台智雲福爾摩沙大模型 FFM

Embedding API 說明文件

Generation usage for curl

1. 設定環境

```
export API_KEY=00000000-0000-0000-0000-000000000000
export API_SERVER=https://ffm-trial05.twcc.ai/embeddings/api
```

2. 透過以下的 curl 指令取得 embedding 結果

使用範例(non-batch)

```
curl "${API_SERVER}/embeddings" \
  -H "X-API-KEY:${API_KEY}" \
  -H "content-type: application/json" \
  -d '{"input": ["search string 1"]}'
```

回傳範例


```
{
 "data": [
   {
     "embedding": [
       0.06317982822656631,
       -0.5447818636894226,
       -0.3353637158870697,
       -0.5117015838623047,
       -0.1446804255247116,
       0.2036416381597519,
       -0.20317679643630981,
       -0.9627353549003601,
       0.31771183013916016,
       0.23493929207324982,
       0.18029260635375977,
       ...
       ...
     ],
     "index": 0,
```

```
        "object": "embedding"
    }
 ],
 "object": "list"
}
```

使用範例(batch)

```
curl "${API_SERVER}/embeddings" \
   -H "X-API-KEY:${API_KEY}" \
   -H "content-type: application/json" \
   -d '{"input": ["search string 1", "search string 2"]}'
```

回傳範例

```
{
 "data": [
    {
       "embedding": [
          0.06317982822656631,
          -0.5447818636894226,
          -0.3353637158870697,
          -0.5117015838623047,
          -0.1446804255247116,
          0.2036416381597519,
          -0.20317679643630981,
          -0.9627353549003601,
          0.31771183013916016,
          0.23493929207324982,
          0.18029260635375977,
          ...
          ...
       ],
       "index": 0,
       "object": "embedding"
    },
    {
       "embedding": [
          0.15340591967105865,
          -0.26574525237083435,
          -0.3885045349597931,
```

```
            -0.2985926568508148,
            0.22742436826229095,
            -0.42115798592567444,
            -0.10134009271860123,
            -1.0426620244979858,
            0.507709264755249,
            -0.3479543924331665,
            -0.09303411841392517,
            1.0853372812271118,
            0.7396582961082458,
            0.266722172498703,
            ...
            ...
        ],
        "index": 1,
        "object": "embedding"
    }
 ],
 "object": "list"
}
```


額外說明

1. non-batch : 一個 request 中只包含一筆 input data
2. batch : 一個 request 中包含多筆 input data
3. input data 限制 ： 每一筆 input data 長度上限為 2048 tokens
4. total data 限制 ： 一個 request 包含的 total input data 上限為 32000 tokens