The Curious Case of Neural Text Degeneration

Nan An, Yannis Fontaine, Newman Chen

**Introduction**

For our project, we chose to study the paper titled “The Curious Case of Neural Text Degeneration” by Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. This was originally published in 2019 which is important to understand the context as we begin to analyze its effectiveness. In brief, this paper brings up a common problem for text generation in which, when using maximization based decoding, the generated output ‘degenerates,’ meaning that it becomes bland, incoherent, or repetitive. The solution proposed is named ‘nucleus sampling,’ which is itself a new approach but clearly heavily inspired by other approaches, especially evident with its similarity to the top-k approach it describes. Essentially, this involves sampling from a ‘nucleus’ of probability samples where the most likely tokens lie and forming this nucleus by taking the probabilities that make up the top p percentage of this mass. For example with p=90, we would take enough different options such that the probabilities of these tokens would represent the top 90% of tokens that could be generated. Other previous methods such as the top-k method in which we choose the top (fixed) k most likely tokens, ordinary sampling, beam search (which explores a number of possibilities and takes the most likely one), and greedy, are all supposedly worse. We will now describe a bit more the problem set-up and motivation.

**Problem and motivation set up**

As noted earlier, at the time of the publishing of this paper, LLMs (and specifically GPTs) were much less widespread and less powerful/effective. At this time GPT-2 was the latest model by OpenAI and the one primarily used for the purposes of experimentation within the paper. Thus, this problem of degeneration in text generation was a much bigger problem and a worthy motivation for the research done by this team. The paper shows that the different approaches related to maximization—namely, greedy, beam sampling, sampling (with temperature), and top-k, all produce different problems. We felt the paper took a good approach to demonstrate the flaws in these approaches, namely by providing concrete examples of how the different approaches perform for the same task, in which some devolve into repeating the exact same phrase multiple times or some drift off-topic and include some nonsensical text within the generation. The paper does so for every single approach, including the one they develop that is supposed to resolve these problems. However, this is a difficult thing to represent as a general case—that is, it is difficult to represent that an approach such as beam sampling will repeatedly devolve into incoherent or repetitive text generation when there are so many different contexts/possibilities to account for. Though it does also include some more unique graphics that demonstrate the possibility of this degeneration, such as showing how the generation of the phrase ‘I don’t know’ leads to a positive feedback loop in which it will continue to repeat itself, we felt it could have done a better job of more objective, varied proof that such degeneration is common for these approaches. Apart from one large table in which it lists the different scores (perplexity, Self-BLEU4, Zipf Coefficient, Repetition %, and HUSE), which are themselves not very well explained/elaborated upon, we mostly only see these one-off examples which demonstrate the point well but are not the most convincing in terms of generalizing this problem. That being said, the motivation is still posed well and we can see it is certainly a problem, especially for an important task like task generation and especially for this time period in which the GPT models were not as powerful.

Furthermore, we think the paper can do a better job of truly justifying what they identify as the real problem: the ‘unreliable tail’ of the probability distribution that causes errant tokens/phrases to be generated. Though the paper repeatedly mentions this as the main problem and makes it the main issue that their approach, ‘nucleus sampling,’ targets, other than a few sentences of theory that state that this is problematic the paper does not consistently/rigorously prove this to be the primary issue for the degeneration of text.

