Реализация семантической поисковой системы **emoji**

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).

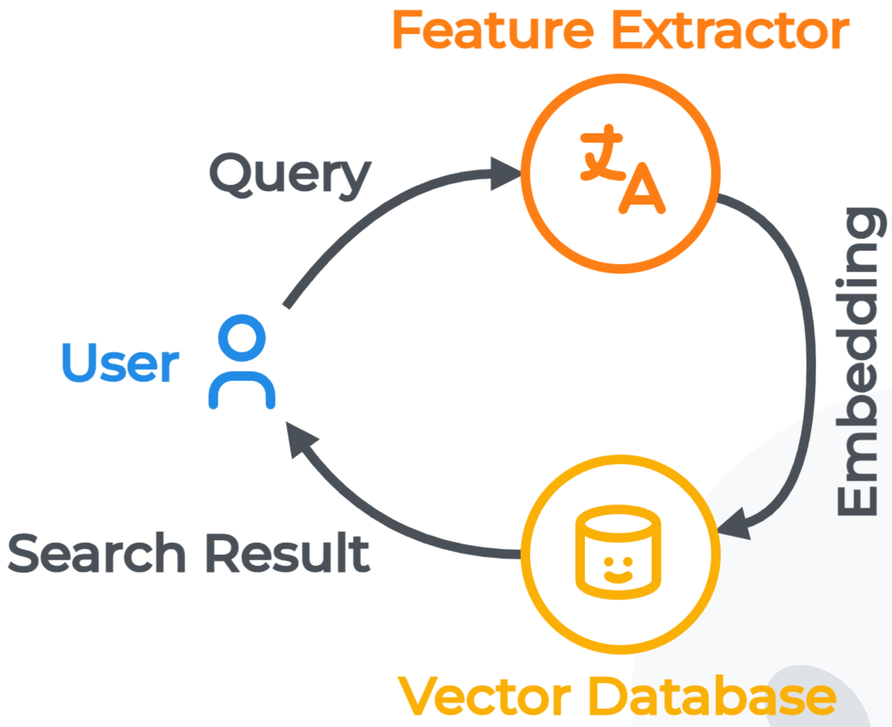
```
~ minikube start
😄  minikube v1.34.0 on Arch 24.1.2
🆕  Kubernetes 1.31.0 is now available. If you would like to upgrade ...
✨  Using the docker driver based on existing profile
👍  Starting "minikube" primary control-plane node in "minikube" ...
🚜  Pulling base image v0.0.45 ...
🔄  Restarting existing docker container for "minikube" ...
❗  Image was not built for the current minikube version.
🐳  Preparing Kubernetes v1.30.0 on Docker 26.1.1 ...
🔎  Verifying Kubernetes components...
    ▪ Using image gcr.io/k8s-minikube/storage-provisioner:v5
🌟  Enabled addons: storage-provisioner, default-storageclass
💡  kubectl not found. If you need it, try: 'minikube kubectl -- get pods -A'
🏄  Done! kubectl is now configured to use "minikube" cluster ...
```

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 запроса от пользователя.
2. Преобразование запроса в векторное представление при помощи **экстрактора признаков**.
3. Поиск k наиболее релевантных ответа в **векторной базе данных** на основе расстояния между векторами.
4. Возврат результата поиска пользователю.



