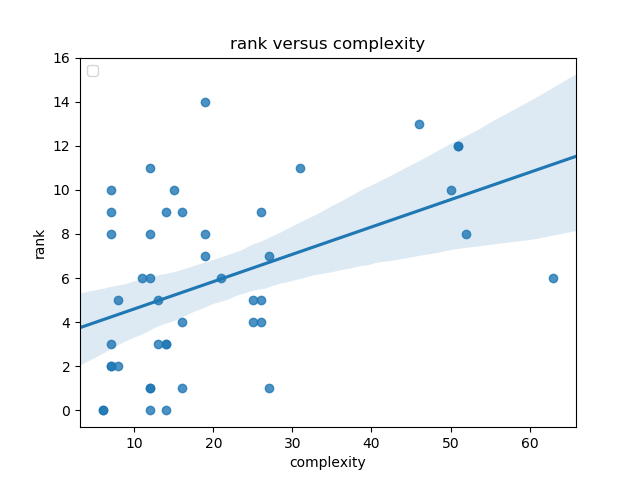
If instead we consider how resistant Wikipedia pages are to compression, we see that low page rank pages, are more resistant to compression. This finding recapitulates the same result as the above figure, where increasing Wikipedia page rank slightly decreases text complexity.

When agregating search results between all possible search engines, and then plotting page rank versus complexity, for particular types of search terms, there was a strong positive correlation between page rank, and reading level.

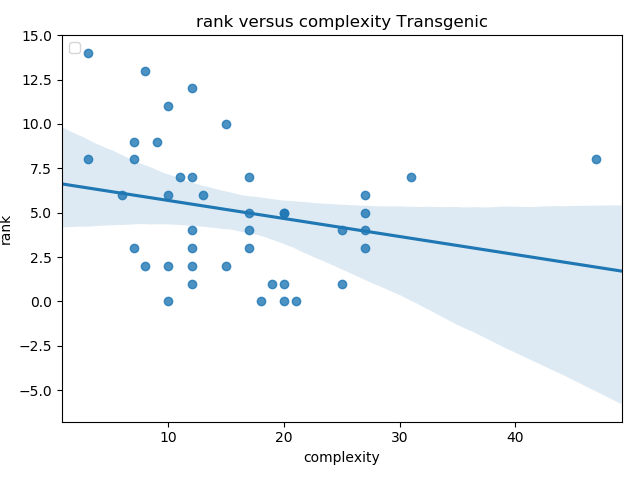
In the case of GMO, and Genetically Modified Organism, positive increasing trends where observed 