**Experimental setup**

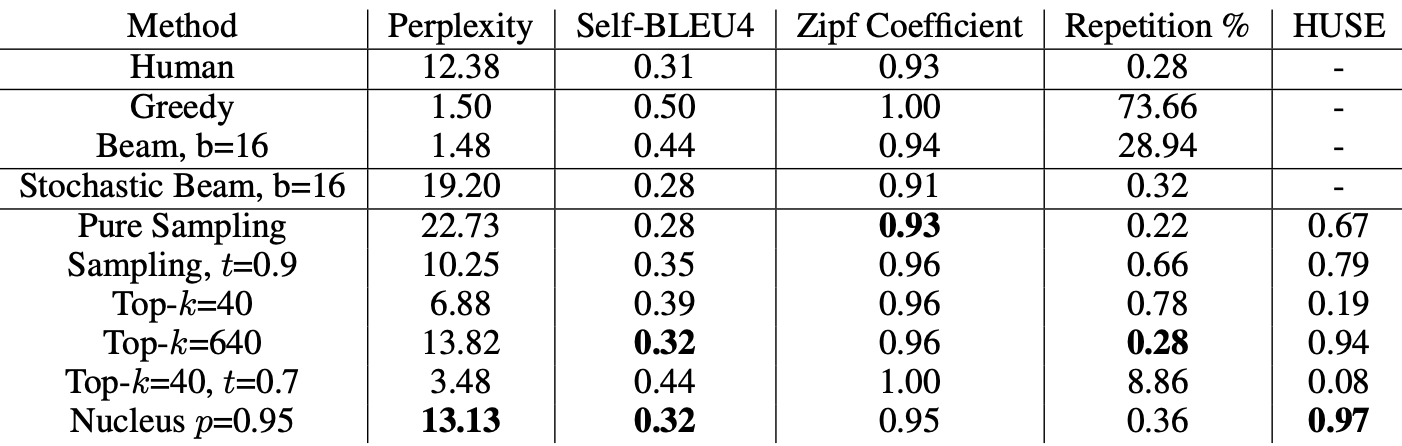
The experimental setup, in our opinion, is mostly sufficient. Their setup involves generating 5,000 text passages up to 200 tokens (or the end of sequence token), which are generated conditionally on the WebText dataset, which is itself scraped from the web. As our target is mostly just producing ‘normal’ text from a variety of different contexts this experimental setup seemed fine to us.

The paper continues by describing different evaluation methods that are important. It first focuses especially on perplexity, which is the certainty of a model of its prediction. They argue that the optimal model should generate text that has perplexity closest to that of the gold text, rather than just being absolutely certain of everything it does. Using linguistic justification such as Grice’s Maxims of Communication to show that normal, human language isn’t always the most obvious, and that perplexity that is too high leads to repetition and low diversity, we feel that this is properly motivated. The paper then continues on to discuss the Zipf distribution which further strengthens this idea of achieving the closest perplexity to the gold, human standard, as well as Self-BLEU, which measures diversity, and a simple count of the amount of repetition. Finally, the paper touches on human evaluation with the Human Unified with Statistical Evaluation (HUSE) score. All these serve as different measures/criteria which justify the effectiveness of an approach and which constitute the evaluation of the different experiments. Though we feel these are mostly good in principle and account for important aspects of text generation, we feel the article does not consistently explain these different criteria well. Especially for the Self-BLEU score, which is simply stated to measure diversity, and the HUSE score, which is stated to include human evaluation, we feel there is a real lack of explanation and intuition behind why these measures are important and even how they work.

**Results**

As we see in our section on the experimental setup, though we feel the experiments run are logical and largely sufficient, there is a lack of description and intuition provided for some of these different scores to evaluate the methods. As mentioned before, there is one large table that states that their method, nucleus sampling, does quite well across all the categories, but apart from the pure numbers that demonstrate this there is not much explanation or intuition provided. As a result the results provided are somewhat convincing in that the examples provided make it seem trustworthy, but we feel there should be more of an effort to make this into something generalizable and evident across many different contexts rather than providing one specific example in which nucleus sampling does well along with some numbers without sufficient explanation.

Furthermore, we see within the table of evaluation metrics that their proposed solution, nucleus sampling, seems to provide mostly similar results to that of top-k sampling with k=640.



As we see here, the numbers for these two methods are extremely similar across the board, with top-k=640 even doing better with the ‘repetition %’ metric. We recognize that they are not saying that their approach is necessarily perfect and is also very similar to the top-k approach in theory, but we feel that these results do demonstrate it is better but could be made more convincing. One thing that could help is perhaps a measure of statistical significance for the improvement offered by nucleus sampling over top-k.