https://www.elastic.co/what-is/semantic-search

Датасеты

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

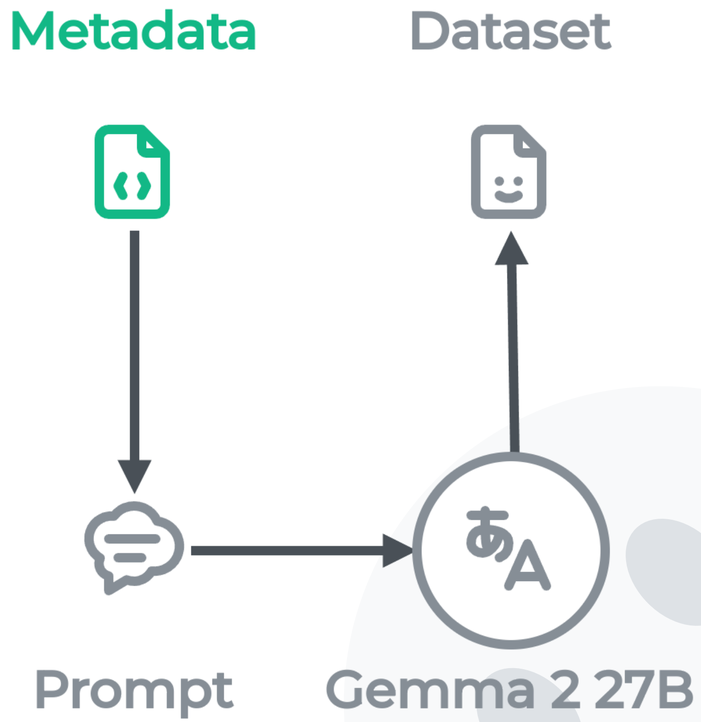
План

1. Создать датасет содержащий базовые метаданные из различных источников.
2. Выбрать **LLM** для генерации семантических метаданных (описание и ключевые слова).
3. Определится форматом входных и выходных данных, создать промпт для **LLM**.
4. Сгенерировать датасет.

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

Генерация датасета | Сбор метаданных

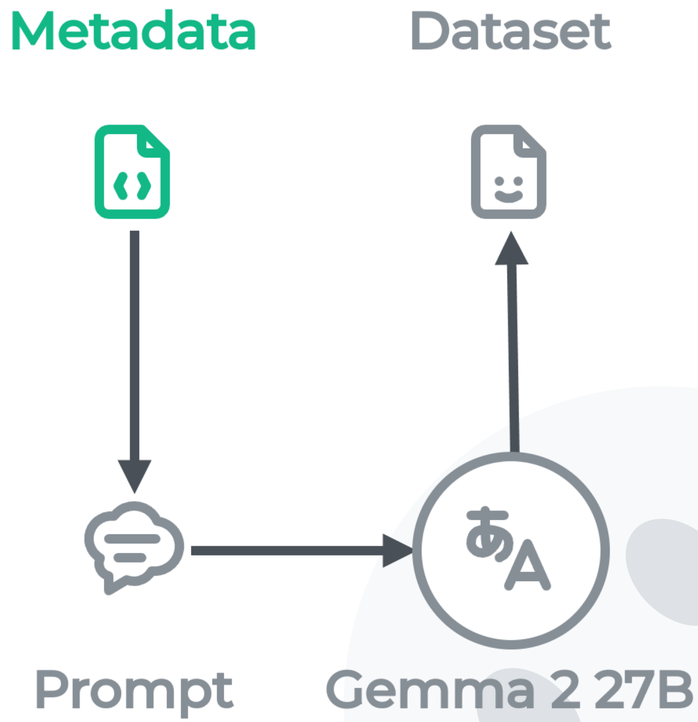
- **Emoji for Python** — список emoji и их короткие коды.
- **Gitmoji** — описание в контексте git, тег из спецификации семантического версионирования (MAJOR, MINOR, PATCH).
- **Unicode CLDR Project** — ключевые слова и информация для Text-to-Speech систем.



