Neural Article Generator

Abstract

I will be looking into a machine learning algorithm with the primary objective being article generation; a system which will help in writing informative articles most likely to catch the reader’s attention. The dataset will be a set of 250,000+ articles on various terms which garner the most traction when a given topic or keyword is searched on Google as well as a comparatively smaller set of articles which do not do so well. I plan to use a combination of both BERT and OpenAI – GPT to address different aspects of the generator.

Introduction

Writing articles plays a very important part of the outreach strategy for almost all organisations, ranging all the way from clubs in schools and departments to major companies around the world.

However, a major issue often comes up is repetitiveness. When most of the writing is done on a similar subject or by a small group of writers, the articles often get stale with similar writing patterns and it becomes very difficult to maintain the attention of the target crowd.

On the other hand, if someone is new to writing articles or is trying to take on a new topic they often lack the idea with respect to what terminology must be used or how an article should be structured to entice readers into going through the entire post.

This generator is aimed to help such problems. After going through thousands of articles on various topics it will be able to predict to what keywords trigger the most clicks on a given article compared to others.

Since there are a lot of topics, I will focus primarily on articles from the fields of Machine Learning, Robotics, Electronics and Software Engineering. Since these topics are often aimed at a similar crowd, the algorithm will be able to pick up on patterns on the structure of articles even better thought the keyword encoding will be domain specific

Problems to Be Addressed

One of the major hurdles that will have to be overcome is that we will not only have to provide topic specific keywords but also incorporate them into well formed sentences/paragraphs. The generated text should not only make grammatical but are also easy to read and attention catching. This can be addressed by implementing a double layered model, one for learning the contextual significance of various words used.

When it comes to keyword prediction it is important that we filter out commonly used words like articles, prepositions and such. To do this we will have to train the generator on all articles irrespective of the topic and training it separately on domain specific text, followed by filtering out the set of words obtained from the generalised training.

An issue which would also be addressed by grouping of articles is article structure. For different genre, articles have to be written in different ways because they are not only about different topics but have a completely different target audience. Given the topics that have been selected the generator will initially be catering mainly to the STEM community, more specifically the Tech niche.

The other big impasse is filtering out clickbait articles. This will be a problem we face will creating the datasets. These articles often get a high number of views but they lack in depth and quality of writing. As a result, these articles will be very good for capturing keywords but will be hinderance to the sentence formation segment. Further more it’ll be a challenge to automatically filter such a huge set of articles into clickbait and quality articles