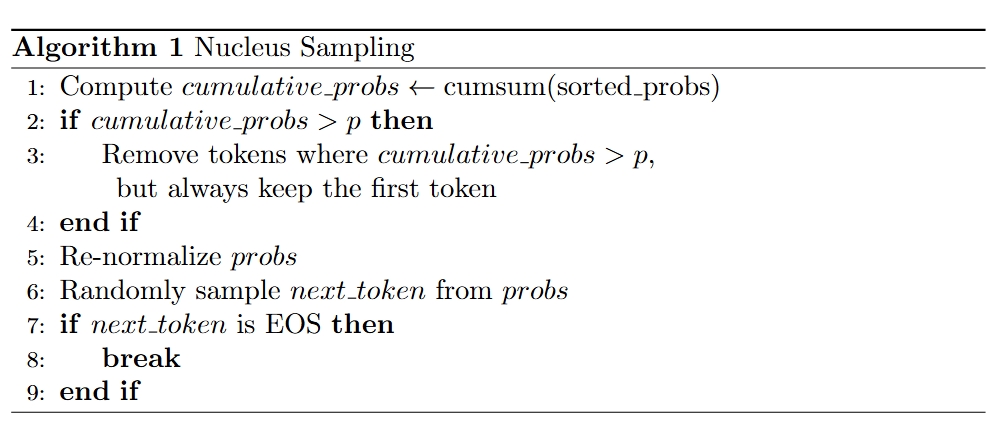
**How this connects to further research and impact in general**

The paper offers an improvement on an important field of research, especially when we regard retroactively the progress made in text generation and NLP and LLMs in general and the importance and ubiquity of text generation today. Thus, we feel that this research not only connects well to further research in a quickly developing field but searches to address an important problem in general, which strengthens its positive impact. We see that text generation is a truly important subproblem in the field of NLP and thus understanding not only the proposed approach of nucleus sampling, but understanding all the different past approaches examined and their pros and cons is important for understanding advancements in GPTs and the research being done in text generation. Though this represents just one optimization of one small part within the entire system, this is an important cog in the context of encoders and decoders in general. As a result, this contributes directly to the building of GPTs, whose influence today cannot be understated and is sure to grow even further. We thus feel that this paper serves an important purpose in an important field and what we have learned in this paper is very significant especially when we see the way the landscape of NLP has evolved today.

**Our overall thoughts and impressions**

Overall, we view this paper positively and feel it has contributed something significant to an important field. Though it mostly occupies a small space within this large, important field of text generation and GPTs in general, this ‘local’ improvement definitely has had a substantial impact on such an important topic today. Furthermore, we felt that the paper was well-done overall. As a simpler problem that we can somewhat abstract away from the larger, more general problem of text generation, it is easy to grasp yet important due to the importance of text generation today. This simplicity causes us to feel that this paper effectively addresses this problem of text degeneration and thus makes a significant impact on this important and relevant topic. It does so while being well-written, structured, clear, and well-motivated. Though some of the evaluation metrics and problem statements/motivations could be better elaborated upon, we feel this paper effectively addresses this small yet important problem succinctly and in a way that is easy to understand. As this approach has stuck around and is used today by many transformers, the impact of this paper is certainly substantial which is helped by the fact that the paper itself is well-crafted.

**Our own code and results**

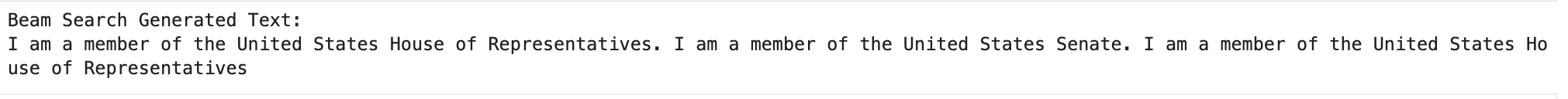
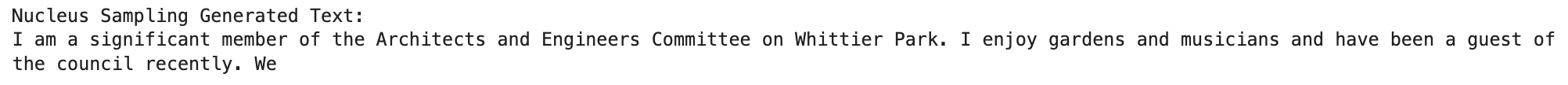
****

For our simple experiment we implement beam search, top-k sampling, and nucleus sampling and compare the outputs to see their effectiveness. Nucleus sampling, as we described, generates text by taking a subset of tokens that has at least a cumulative probability p.

For the implementation of nucleus sampling, the model first computes the cumulative probability distribution of the sorted token probabilities. We then ‘truncate the tail’ by removing all tokens whose cumulative probabilities surpass this probability p. The remaining probabilities are then re-normalized, and the next token is sampled from this set. We repeat this process until the EOS token is generated or we reach the maximum length.