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

Генерация датасета | Сбор метаданных

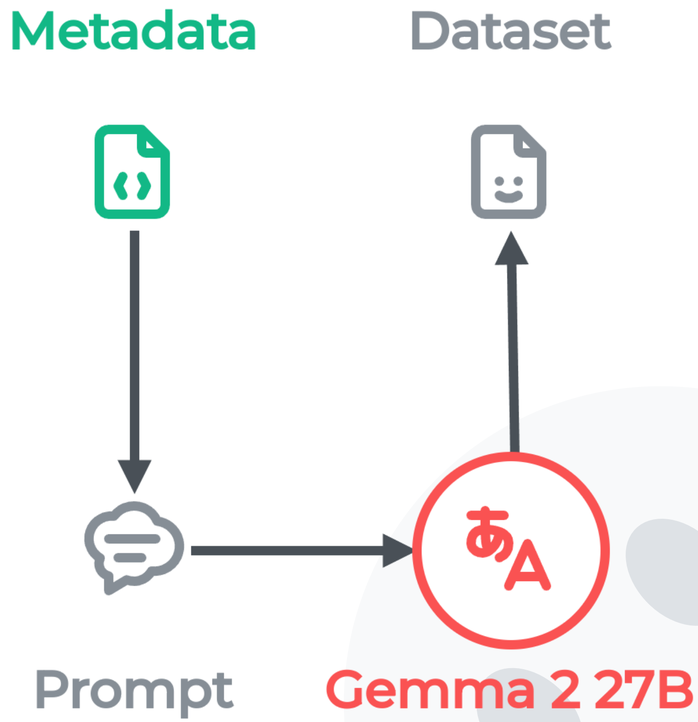
emoji	shortcodes	keywords	tts	git-description	git-semver
🦺	[':safety_vest:']	['emergency' 'safety' 'vest']	['safety vest']	Add or update code related to validation.	
♻️	[':recycle:']			Refactor code.	
📄	[':page_facing_up:']	['document' 'facing' 'page' 'paper' 'up']	['page facing up']	Add or update license.	
📸	[':camera_flash:']	['camera' 'flash' 'video']	['camera with flash']	Add or update snapshots.	
🍳	[':egg:' ':fried_egg:']	['breakfast' 'cooking' 'easy' 'egg' 'fry' 'frying' 'over' 'pan' 'restaurant' 'side' 'sunny' 'up']	['cooking']	Add or update an easter egg.	patch
🔀	[':twisted_rightwards_arrows:']	['arrow' 'button' 'crossed' 'shuffle' 'tracks']	['shuffle tracks button']	Merge branches.	
🔧	[':wrench:']	['home' 'improvement' 'spanner' 'tool' 'wrench']	['wrench']	Add or update configuration files.	patch
🙈	[':see_no_evil:' ':see_no_evil_monkey:']	['embarrassed' 'evil' 'face' 'forbidden' 'forgot' 'gesture' 'hide' 'monkey' 'no' 'omg' 'prohibited' 'scared' 'secret' 'smh' 'watch']	['see-no-evil monkey']	Add or update a .gitignore file.	
🗑️	[':wastebasket:']			Deprecate code that needs to be cleaned up.	patch
💄	[':lipstick:']	['cosmetics' 'date' 'lipstick' 'makeup']	['lipstick']	Add or update the UI and style files.	patch



Генерация датасета | Выбор LLM

Для создания датасета была выбрана модель **Gemma 2 27B**. Она обладает небольшим размером и предоставляет неплохую точность по сравненю с более большими моделями.

			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-2-27b-it

Генерация датасета | Создание промпта

Входные данные

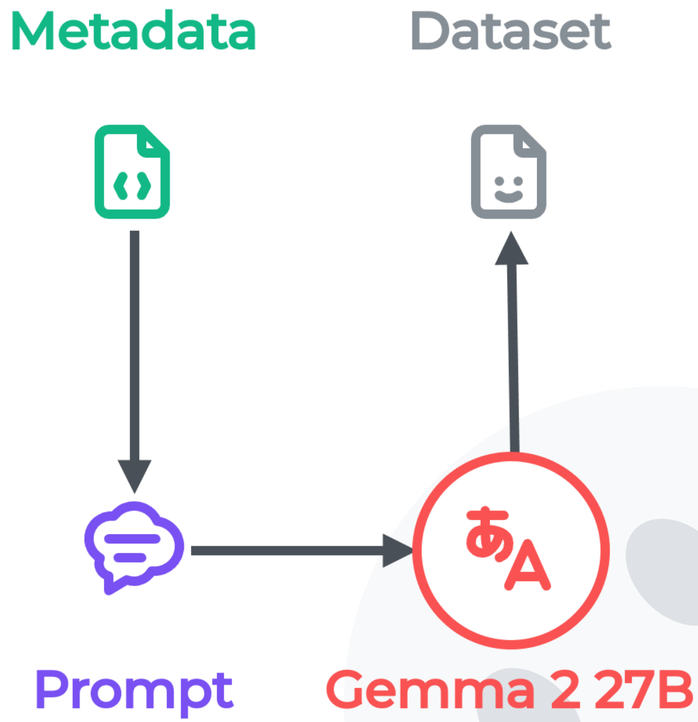
```
[emoji-whale]
emoji = "🐳"
shortcodes = [":whale:"]
keywords = ["animal", "beach", "face", "ocean", "spouting", "whale"]
description = ["spouting whale"]
```

