arXiv vs. snarXiv

Tyler Blanton and Sam Kowash
University of Washington, Department of Physics

Background

The arXiv is a popular e-print repository for publications in physics, astronomy, and other quantitative sciences. It hosts nearly 1.5 million papers, of which ~120,000 are in theoretical high-energy physics (`hep-th`). Physicist David Simmons-Duffin developed a program known as the snarXiv, which procedurally generates abstracts in the style of `hep-th` from a context-free grammar.

Humans (even physicists) have surprising difficulty determining whether a given abstract is from the arXiv or the snarXiv. Over 750,000 guesses collected through an online sorting game, players succeeded at picking the genuine paper from a pair only 59% of the time. Try it for yourself below!

Abstract A
In the 20th century, a fair amount of work was done demystifying QED in the presence of a stack of canonical co-isotropic branes. In this paper, we make contact with analyzing heterotic strings, consequently reconstructing perturbation theory on $\mathbb{C}^n$, and classify anomalous dimensions in loop models with sleptons. Our computation of the solution of magnetic dualities in models of hadrons provides a certain notion of perturbation theory (taking into account cosmic rays at Λ_{QCD}). Our results prove that decay constants turn out to be equivalent to an instanton at the Planck scale. Finally, we establish that sleptons can be brought to bear in reformulating heavy ions.

Abstract B
We study the effective action of the heterotic string compactified on particular half-flat manifolds which arise in the context of mirror symmetry with NS-NS flux. We explicitly derive the superpotential and Kähler potential at lowest order in α' by a reduction of the bosonic action. The superpotential contains new terms depending on the Kähler moduli which originate from the intrinsic geometrical flux of the half-flat manifolds. A generalized Gukov formula, valid for all manifolds with SU(3) structure, is derived from the gravitino mass term. For the half-flat manifolds it leads to a superpotential in agreement with our explicit bosonic calculation. We also discuss the inclusion of gauge fields.

We explored computational approaches to this classification problem to investigate whether its difficulty stems from snarXiv's genius, or humans' bewilderment in the face of unfamiliar jargon.

n-gram model

One approach to text classification is to develop a *language model*. Given a class $Y \in \{-1, 1\}$ (with the negative sign referring to arXiv and the positive to snarXiv) and a document X consisting of words $\{w_i\}_{i=1}^N$, we want to characterize the probability

$$\mathbb{P}(X|Y) = \prod_{i=1}^{N} \mathbb{P}(w_i|w_1^{i-1}, Y),$$

where w_1^{i-1} is the sequence of the first $i-1$ words. The sample space dwarfs the available data and this can never be satisfactorily trained. Instead we assume that a word's probability depends only on the preceding $n-1$ words, reducing the training task to estimating the probabilities of so-called n-grams from their frequencies in the training corpora.

$$\hat{\mathbb{P}}(X|Y) \equiv \prod_{i=1}^{N} \hat{\mathbb{P}}(w_i|w_{i-n+1}^{i-1}, Y) \equiv \frac{C_Y(w_{i-n+1}^{i}) + 1}{C_Y(w_{i-n+1}^{i-1}) + V},$$

where $C_Y(w_{i-n+1}^{i})$ is the number of times that n-gram occurs in the corpus for class Y and V is the size of our vocabulary. This definition inherently gives low weights to longer abstracts, so for classification we compare a geometric mean, the perplexity:

$$PP(X|Y) \equiv -\frac{1}{N} \sum_{i=1}^{N} \log \hat{\mathbb{P}}(w_i|w_{i-n+1}^{i-1}, Y).$$