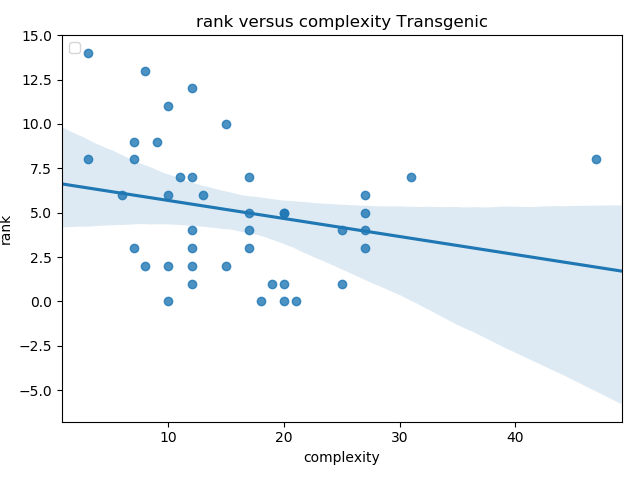
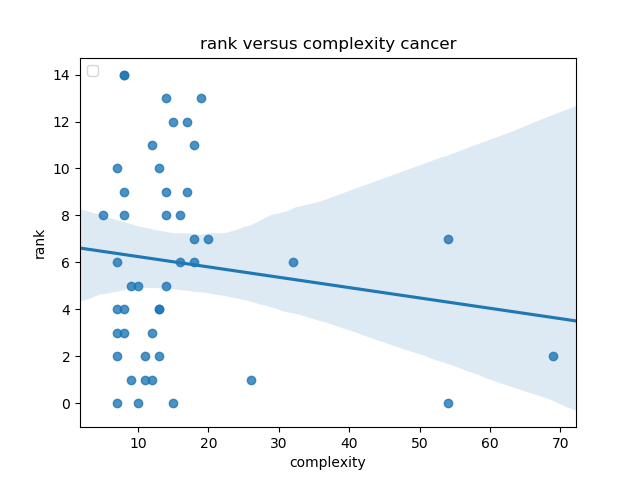
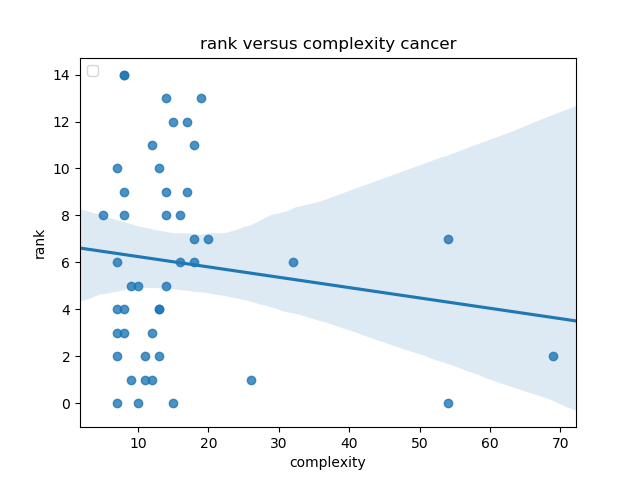


Searches for soccer also showed this trend



However, for as many examples of search queries where complexity increases with page rank, there were just as many counter examples. Prominent counter examples include Transgenic, and cancer





**Points of discussion:**

1. Text complexity vs. site ranking within and between searches

***Are simpler texts ranking higher in Google?***

***How does the language used by scientists compare to these rankings?***

(code is part implemented, the data is harvested, however plotting methods and analysis need developing)

* 1. For various scientific searches and various non-scientific searches

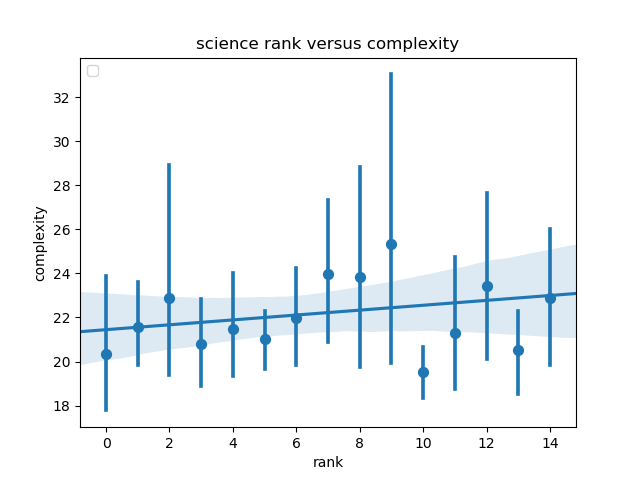
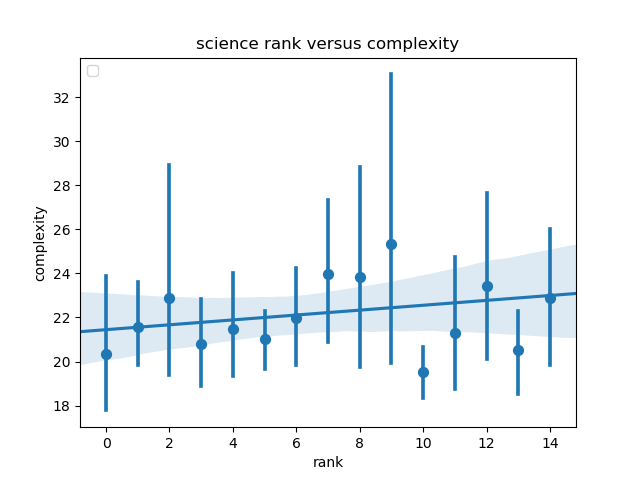
science\_searchss = [ 'evolution', 'photosysnthesis' ,'Transgenic', 'GMO', 'climate change', 'cancer', 'Vaccines', 'Genetically Modified Organism']

culture\_keys = ['reality TV', 'prancercise philosophy', 'play dough delicious deserts', ' unicorn versus brumby', 'football soccer']