Ожидаемый результат

emoji-whale

The spouting whale emoji symbolizes the majesty and vastness of the ocean. It is frequently used in discussions about marine life, conservation, and the beauty of nature, and can indicate significant environmental updates in a git commit message. This emoji can also express a sense of wonder or awe, reflecting the grandeur of the natural world.

["marine", "ocean", "animal", "beach", "whale", "spouting", "nature", "aquatic", " wildlife", "ecosystem", "conservation", "water", "sea", "symbolism", " environment"]



Генерация датасета | Результат

Среднее количество ключевых слов: **12**.

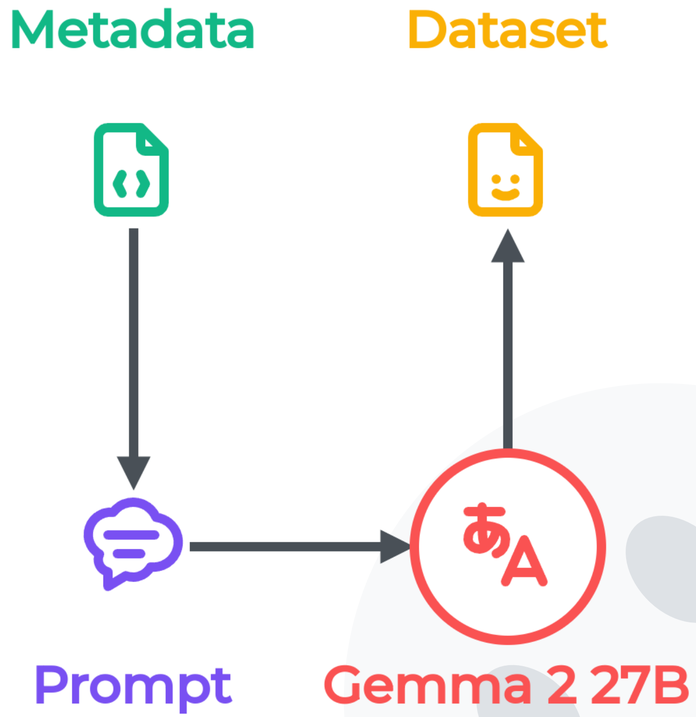
Среднее количество символов в описании: **250**.

Пример

🤔

Описание: This emoji depicts a face deep in thought, conveying contemplation, pondering, or seeking solutions. It can be used to express consideration of ideas, confusion, or simply taking time to think before responding.

Ключевые слова: analyzing, brainstorming, confused, considering, contemplation, decision making, meditation, pondering, problem solving, questioning, reflecting, thinking, thinking face.



Поисковая система v1

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

Плюсы

- Очень прост в реализации.
- Быстрая скорость работы.

Минусы

- Большой размер векторов.
- Поддерживает поиск только точному совпадению слов.

