Thomas Garvis, Ngram Word Prediction Algorithm

IT1. Artificial Intelligence; Machine Learning; Natural Language Processing

Key Words: n-gram, natural language processing, computational linguistics

Overview: The goal of this project is to create an algorithm that can shift through numerous ngrams to predict which word will most likely be typed next. The Ngram Word Prediction Algorithm will predict a user's next word in any number of given circumstances. Using almost all the books in the world as a collection of data, this algorithm will be able to predict the most common word that would appear next in real time. The user would put in a phrase one to five words long. This algorithm reads through the collection of words then provides the user with a couple of options the algorithm believes should be the next word. Ngrams make this possible. An ngram is a contiguous sequence of n items. In this case, words and sentences from books are used as ngrams. These ngrams can be used to create a new word prediction and query completer algorithm. Once this algorithm is developed, it can be put into many different applications to solve different problems.

Intellectual Merit: This Small Business Innovation Research Phase I Project creates a new way to word predict and opens the possibility to find better prediction algorithms than the ones we have now. This algorithm would help spark interest in the new ngram technology. Google released a large amount of data, through ngrams and Google Books, to the public for any use. Since all this data is still new, this is the first autocomplete project utilizing this valuable information to produce results. Some difficulties that could arise would be the speed of the algorithm. When shifting through the copious amounts of data sets, running a precise and speedy program will be ideal. After this algorithm is done, it will be implicated into different application to solve different problems.

Broader Impact: This algorithm could be applied to countless applications to perform and solve different problems. An in-time word prediction algorithm would fit into a writing program to help authors and writers with their work. With the proper collection, this algorithm could rival search engine's word predictions. When added to a children learning application, this algorithm could enhance the teaching of children. Kids can type and see possible completion of sentence, thus learning sentence structure. They would be able to substitute words to see how it would affect a sentence. Ngrams are a powerful technology and this project will promote them and encourage other people create more products that use Ngrams. This algorithm is very flexible and would be able to support different collection such as number or characters, not only words.

Elevator Pitch

Why do we study history, or look to the past? To improve our future and learn from our mistakes. This question reaches farther than just social sciences, it is a pressing question in tech development as well. My algorithm will look at 5 million book over the past 500 years to answer this question in its relation to word prediction. In 2010, Google broke down their collection of online books into ngrams. Starting at unigrams all the way up to fivegram. Google attached two pieces of data to each gram: word frequency, and book frequency sorted by year. Moving forward, I will be the first one to create an in real time word prediction algorithm based on Google's Ngram data. The most important aspect of this project is the flexibility of my program, due to its ability to interpret different and new data sets. Children's education is rapidly digitizing, and this application will be able to revolutionize interactive typing programs by using predictive algorithms. Furthermore, due to the globalization of current industries, accurate foreign translation needs are at an all time high, using Google's foreign language data set this program has the ability to improve real time translation. Additionally, Google's auto complete search algorithm keeps track of search terms by region, this program will be able to interpret this data and even improve upon Google's own search algorithm. I am currently on the forefront of predictive word programing, and the possibilities are endless. These possibilities have not been fully realized yet or the programs that have tried predictive word programming have failed to come to market. So why do we look to the past? To improve the future

Commercial Opportunity and Social Impact

My Auto Complete Algorithm not only has commercial opportunities, but also impacts society depending on how the algorithm is used. My purpose for this algorithm is not only to create a top of the line auto complete algorithm, but also to spark interest in a technology and certain types of data. To improve my chances at future success, I have a business model set up and how I would go about getting involved in the market. This algorithm can be used to better life and society. Most products created using this algorithm would be improvements on what we already have. They will make jobs simpler and more efficient. My goal is to enrich the world with products that can improve upon what we have today by using the data we have collected over our lifetimes, and to create these products, one must have a sound business plan to follow.