* + 1. Science searches may be: Genetics, evolution, cancer, vaccine, GMO, climate change, photosynthesis
    2. Non-science searches may be: Soccer, culture, reality television, ???
    3. Compare text complexity between groups i and ii , but this is an overall comparison, as each group will have top vs. lower rankings and thus cannot be compared directly for rankings.
    4. Overall comparison:

(code is part implemented, need more ‘ideals’)

* 1. Also perhaps targeted comparisons of ideal educational websites vs average?
     1. Average: Wikipedia because it’s a classic go-to? Good. Also I think Wikipedia can be searched programmatically via duckduckgo’s bang expansion syntax
     2. Ideal Upgoer5, is an interesting ideal. Also the classic readability texts ‘The Readability of Scientific texts … is declining over time’, ‘the Science of scientific writing’.
     3. What else? Can someone create a list of ideals.



In the search results resulting from scientific engine queries, there is a weak but present relationship between page rank, and the reading level, necessary to comprehend webpages, the readability of the science based queries. A relatively stable mean value of 21.5 readability was observed.



# Discussion:

In this document, we have described variation in the readability of search engine results, and in some cases, we have characterized reduced readability across variety of web-based documents of cultural and scientific content. In addition to describing the nature of the readability problem facing scientific publication, we feel obligated to suggest possible remedies. One such remedy is the development of a browser-friendly interface to Stanford textstat metrics.

The writing of more readable scientific texts, could be increased by creating more incentives to write in simple, and low entropy manner. One way to incentivize readability is to create a leader board to foster social competition. Authors could have the opportunity, to submit there documents, to a word complexity ranking tool.

It’s possible that many science writers, agree with the goal of publishing accessible descriptions of their work, however, they may simply not find it convenient to maximize the accessibility of their writing, as doing so may conflict with other writing priorities. The convenience of accessible scientific writing could probably be improved with the introduction of new tools.

Many academic writers are familiar with MS Words spelling and grammar suggestion, where possible spelling and grammar transgressions are highlighted, aiding the writer's awareness, and review of their document. Additionally, a newer tool, Grammarly[ref to ext website], is able to make similar suggestions, however, it can also act on text entered into web pages. Grammarly operates as a browser plugin, and it cleverly acts on text fields of web pages. Additionally, Grammarly builds on the familiar, suite of suggested grammar based writing improvements, seen in word.

By analogy, Stanford readability text stat metrics could live inside Word, LaTeX and web browser based editors, nested inside the aforementioned traditional spelling and grammar tools. For instance, during the development of this document, text stat metrics as they pertain to this text body, where routinely computed, as a way of objectively previewing, the impact of new phrases on readability.

An existing tool, the simplewriter editor[https://xkcd.com/simplewriter/], in some ways occupies this new niche in readability tools, however, the simplewriter editor, is in part intentionally trivial, it is not intended for publishable scientific documents. The function of the simplewriter editor, merely restricts vocabulary to 10,000 of the most commonly occurring English words, it’s goal is to generally raise awareness to the general deficiency of readability in scientific writing, by exposing readers to a delightful and entertaining alternative, where we can experience texts of extremely low reading levels applied to highly technical subject matter.

Another recommendation to incentivize enhancing the readability of published articles, is to create a website, that acts as a readability ‘leaderboard’. That is when someone publishes something, they can voluntarily upload their text to a website, which enters their text into a textstat analysis pipeline, which gives there publication a readability score, if the writer is proud of their score, they can then opt for their publication to be entered into a hierachally sorted list, where the most readable text occupies the leading position, much like a video games ‘hall of fame’ page.