"cat smiling": 😻,🐱,😇,😽,😼
"remove code or files": 🗑️,🔥,✖️,✒️,0️⃣
"walking the dog in rainy weather": ☔,🐕,🚶‍♀️,🚶,🚶‍♀️

Поисковая система v2

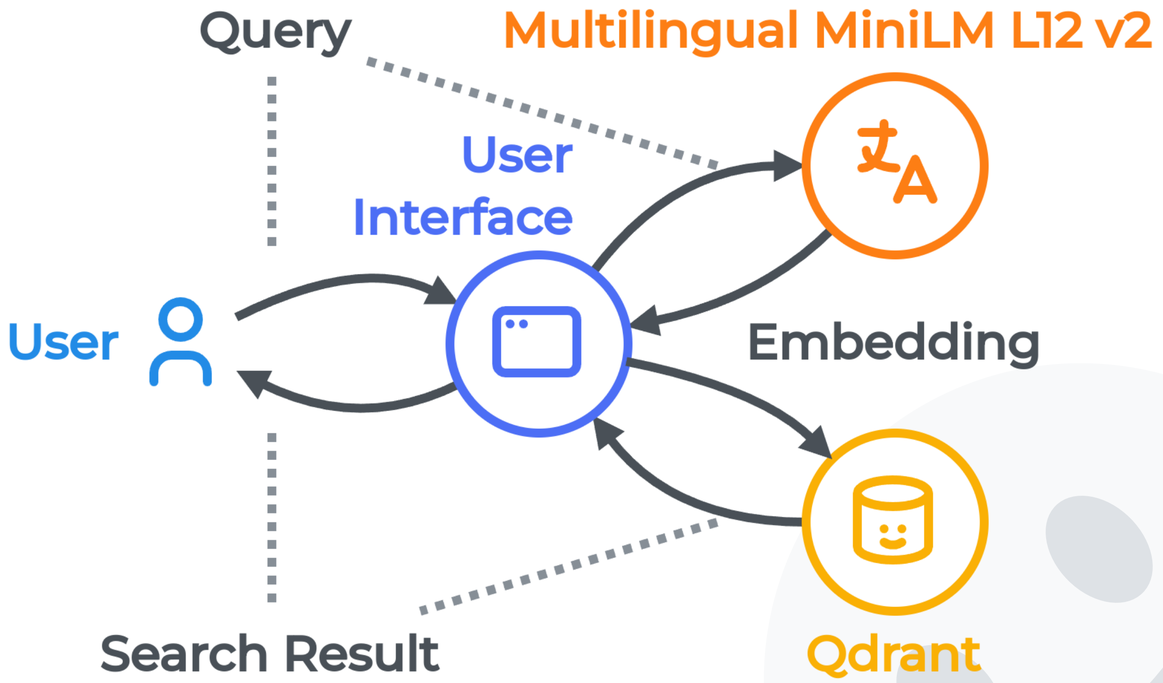
Итоговое решение семантической поисковой системы реализованно на основе **Multilingual MiniLM L12 v2**.

Плюсы

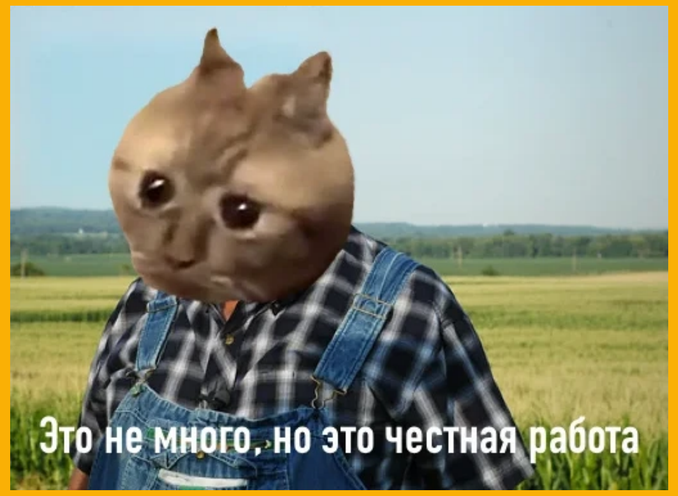
- Учитывает семантику запроса.
- Позволяет производить поиск на 50 различных языках.

Минусы

- Сложность реализации.
- Более долгая скорость работы.



Спасибо за **Внимание**!



2024