My business model would include a combination of two different plans. The first would be selling the products I create with the algorithm as the core. The second plan, people would involve a becoming a contractor. People would come to me with a product in mind that requires my algorithm, and I would create applications that solves their problem. Schools could ask for a learning tool and I would sell my product by computer. The second plan would be selling the idea of the auto complete algorithm itself. The possibilities for this algorithm are so large that I would never be able to cover and create everything this algorithm can be applied to. To expand more, I would sell my algorithm for other people to create their own products. By selling my algorithm to others, I will have different forms of revenue. This algorithm by itself has value and can be marketed, and coupled with producing products, I would be an active player in any field or market my algorithm is being used in. Contributing to other's and creating my own products will give me more influence in the markets I target and the ones I am not directly related to.

This algorithm applies to multiple markets. The flexibility of the algorithm will allow numerous applications in different field and markets to be created. Not only will I be addressing the markets for auto complete applications, I will be selling the algorithm to entrepreneurs who want to make their own products that will benefit from my idea. For dealing with data on such a large scale, the market for auto complete algorithms is small and single track focused. You can purchase Google's auto complete, but it is only limited to search queries. I am creating something that can be applied to different fields depending on the collection of data you are starting with. By selling the algorithm and contracting applications out, I run into some risks with my markets.

One of the issues of supplying the market with an algorithm and products that use it, is that I would be selling to some of my future competitors. I would not find this to be an incredible pressing issue as the flexibility will promote people to explore other fields and options and my technical knowledge of my own algorithm would give me the competitive advantage. The most important risk is being one of the first people to create something like this. I must pave the way and do not have anyone else's mistakes to learn from. Another risk is that most markets are entrenched with their current technology and may not see the value in my algorithm or how it can improve upon what they already have. Most of these products will be a quality of life or something that can be better, but there may be a cheaper option. For example, a school could approach me giving me requirements for a learning tool that requires my algorithm. Even though this tool could benefit the students more, not paying for the product and only having a teacher will always be the cheaper alternative.

Not only and I trying to create an algorithm that can be used in almost every field, I am trying to spark interest in new data sets and new ways of thinking in Natural Process Learning. Part of this algorithm is to encourage people to think about what can be accomplished with the right data set. For me specifically, I am trying to use Google Ngram data. In the past this data set was only used by historians to plot the trajectories of words and phrases over time or the word frequency of historical words frequency in rap lyrics. For a collection of data derived from almost all the English books for the past 500 years, why is this not being used for more? In today's world, we have more data than we know what to do with, and creating

an algorithm that's function can change depending on the data set, will promote the use of all the data we have collected over the vast years. I am trying to create products that will increase the quality of life, which can only be used to make things simpler and easier. Depending on the application, this algorithm can make learning in school easier and learning languages better. A major issue we have in the world is the language barrier. Most class and learning applications are too structured and have rigid vocab. One can sit in class all you want and learn, but never become a fluent speaker. The relevance of material taught in today's foreign language classes are questionable. Google Ngram data is not only limited to the English language. By using the other language collections, one can find the most common vocabulary for a language. With my algorithm, you would learn most important vocabulary and construct sentence with auto complete. Women are illiterate all over the world, which is a problem. The algorithm can be used in literacy tools in underdeveloped area. Studies show that one percentage point increase in female education raises the average gross domestic product (GDP) by 0.3 percentage points and raises annual GDP growth rates by 0.2 percentage points. Changes like this would have a global impact benefiting countries everywhere.

This auto complete algorithm is nonintrusive that impacts only what we opt in. It does not force people or change someone's way of life. It focuses on more of a quality of life system. I cannot see how an algorithm that can predict a gram or construct a sequence of characters can be used to cause harm to the environment and others. On the contrary, it can be used more in everyday life to improve what we already have. The negative impacts are very small, but the potential impacts on improving daily life and common applications can be huge.

This algorithm has the ability to create great positive movement for different markets and the globe. This is all made possible through the flexible nature of the algorithm. Being able to use different collections of data changes this auto complete word algorithm into something entirely different. When put into the proper application and with the right data set, this algorithm can be the core in teaching people how to read or learn different languages. It can address larger global issues as well as small community issues. The algorithm also is not contained to only one market. I intend to have a contracting business and create products that will solve problems with this algorithm as its base. My business will also sell out the algorithm to allow more people to use it and spark interest in Natural Learning Processes as well as show how big data sets can be used efficiently to solve different problems. One would think that this algorithm must have large impacts for being able to expand into many different markets, but the negative impacts are rather small. This algorithm has the chance to make a difference once it gets out there, and that is my intention with my product.

