Word Complexity Project:

Testable hypothesis:

* Conventional scientific writing, has an extremely small variance about sentiment.
  + Academic writing has close to neutral sentiment, with a consistently small negative sentiment.
  + Non academic writing is very polarized about sentiment, it has large values of positive, or negative sentiment, and overall there is large variance.
  + Non academic writing has high subjectivity as measured by NLTK sentiment analysis, but academic writing is more objective.
  + A simple classifier (K-mean’s) can discriminate between academic and non academic writing, when it is fed data points situated in dimensions of writing complexity, and sentiment.
* Upgoer5 has been described as Randall Munroes ‘artificial language’, However, I argue that Wikipedia articles, also constitute an artificial language, with a writing style, that lends its self to automated classification. Wikipedia articles are highly edited and curated, by editors striving to enhance particular writing values, this gives articles a homogenous and detectable writing style.
* The Page rank of a website, is related to how many websites link to that website. Web pages with lower page rank (presented as most immediate search results) are consumed by readers more frequently, because they are easier to read. More easily read web pages, link to other more easily readable web pages, thus web pages that are more readable and less factual will have a lower page rank.
  + We found inconsistent and counter factual support for this hypothesis:
    - Web pages that scored high in complexity metrics sometimes had favorable page rank.
  + I don’t want to omit the refuted hypothesis. I would rather state this as a negative result. As it’s interesting to write about why.
    - Search engines with a liberal bias, might be actively intervening to promote fact based web content, and this could be good.

General hypothesis:

* The way that scientists communicate their work harms their ability to spread useful, factual information to the general public.
  + The language scientists and many science educators use online - a primary medium for the sharing of scientific information - is very complex
  + The language scientists and many science educators use online is also more complex than language used by many non-scientists and science deniers.

**Problem:**

Because scientists write at such a complex level, the most readable and findable online information for the general public is potentially less accurate (especially regarding controversial issues), while the most accurate information is likely more difficult to find and understand.

**Impact:**

The difficulty of finding and understanding accurate scientific information online contributes to a general misunderstanding of scientific fact.

**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 implemented)

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

(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.

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 sort, 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 decaying 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.

**Recommendations:**

Incentivizing Writing for enhanced Readability, via convenience tools.

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. In addition to describing the nature of the problem, we feel obligated to propose possible remedies. One such remedy is the development of a browser-friendly interface to Stanford textstat metrics

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[], 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 augments the familiar, range 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, in a way that coincides with 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 upgoer5 editor, in some ways occupies this new niche in readability tools, however, the upgoer5 editor is not intended for publishable scientific documents. The function of the Upgoer5 editor is to generally raise awareness, to the general deficiency of readability in scientific writing, by exploring the application of a very low reading level, to highly technical subject matter.

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.