CSE 546 Project Proposal: arXiv or snarXiv?

Tyler Blanton   Sam Kowash

November 15, 2018

The arXiv is a free online repository maintained by Cornell for preprints of scientific papers (many of which go on to be published in journals) from fields such as mathematics, physics, and computer science. Submissions are moderated via an endorsement system to ensure that only legitimate papers are posted, making it a valuable resource for scientists in many disciplines. In many areas of theoretical physics, almost all new papers are posted to the arXiv before publication.

The snarXiv is a computer program created by physicist David Simmons-Duffin that generates titles and abstracts in the style of arXiv's `hep-th` section using a context-free grammar. Simmons-Duffin created a game, "arXiv vs. snarXiv," in which readers are presented with two title/abstract pairs — one from `hep-th`, the other generated with snarXiv — and are asked to choose which one is real. After 750,000 guesses, the success rate was only 59%; i.e., people mistakenly chose the randomly generated snarXiv title/abstract to be the real one 41% of the time, which is pretty remarkable (and amusing).

Our goal is to use machine learning techniques to write a program that accurately classifies `hep-th` title/abstract pairs as real (arXiv) or fake (snarXiv). The arXiv has an API[1] for downloading metadata (including titles and abstracts), and Simmons-Duffin has made his snarXiv code available on GitHub[2], so we can sample large data sets from it.

We think this poses an interesting classification problem, as each corpus follows a relatively constrained syntax — arXiv the conventions of academic writing, and snarXiv the grammar constructed to parody those conventions — but one carries semantic cargo while the other is senseless. Discriminating between them can be difficult for humans, including (based on an informal survey of nearby offices. . .) those with substantial domain knowledge.

Our plan is to explore the performance of a variety of learning approaches on this task; for example, we predict that sequence models like recurrent neural networks may outperform n-gram models due to the grammatical syntax that snarXiv uses, and we would like to test this. We will also need to consider various methods of word embedding to transform text into usable features. Although the scope of our project is restricted to the admittedly impractical field of faux physics literature, the problem of distinguishing real vs. simulated writing does have real-world applications such as spam-filtering.

By the project milestone, we expect to have have calculated the classification accuracy for at least one method, and hopefully two: one n-gram model and one sequence model.

References

[Aggarwal and Zhai(2012)] Charu C. Aggarwal and ChengXiang Zhai. A survey of text classification algorithms. In *Mining Text Data*. Springer, Boston, Massachusetts, 2012. URL `https://doi.org/10.1007/978-1-4614-3223-4_6`.

[Cavnar and Trenkle(2001)] William Cavnar and John Trenkle. N-gram-based text categorization. 05 2001.

[Karpathy(2015)] Andrej Karpathy. The unreasonable effectiveness of recurrent neural networks, May 2015. URL `http://karpathy.github.io/2015/05/21/rnn-effectiveness/`.

[1] `https://arxiv.org/help/api/index`

[2] `https://github.com/davidsd/snarxiv`