Technical Discussion and R&D Plan

Before talking about the technical innovation and the R&D plan, there are some key challenges and risks in bringing this algorithm to market. The challenges lay in the data processing, crafting a speedy algorithm, and data storage. For data processing, finding the best way to compress and filter the original information from Google will greatly improve the performance the of algorithm. With proper data processing, the algorithm will not have to do as much work, and overall will take less time. Without a fast algorithm, the customers will be unhappy with any product that can be created using this algorithm. Maximizing an algorithm is always a technical challenge because it takes clever techniques and an expert's eye to produce results. The last of the challenges is the way the data will be stored. Picking the right database will be crucial in the speed of the overall algorithm. With proper research and knowledge about the data storage, the algorithm will be able to pull the necessary words the user will want faster. The major risk bringing this algorithm to market, is the potential of it sitting on the shelves for years until someone tries to use it to solve problems in a way that consumers will buy. I want to prevent that from happening by creating my own applications and sparking interest that way.

The first innovation is the data set that the algorithm will be using. Google Books has almost every book from the last five-hundred years. Google scanned each book to create their ngram data set. This data set tell us how many times a sequence of word, length one to five, appear in books. What better way to

predict the next word while writing than the words authors have been using for years? Another element to this innovation is the not only using the frequency of these phrases, but also the part of speech of words. Google also provides part of speech for most words. Finding a way to combine word or phrase frequency with part of speech will make this algorithm extremely unique. The next innovation is how I manipulate the data Google has provided. The data set is not one that can be immediately used for this algorithm due to its size and set up of the data. With some clever filtering and formatting, this once unusable data becomes the critical aspect of word prediction. This data set was originally used to track popularity of words or phrases by year. No one is using this data in the way that I am. I will be the first to utilize Google's ngram data for more than its original design.

The key objectives start with finding the correct database and way to store the data. This is important in the research stage because proper indexing will be important when searching for the next word or set of words. Following along with this is finding the best structures to hold the data and formatting the data. Researching the best structures to hold the data will help on the data storage side, but also the algorithm filtering through all the data. Given the scope of having to do queries on one grams all the way up to five grams, the possible answers for the next word will not be small. Another objective is the algorithm itself, so research on best practice will help formulate the desired algorithm. Doing research on data formating will also be beneficial to the project. Answering question about what type of files the data should be in and how they should be formated in those files are important to the whole data processing part of the project. This project has a couple more questions that will determine the feasibility of this algorithm. I will need to answer the question about the amount of storage I will need. This will be tricky since I will be doing a fair amount of data manipulation and there is no way to exactly calculate it without trying. Considering how much computer memory I can use and how much data I can load onto the computer at a given time is also a factor of usability.

This project has a couple of milestones that mark important gains in the project. The first one is the completing a clear data processing. This will take some testing as the project continues, but finding new ways to clean and format the data to better fit the algorithm is beneficial on all fronts of this project. Load and storage testing will also be important to project. Once the algorithm process all the data and does not require a large amount of computer memory, I will have reached another milestone. Having a base algorithm is also another milestone. The algorithm has to start from somewhere, so creating an algorithm that works but not optimally is the one of the first steps. The last milestone would be to take this base algorithm and improve one it. I would make it operate better and faster than before trying to reach speed quotas. Once these speed quotas are met and the algorithm produces result, we are approaching the end of the project.

My R&D plan consists of completing my technical challenges as they pose the greatest difficulty to me and the testing of these challenges. The first couple months will be the data processing and data storage. I will be done this process in December and be able to start working on the algorithm in the winter. In January, I should see some good progress with the algorithm and continue to work on it. By the time March comes around, the algorithm will finished or close to being finished. The first part of this timeline, the data processing and data storing, will take experiments to find optimal ways to format and store the data. This will be done with a basic algorithm and the time to retrieve data will be recorded. When getting into the algorithm portion, I will be testing not only time it takes to produce results, but also strain on the computer. The algorithm needs to be able to produce quick results, but if a computer cannot handle the program, there is no point.