УДК 004.912

ГЕНЕРАЦИЯ ДОМЕННЫХ ИМЕН НА БАЗЕ НЕЙРОННОЙ СЕТИ

К. В. СЕРГУН, канд. техн. наук, доц. К. Я. РАХАНОВ
(Полоцкий государственный университет
имени Евфросинии Полоцкой, Беларусь)

Аннотация. *В докладе представлен метод генерации доменных имен с использованием модели GPT, изучены потенциальные источники ключевых слов, которые можно использовать при создании набора данных для обучения модели.*

Ключевые слова: *модель GPT, обучение, сбор данных.*

В связи с тем, что в Интернете насчитывается более 1,5 миллиарда веб-сайтов и зарегистрировано более 370 миллионов доменов [1], выбор уникального доменного имени для веб-ресурса может быть очень труден. Для того, чтобы получить идеальное и запоминающееся доменное имя, необходимо объединить два или три слова. Наряду с этим, следует учитывать некоторые советы, связанные с доменным именем:

– необходимо сделать доменное имя простым;

– лучше использовать короткое имя;

– доменное имя должно быть уникальным, так как оно поможет создать себе бренд;

– не стоит использовать какие-либо специальные символы, такие как цифры или дефисы. Допускается использовать ключевое слово, связанное веб-сайтом.

Запомнить все эти советы сложно каждому. Поэтому задача автоматической генерации уникальных доменных имен так актуальна. Самым простым методом ля выполнения данной задачи будет использование метода генерации случайных слов. При генерации случайных слов из заранее собранного словаря можно получить уникальное доменное имя, но это имя не будет говорить пользователю о том, что из себя представляет веб-ресурс. Чтобы этого избежать, можно отсеивать неподходящие по тематике ресурса слова путем ввода ключевых фраз, описывающих суть веб-сайта. Для реализации подобного лучше всего подходит использование нейронных сетей.

При выборе модели искусственного интеллекта необходимо учитывать различные критерии. К примеру, такие как: способность генерировать связные и информативные тексты, предобученность модели, ее доступность и уровень потребления ресурсов.

После анализа моделей искусственного интеллекта, таких как RNN, LSTM, GRU, GPT и BERT, была выбрана модель GPT. Данная модель специализируется

на генерации текста и может создавать связные и качественные тексты. Она хороша тем, что обладает обширными знаниями и лексическим разнообразием, так как была заранее обучена на огромном количестве текстов в сети Интернет. В связи с этим, для генерации доменных имен была выбрана модель GPT.

Обучение модели GPT

Обучение – это процесс предоставления больших объемов текстовых данных в модель на протяжении всего этапа обучения, чтобы помочь ей распознавать закономерности и связи между словами, фразами и предложениями в тексте. Модель использует алгоритмы глубокого обучения для распознавания закономерностей и корреляций между словами во время обучения, чтобы понимать и создавать язык, напоминающий человеческую речь. Обучение является важным шагом в разработке эффективных моделей обработки естественного языка, поскольку оно позволяет модели учиться на огромных объемах данных и повышать ее точность и эффективность при выполнении задач на основе НЛП, таких как языковой перевод, генерация текста и ответы на вопросы.

Процесс обучения модели GPT

Сбор данных. Первым шагом в обучении модели GPT является сбор большого количества текстовых данных. Эту информацию могут предоставить несколько источников, включая книги, журналы и веб-сайты. Чем больше и разнообразнее данные, тем лучше модель генерирует текст на естественном языке.

Для выполнения задачи генерации доменных имен необходимо собрать большой объем данных в виде имен доменов и ключевых слов, описывающих их. В качестве этих ключевых слов можно использовать данные, записанные в метатегах веб-страниц.

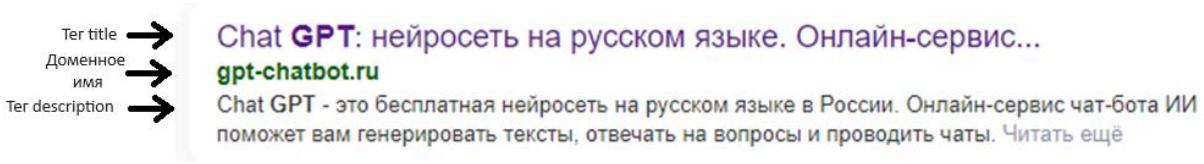


Рисунок 1. – отображаемая поисковым сервисом информация о ресурсе

Как видно, эти данные достаточно хорошо описывают суть веб-страниц, а также что доменные имена вполне соответствуют содержимой информации. Именно поэтому, для сбора данных для обучения модели, и были выбраны метатеги.

Метатеги, или мета-теги, это элементы HTML, которые предназначены для предоставления дополнительной информации о веб-странице. Метатеги не видны пользователям на странице, но используются поисковыми системами, браузерами и другими средствами для определения и отображения информации о содержимом страницы.

Для SEO-продвижения наибольшее значение имеют мета-теги title и description. Эти теги содержат информацию о сайте, с помощью которой поисковый робот может правильно определить тематику веб-ресурса и корректно ранжировать, то есть сортировать его в выдаче. При заполнении мета-тегов используются ключевые слова и фразы, которые набирает в поисковой строке целевой пользователь.