Unlike, in the case of the science/pseudo-science based corpus, in the plot of page rank, versus reading level complexity, across all of the cultural keys, a clear overall trend, of low complexity reading levels, with decreasing page rank is visible. Webpages of higher complexity, tend to have higher page rank, indicating that

1. Use of scientific vs. non-scientific terms

***Are simpler, yet less scientifically precise, terms ranking higher in Google (Yahoo, Bing, etc.)?***

* 1. Examples, based on currently ongoing conversation around science \*can be batch processed

(code is implemented)

* + 1. GMO vs. transgenics
    2. Global warming vs. climate change vs. anthropogenic climate change
    3. (though non-scientific, perhaps) Intelligent design vs. evolution

A list of appropriate blogs needs to sought, or the search terms for finding these blogs need to be developed.

* 1. Case study: targeted comparisons of scientist-led blogs vs. public-led blogs covering specific scientific subjects? \*can’t be batch processed

1. Text complexity vs. text sentiment (code is implemented)

***Are more neutral/factual websites more complex?***

* 1. Rank pro, anti, and neutral websites for text complexity
     1. Vaccines
     2. GMOs
     3. Climate change

Code for doing this is unimplemented:

* 1. Non-neutral sites also have more self-links rather than external links?
     1. Wikipedia would need to be excluded here

**Predictions:**

1. Websites with simple text will be closer to the top of a google search result.
   1. Overall, scientists will be more likely to write at a level of complexity that is higher than the average ranking of a top google result (above the average reading level for a person).
      1. This means that their work is less likely to be seen by the public

Partly Supported by graphs, although its not at all a simple linear relationship, of complexity neatly increasing with increasing page rank.

1. Simpler, broader terms are more likely to be higher in google rankings.
   1. Scientists are less likely to use these terms due to their
      1. lack of precision in terms of describing the scientific topic being discussed
      2. the thought of avoiding terms that are known and potentially controversial nature (e.g. evolution)

I think b, will require a lot of non code work.

* 1. a case study of scientific blogs will show that scientific topics discussed by the public will use simpler terms (and also likely less complex language (tying in point 1) relative to a scientific blog. Do scientific blogs have better (smaller) page ranks, relative to other scientific documents?

1. More neutral/factual websites will tend to be more complex.
   1. Factual websites will be more complex in comparison to websites who may take a stronger (cultural or social) position on a position, whereby they would rely on opinions and other values and less on a scientific rationale.
   2. The range of words utilized in subjective writing, is likely to at a lower reading grade level, additionally subjective writing may be more engadging to read, and more frequently written in an active voice which would be simpler in terms of their complexity, but more compelling with their regard to swaying opinion.

Additional questions:

-In Russell’s general search graphs, two clusters of websites seemed to fall out in the graphs. How do we figure out what is causing this? I am unsure if this is still true.

Issues to consider:

-Are the first few, super successful sites outliers? Should we run these with and without the first page of results to see the differences?

* Solution: run data collection script to include all data. And then run analysis and report metrics with and without these outliers. And tag them in graphs with a different color.

Also visit these websites.

Grammarly is open source, extensible and implemented in java script. Gunning fog can also be implemented in java script

Kuhn, Tobias. "The Controlled Natural Language of Randall Munroe’s Thing Explainer." *International Workshop on Controlled Natural Language*. Springer, Cham, 2016.

[The Controlled Natural Language of Randall Munroe's Thing Explainer](https://link.springer.com/chapter/10.1007/978-3-319-41498-0_10)

Grammarly

Japos, Genaro V. "Effectiveness of coaching interventions using Grammarly software and plagiarism detection software in reducing grammatical errors and plagiarism of undergraduate researches." *JPAIR Institutional Research* 1.1 (2013): 97-109.

<http://gunning-fog-index.com>

<https://github.com/words/gunning-fog/blob/master/index.js>

https://github.com/grammarly/eslint/blob/master/docs/rules/README.md