# Реализация поисковой системы **emoji** по текстовому описанию

Проектная работа по курсу NLP

МИСИС
УНИВЕРСИТЕТ
НАУКИ И ТЕХНОЛОГИЙ

2024

# Описание **проекта**

**Цель:** упростить поиск emoji для различных текстовых материалов, что позволит повысить наглядность и эмоциональную насыщенность представляемой информации.

## Варианты использования:

- Для презентаций.
- Для интернет-статей.
- Для проектов (оформление README.md, коммитов, merege/pull requests, issues, документации).
- Для CLI (оформление выводимой информации, вкладки помощи).

```
~ minikube start
😄  minikube v1.33.1 on Arch 24.0.8
🎉  minikube 1.34.0 is available! Download it: https://github.com/kubernetes/minikube/releases/tag/v1.34.0
💡  To disable this notice, run: 'minikube config set WantUpdateNotification false'

✨  Automatically selected the docker driver. Other choices: none, ssh
📌  Using Docker driver with root privileges
👍  Starting "minikube" primary control-plane node in "minikube" cluster
🚜  Pulling base image v0.0.44 ...
💾  Downloading Kubernetes v1.30.0 preload ..
```

https://towardsdatascience.com/semantic-search-for-emojis-in-50-languages-using-ai-f85a36a86f21

# **Семантическая** поисковая система

— это система, которая осуществляет поиск информации на основе понимания смысла запросов пользователей и содержания документов.

1. Получение текстового описания.
2. При помощи **Language Model** текстовое описание преобразуется в векторное представление (эмбеддинг).
3. На основе полученного эмбеддинга осуществляется поиск в векторной базе данных.
4. Возвращене результата поиска.

**LM** (**DistilBERT** or **MiniLM**) for obtaining embeddings

**Vector search engine** (based on KNN of ANN)

https://clickhouse.com/blog/vector-search-clickhouse-p1, https://www.elastic.co/what-is/semantic-search, https://huggingface.co/docs/transformers/model_doc/distilbert, https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

# Датасеты

Готовых наборов данных для решения этой задачи практически нет. Однако их можно сгенерировать с помощью **LLM**.

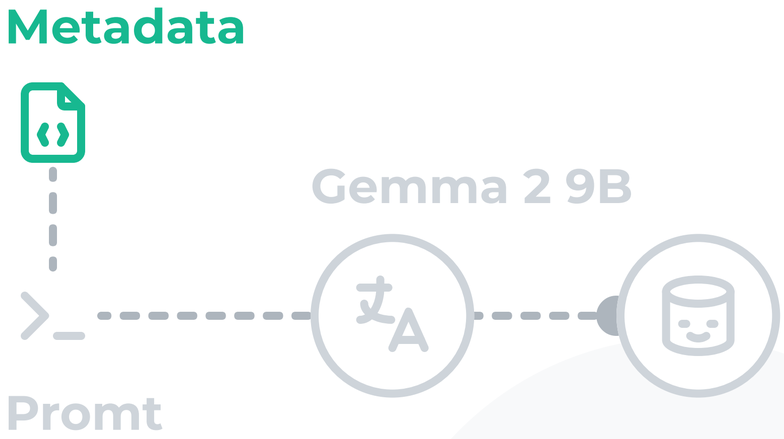
- **Emoji Metadata Dataset** созданный при помощи **Llama3**.
- **Gitmoji** — это инициатива по стандартизации и объяснению использования emoji в сообщениях коммитов GitHub.


https://huggingface.co/datasets/badrex/llm-emoji-dataset, https://gitmoji.dev/, https://www.kaggle.com/datasets/rtatman/emojinet/data

# **Создание датасета** | Сбор метаданных emoji

## Источники

- Emoji for Python — список emoji и их короткие коды.
- Gitmoji — описание в контексте git, тег из спецификации семантического версионирования (MAJOR, MINOR, PATCH).
- Unicode CLDR Project — ключевые слова и информация для Text-to-Speech систем.

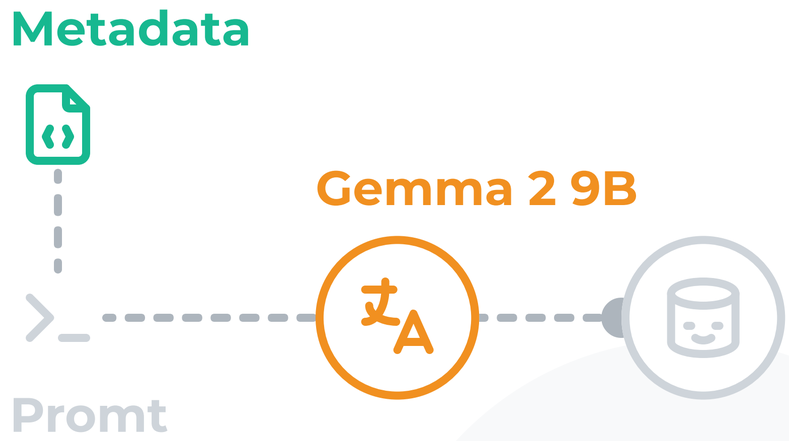



https://gitmoji.dev/, https://semver.org/, https://cldr.unicode.org/,

# **Создание датасета** | Выбор модели

Для создания собственного датасета была выбрана модель **Gemma 2 9B**. Она предоставляет более точность в обработке естественного языка по сравнению с **Llama 3**.