Тег Title. Мета-тег title представляет собой название или заголовок веб-страницы. Его рекомендуется указывать на каждой странице сайта. В поисковой выдаче title представлен как крупная синяя ссылка и отображается как название вкладки браузера в момент открытия страницы. В случае отсутствия тега title, могут быть использованы URL-адрес страницы или заголовок h1.

```
<title>Chat GPT: нейросеть на русском языке. Онлайн-сервис в России</title>
```

Рисунок 2. – данные из метатега Title

Тег Description. Тег description служит для предоставления краткого описания содержания веб-страницы и является логическим дополнением к title. Поисковые системы используют информацию из этого мета-тега для формирования сниппета – небольшого текстового блока, который дает представление о содержании страницы в результатах поиска. Description должен содержать несколько коротких предложений с ключевыми словами и обычно ограничивается от 100 до 200 символов. Как и title, мета-тег description должен быть уникальным для каждой страницы сайта.

```
<meta name="description" content="Chat GPT - это бесплатная нейросеть на русском языке в России. Онлайн-сервис чат-бота ИИ поможет вам
генерировать тексты, отвечать на вопросы и проводить чаты">
```

Рисунок 3. – данные из метатега description

Тег Keywords. Мета-тег keywords предназначен для указания списка ключевых слов, связанных с конкретной страницей. В прошлом он играл важную роль в оптимизации для поисковых систем, но в настоящее время его значение снизилось. Google, в частности, официально объявил в 2009 году, что не учитывает ключевые слова из этого тега при ранжировании страниц. Как следствие, разработчики уделяют больше внимания тегам Title и Description, заполняя тег keywords только по желанию.

```
<head>
 <meta name="keywords" content="ключевое слово 1, ключевое слово 2, ключевое слово 3"
</head>
```

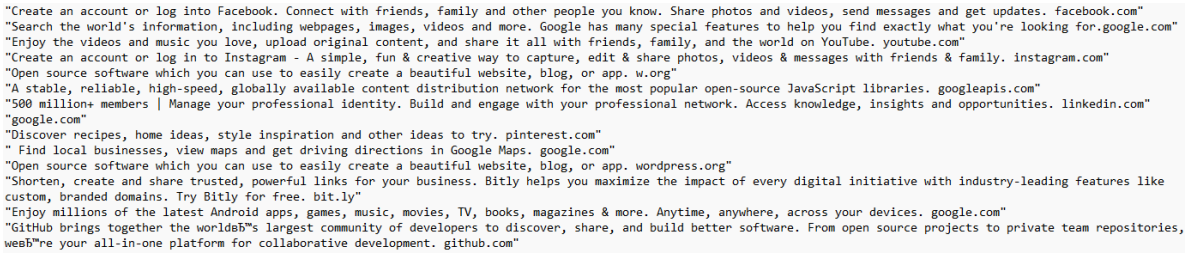
Рисунок 4. – данные из метатега keywords

В итоге, в качестве данных для обучения могут быть выбраны данные из тегов «decription» и «keywords», но ввиду того, что тег «keywords» потерял свою популярность и эффективность, разработчики почти полностью перестали им пользоваться. Из-за этого собрать достаточно данных из тега «keywords» будет затруднительно.

Обучение моделей GPT на данных из метатегов обосновывается большим объемом и разнообразием информации, эти данные могут сохранять различные типы информации, что помогает моделям GPT изучать широкий спектр языковых структур и связей. Это позволит моделям обучаться в разных стилях текста и тематике. Данные из метатегов могут включать в себя информацию из различных областей знаний, что позволяет моделировать прогресс своих контекстуальную адаптацию при генерации текстов. Метатеги могут рассчитывать контекстную разметку для данных, что помогает моделям лучше понимать связи и отношения между различными текстами. Это соглашение поддерживает возможности моделей по созданию смысловой целостности текстов.

Таким образом, выбор данных из метатегов для обучения моделей GPT обусловил стремление к расширению моделей знаний, продолжению ее способностей к обобщению и контекстуализации, а также обеспечению разнообразной и информативной основы для обучения.

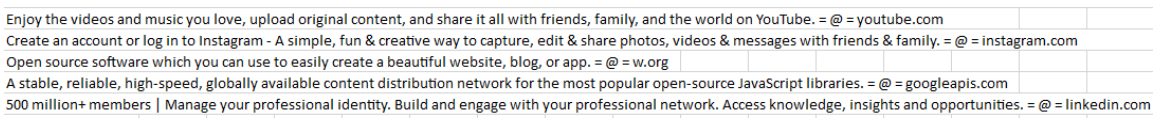
Вручную процесс сбора данных для обучения будет очень трудоемким и времязатратным, поэтому необходимо написать парсер, который сможет собрать необходимую информацию с большого количества веб-сайтов.