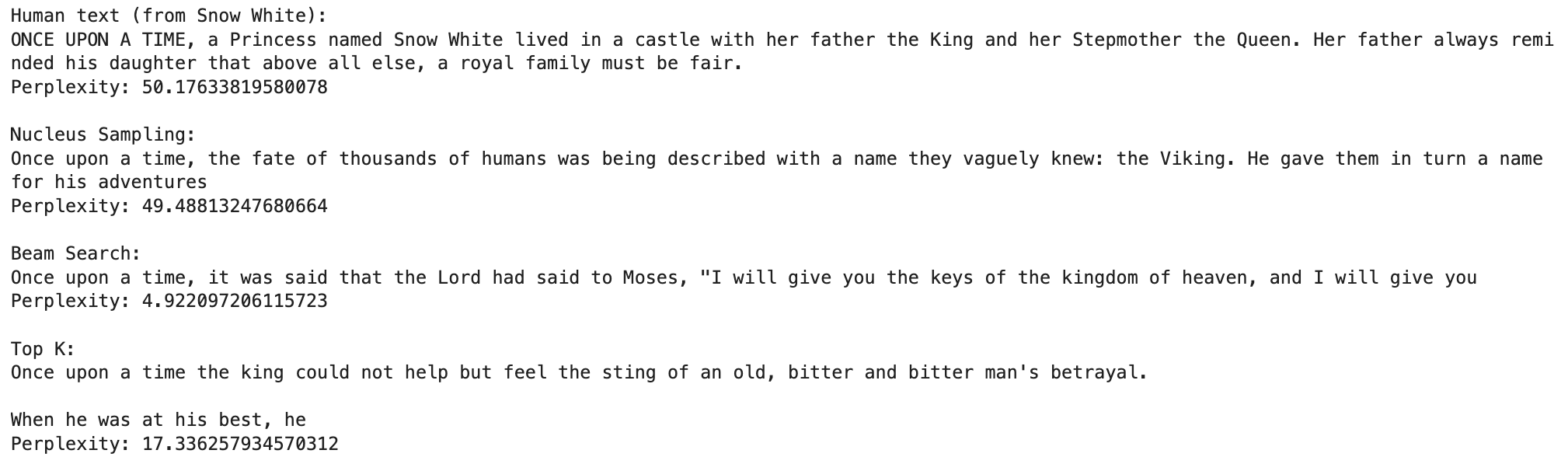
An example to help understanding:aAssume we have a probability distribution of 4 tokens, probs = [0.6, 0.3, 0.1, 0.1], with p =0.9. Generating a cumulative probability distribution gives us [0.6, 0.9, 1.0, 1.0], and the third and fourth indices are removed. After re-normalizing, we have probs = [0.6/0.9, 0.3/0.9, 0, 0] = [0.67, 0.33, 0.00, 0.00]. We then sample from just these top 2 tokens.

We use our human judgment along with their suggested measures of perplexity and the Zipf graph and coefficient to evaluate. Here are some of our results:

Generated texts:

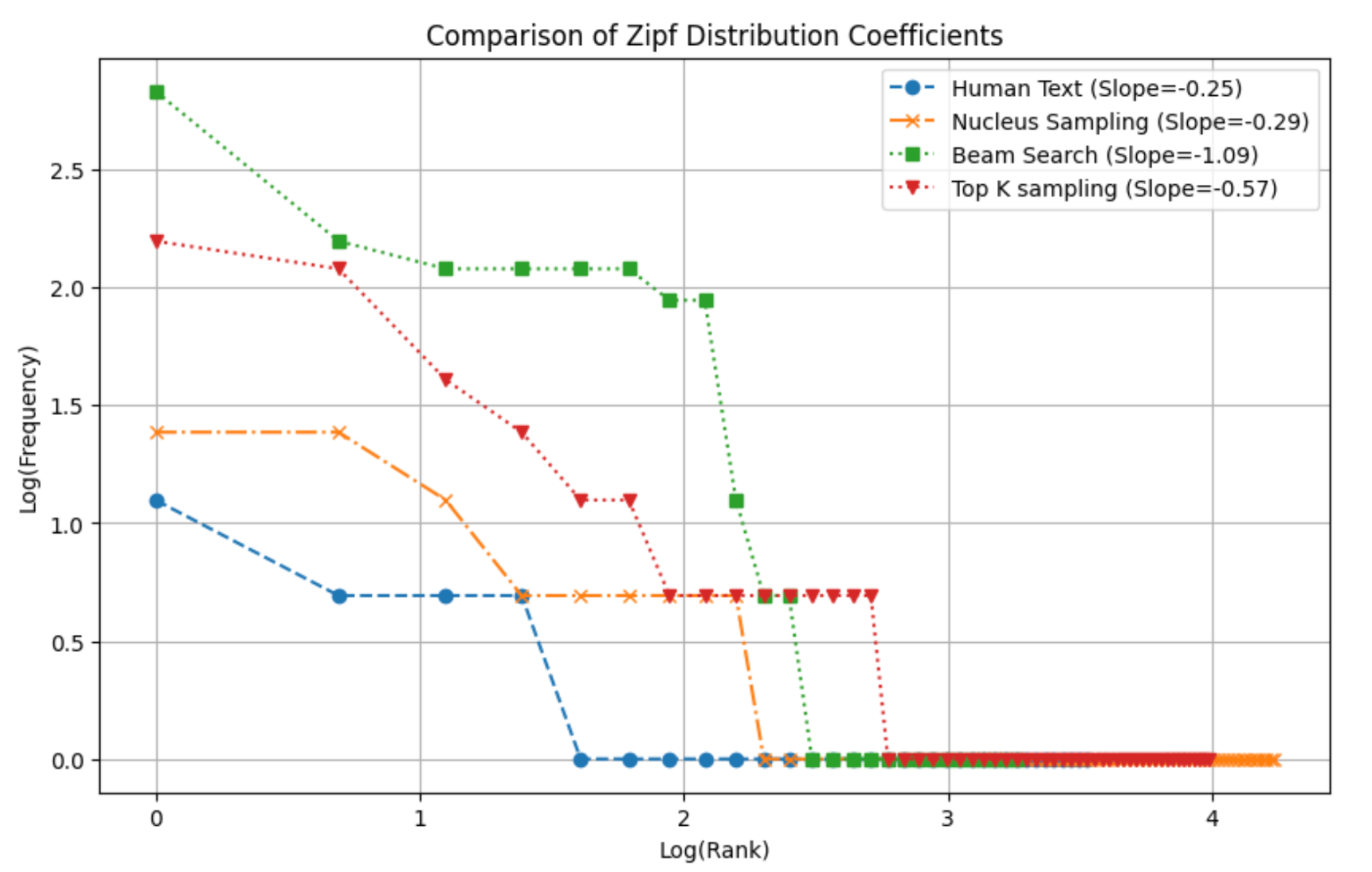
Because many fairy tales, such as Cinderella, begin with the sentence “once upon a time,” we chose this as our “human text.”

Perplexity comparisons:



Interestingly, human-written text has the highest perplexity, which shows its inherent complexity and diversity that is difficult to fully replicate especially with this open-ended prompt. Nucleus sampling is the closest to human perplexity but still maintains high quality generation.

Zipf coefficient:

****

The Zipf coefficient shows that nucleus sampling is the closest to the human distribution.

We see some of the benefits of nucleus sampling from our experiments. As we can see from the generated texts, we have a better balance of diversity and quality. We also note that adjusting the p value can allow for more tinkering with this balance between diversity and quality; a higher p value would allow for more varied/creative outputs, perhaps in a task like story generation, whereas with tasks that value accuracy more we can lower the p value to only use more likely tokens. We see that this algorithm both promotes diversity with the aspect of random sampling, but also accommodates for quality by truncating the less likely, ‘lower quality’ tokens. Thus, the model more successfully approximates human text overall.

In conclusion, as we can see, we do notice an increase in effectiveness compared to beam search and top-k sampling. Not only is the text generated more logical and devolve/repeat less, but we see that the perplexity and Zipf coefficients are much closer to human text. Our experiments match the paper’s results and suggest that nucleus sampling is indeed a more effective method for maximization during text generation.

**Use of AI assistants**

We used AI assistants along with some help from the slide/code from the course and labs in order to help us generate the code for our experiment.