Name: Eljan Mahammadli
Course: Guided Research Methods
Project Title: Name Generation with Autoregressive Character-level Language Modeling

1. What am I going to do?

The aim of this research project is to develop an AI-powered system that can generate new and unique names. The system will utilize the power of character-level language techniques to build simple model such as n-gram all the way up to the RNN and Transformers.

The proposed system is designed to generate novel and appropriate name suggestions by taking a list of input names, whether they belong to baby names or company names. The system's core objective is to generate distinct and unique names that do not already exist within the given context. Notably, the system possesses the capability to generate outputs in multiple languages and domains.

For instance, if provided with a dataset of company names, the system will utilize autoregressive character-level language model to generate cool and unique company names that are appropriate and linguistically relevant. Similarly, when given a set of baby names in English, Azerbaijani, or any other language, the system will be capable of generating new and name-like baby names specifically tailored to that language.

This versatility is achieved by modifying the character set (alphabet) and leveraging the input names as the basis for generating new and distinctive ones. By dynamically adapting to different languages and contexts, the system has the potential to offer a wide range of creative fitting name suggestions across various domains.

2. How is it done today? Current Limitations?

Currently, generating new names often relies on manual brainstorming or consulting existing name databases. These methods are time-consuming, subjective, and may not guarantee the creation of truly unique names. Additionally, the process heavily relies on human creativity and domain expertise, which can limit the scope and diversity of the generated names. Nowadays, there are also n-gram models some of which will also be tested in this project but more advanced methods such as MLP, RNN, GRU, and Transformers will be utilized to perform better and generate more name-like and reasonable outputs.

3. What is your idea to do something better?

The proposed approach leverages the power of autoregressive character-level language models, specifically Transformer-based models like GPT. By training the model on a dataset of existing names, the system can learn patterns and linguistic structures specific to language. This enables the model to generate new names that are appropriate, linguistically coherent, and distinct from the existing names in the given context. The system will also be able to scale to different domains and languages.

4. Who will benefit from your work? Why?

The research project will benefit individuals and organizations who are in need of new and unique names. For parents looking for distinctive baby names, the system will provide a valuable resource. Similarly, entrepreneurs seeking memorable and creative company names will find the system beneficial. By automating the name generation process, the system will save time and effort, while offering a wide range of relevant and original name suggestions.

5. What risks do you anticipate?

One potential risk is that the generated names may inadvertently resemble or overlap with existing names, leading to trademark or copyright issues. Careful evaluation and validation processes will be required to ensure the uniqueness and originality of the generated names. However, the payoffs are significant, as the system can provide a valuable tool for individuals and businesses, saving them time and offering a fresh and diverse set of name ideas.

6. Out of pocket costs? Complete within 11 weeks?

The cost of the project will depend on the computational resources required for training and deploying the language model, as well as the effort required for data collection and preprocessing. The timeline for the project will involve several stages, including dataset preparation, model training, fine-tuning, and evaluation. The duration of the project will depend on the complexity of the selected models that I proposed and the size of the dataset. I am planning the model to train on different set of datasets which some of them in Azerbaijani and others in English. In case of the training on the names in Azerbaijani the computation will not be huge issue but if the model would be to train on the very large dataset to perform better then that would be bottleneck. It is estimated to build all the system, develop the proposed different models and train them on the dataset would fit in the duration of the course.

7. Midterm results?

The midterm evaluation will involve training and fine-tuning the language model on a dataset of existing names in Azerbaijan and if the timeline would be enough I am planning to also collect data of global company names and train to evaluate the system. This will ensure that the system is able to scale to different set of languages and domains. The generated names will be assessed for appropriateness, linguistic coherence, and uniqueness. Feedback from users can also be collected to further refine the model. The final evaluation will focus on the system's ability to generate names that are distinct from the existing names in the given context while maintaining relevance and linguistic coherence. More technically the system will be evaluated using a specific loss function (more about this in the next paragraph).

Evaluation of the Model(s)
If I would give brief background, in an autoregressive character-level language model, the input to the model is a sequence of characters, and the model's task is to predict the probability distribution over possible next characters. It does this by learning the statistical patterns and dependencies present in the training data.

So, in order to evaluate this system specific loss function will be implemented. In particular, the given dataset will be splitted to train, develop, and test sets using context length parameter. For instance, if the context length is set to 3 then it means we take 3 characters as input to predict the next one. Then cross-entropy loss function will be used to calculate the loss between the predicted probabilities and true labels that we created.

The cross-entropy loss function first applies a softmax activation to the logits (predicted by the model), which transforms them into predicted probabilities for each class. Then, it calculates the negative log-likelihood loss between the predicted probabilities and the target labels. The loss penalizes the model when its predicted probabilities distinguishes from the true labels, encouraging the model to adjust its parameters to improve its predictions. By doing so, the model will be able to learn and generate name-like outputs according to what it has been given.

8. Final Demonstration?

Finally, the system performance will be displayed in terms of its loss function and the sample outputs that it generates. Specifically, the automated pipeline will be developed to make this system to use easily on the different set of dataset (different domain or language). In simple terms, system will be such that the user will only submit the dataset (this can be names, words from same domain) and it will train this on different model to get best result to display to the user. The user will also have control over which model to use out of proposed models above.

Final Words

The development of advanced language models like GPT has revolutionized natural language processing tasks, including name generation. These models have the potential to generate highly creative and context-specific output. With each language’s unique linguistic characteristics, now is an opportune time to leverage these advancements and develop an AI-powered system that can generate new and appropriate names. The proposed research project aligns with the current state-of-the-art in language modeling and can make a significant impact in the field of name generation.