| | | | Gemma 2 | | Llama 3 | | Grok-1 |
|---|---|---|---|---|---|---|---|
| | BENCHMARK | METRIC | 9B | 27B | 8B | 70B | 314B |
| General | MMLU | 5-shot, top-1 | 71.3 | 75.2 | 66.6 | 79.5 | 73.0 |
| Reasoning | BBH | 3-shot, CoT | 68.2 | 74.9 | 61.1 | 81.3 | – |
| | HellaSwag | 10-shot | 81.9 | 86.4 | 82 | – | – |
| Math | GSM8K | 5-shot, maj@1 | 68.6 | 74.0 | 45.7 | – | 62.9 (8-shot) |
| | MATH | 4-shot | 36.6 | 42.3 | – | – | 23.9 |
| Code | HumanEval | pass@1 | 40.2 | 51.8 | – | – | 63.2 (0-shot) |



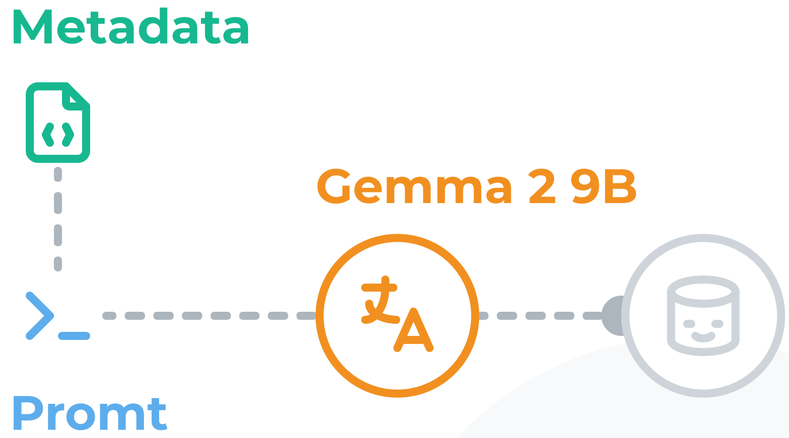

https://blog.google/technology/developers/google-gemma-2/, https://openrouter.ai/google/gemma-2-9b-it:free

# Создание датасета | Промпт-инжиниринг

```
[emoji-zap]
emoji = "⚡"
shortcodes = [":zap:"]
keywords = ["danger", "electric", "electricity", "high", "lightning", "nature", "thunder", "thunderbolt", "voltage", "zap"]
text-to-speech = ["high voltage"]
git-description = "Improve performance."
git-semver = "patch"

[emoji-zany-face]
emoji = "🤪"
shortcodes = [":zany_face:"]
keywords = ["crazy", "eye", "eyes", "face", "goofy", "large", "small", "zany"]
text-to-speech = ["zany face"]
git-description = ""
git-semver = ""
```

Входные данные




```
⚡
Symbolizes electricity and high energy. This emoji is often used to express excitement, danger, or a sudden burst of power...
["electricity", "performance", "danger", "high voltage", "energy", "lightning", "improvement", "excitement", "thunderbolt"]

🤪
Represents a playful or goofy expression, often associated with silliness and fun. This emoji is commonly used to convey...
["playful", "silly", "crazy", "goofy", "fun", "zany", "humor", "light-hearted", "expression"]
```

Ожидаемый результат

# **Создание датасета** | Примеры

🥇

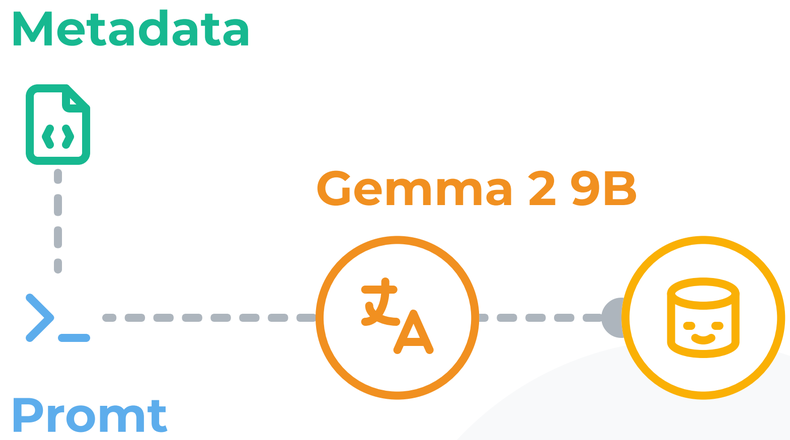
**Ключевые слова:** win, victory, achievement, first place, gold medal, success, excellent, milestone, top.

**Описание:** Symbolic of winning first place, representing achievement, victory, and excellence. This emoji is often used to celebrate successes and highlight top accomplishments. In Git, it can signify achieving a significant milestone, resolving a challenging issue, or implementing a groundbreaking feature.

🍁

**Ключевые слова:** autumn, fall, leaf, maple, nature, season, red, orange, yellow.

**Описание:** Represents a maple leaf, often associated with autumn, fall, and nature. This emoji symbolizes the changing seasons and the beauty of nature's transformation.



# Базовое решение

В качестве базового решения был реализован поиск на основе **TF-IDF** векторных представлений.

"cat smiling": 😻,🐱,😇,😽,😺

"remove code or files": 🗑️,🔥,✖️,✒️,0️⃣

"walking the dog in rainy weather": ☔,🐕,🚶‍♀️,🚶,🚶‍♀️

Спасибо за **Внимание**!

2024