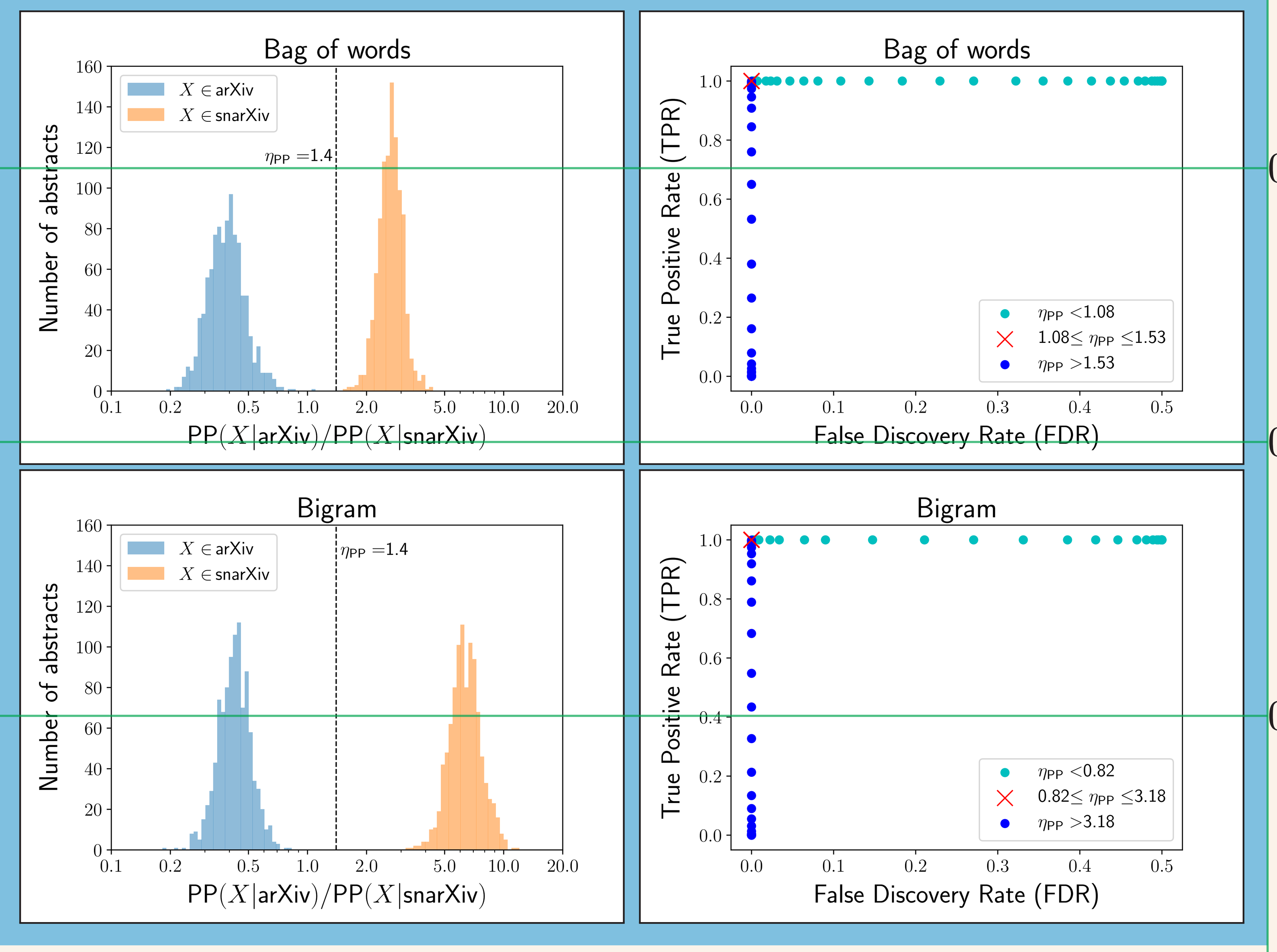
Likelihood-ratio classification

We classified documents based on n-gram frequencies in two ways: in *naive Bayes classification*, we classify according to

$$\hat{Y} = \arg\max \mathbb{P}(Y)\hat{\mathbb{P}}(X|Y),$$

where $\mathbb{P}(Y)$ is the fraction of training abstracts in class Y.

Results



References