"Create an account or log into Facebook. Connect with friends, family and other people you know. Share photos and videos, send messages and get updates. facebook.com"
"Search the world's information, including webpages, images, videos and more. Google has many special features to help you find exactly what you're looking for.google.com"
"Enjoy the videos and music you love, upload original content, and share it all with friends, family, and the world on YouTube. youtube.com"
"Create an account or log in to Instagram - A simple, fun & creative way to capture, edit & share photos, videos & messages with friends & family. instagram.com"
"Open source software which you can use to easily create a beautiful website, blog, or app. w.org"
"A stable, reliable, high-speed, globally available content distribution network for the most popular open-source JavaScript libraries. googleapis.com"
"500 million+ members | Manage your professional identity. Build and engage with your professional network. Access knowledge, insights and opportunities. linkedin.com"
"google.com"
"Discover recipes, home ideas, style inspiration and other ideas to try. pinterest.com"
" Find local businesses, view maps and get driving directions in Google Maps. google.com"
"Open source software which you can use to easily create a beautiful website, blog, or app. wordpress.org"
"Shorten, create and share trusted, powerful links for your business. Bitly helps you maximize the impact of every digital initiative with industry-leading features like custom, branded domains. Try Bitly for free. bit.ly"
"Enjoy millions of the latest Android apps, games, music, movies, TV, books, magazines & more. Anytime, anywhere, across your devices. google.com"
"GitHub brings together the worldвЂ™s largest community of developers to discover, share, and build better software. From open source projects to private team repositories, wevвЂ™re your all-in-one platform for collaborative development. github.com"

Рисунок 5. – Пример собранных данных для обучения модели

Очистка и предварительная обработка данных. После сбора данных их необходимо подготовить путем очистки и предварительной обработки. Для этого удаляются посторонние данные, включая элементы HTML, знаки препинания и специальные символы, текст приводится к виду, удобному для работы моделью GPT. Кроме того, для упрощения данных они делятся на управляемые фрагменты, такие как слова или подслова.



Enjoy the videos and music you love, upload original content, and share it all with friends, family, and the world on YouTube. = @ = youtube.com
Create an account or log in to Instagram - A simple, fun & creative way to capture, edit & share photos, videos & messages with friends & family. = @ = instagram.com
Open source software which you can use to easily create a beautiful website, blog, or app. = @ = w.org
A stable, reliable, high-speed, globally available content distribution network for the most popular open-source JavaScript libraries. = @ = googleapis.com
500 million+ members | Manage your professional identity. Build and engage with your professional network. Access knowledge, insights and opportunities. = @ = linkedin.com

Рисунок 6. – Пример обработанных данных

По окончанию процесса обучения модели GPT, с использованием ее внутренних методов (таких как generate), можно протестировать ее работу сгенерировав доменное имя, введя максимально полное и точное описание. Путем настройки параметров метода generate (length, temperature, truncate, include_prefix, batch_size) можно добиться более качественного результата.

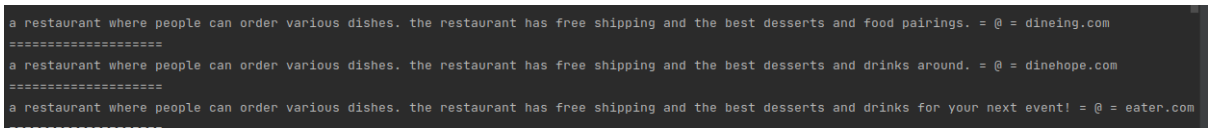


Рисунок 7. – Пример генерации доменного имени

Вывод. В результате проведенной работы по генерации доменных имен с использованием нейронной сети стало ясно, что модель GPT может показывать хорошие результаты в генерации. Благодаря модели, могут быть получены уникальные и оригинальные доменные имени. Качество результатов напрямую будет зависеть от количества и качества данных для обучения модели. Также, важным фактором будет являться контроль качества результата с целью дальнейшего улучшения путем возможной донастройки и дообучения модели.

ЛИТЕРАТУРА

1. Sahoo M. 9 Best Domain Name Generator Tools to Discover a Domain. – URL: https://medium.com/knoansw/9-best-domain-name-generator-tools-to-discover-a-domain-4c1daf87a7a. – 16.06.2021.
2. GPT Books // GPT Architecture. – https://whoisdsmith.gitbook.io/gpt-books/gpt-books/customgpt-training/gpt-architecture. – 05.2023.
3. Wikipedia // Generative pre-trained transformer. – https://ru.wikipedia.org/wiki/Generative_pre-trained_transformer. – 2024.
4. OpenAI // ChatGPT – https://chat-gpt.org/. – 2024.
5. Karkhane P. – Unveiling Language Model Architectures: RNN, LSTM, GRU, GPT, and BERT. – https://medium.com/@kpradyumna/unveiling-language-model-architectures-rnn-lstm-gru-gpt-and-bert-c9efdf4eb8cc. – 19.07.2023.
6. Leonard // From RNN to ChatGPT – https://wzdlc1996.github.io/artic/datascience/rnnchatgpt/. – 03.2023.
7. Shree P. – The Journey of Open AI GPT models – https://medium.com/walmartglobaltech/the-journey-of-open-ai-gpt-models-32d95b7b7fb2. – 10.09.2020.