
3. CLOVA Studio - 튜닝

작성자 : Master seo
topasvga@naver.com

튜닝

NO	설명
1	기업의 데이터셋으로 특화 AI 언어 모델 구축 기업의 데이터셋 기업 서비스에 최적화된 AI 모델 특정 Domain에 맞게 미세 조정 가능
2	튜닝 = 기업의 데이터 셋으로 특화 AI 모델 구축 = 기업의 데이터셋으로 추가 학습 시킴. 추가 학습이다. 기업 데이터만 사용하는게 아니다!! 기존 데이터와 합쳐져서 결과가 나오는 것이다.
3	계정의 액세스키, 시크릿키를 사용한다. hcx-003으로 변경 튜닝을 하면 엔진이 변경된다~ test model 이름도 변경됨

튜닝의 필요성

NO	설명
1	플레이 그라운드에서 작업 시 토큰 제약이 있으나 튜닝 이용 시 토큰 제약 없이 더욱 다양한 예제 구성 가능
2	원하는 방식으로 데이터셋을 구성하여, 니즈에 적합한 방향으로 출력 가능
3	모델이 가벼워지며 태스크의 수행 시간이 단축되고 비용 절감이 가능

튜닝 - 작업 종류

사용자가 원하는 답을 얻을수 있도록 사용자 데이터셋을 학습시킨다.
정의된 작업 종류를 제공한다.

예) 튜닝 > 대화 > 생성하기 > LK-D2 > 생성

The screenshot shows the '튜닝' (Tuning) section of the CLOVA Studio interface. The '튜닝' menu item is highlighted in the top navigation bar. Below it, the '새 작업 생성하기' (Create New Job) section is visible, with a sub-header '튜닝을 통해 사용자 데이터에 최적화된 모델을 만들고 테스트해볼 수 있습니다. 튜닝을 진행할 작업의 종류를 선택해주세요.' (You can create and test a model optimized for user data using tuning. Please select the type of job to proceed with tuning.)

There are six job type cards displayed in a grid:

- 튜닝 API (HyperCLOVA X)**: API를 호출하여 직접 데이터를 학습합니다. 학습 파라미터를 조정하여 최적화된 학습을 수행할 수 있습니다. (Learn by calling the API and training directly. You can adjust learning parameters to perform optimized learning.)
- 문서 분류 (Binary Classification)**: 이진 분류 (Binary Classification) 방식으로 데이터를 학습합니다. 분류 기준이 두 가지인 경우 (예, 긍정, 부정) (Learn data in a binary classification manner. In the case of two classification criteria (e.g., positive, negative))
- 문서 분류 (Multi-class Classification)**: 다중 분류 (Multi-class Classification) 방식으로 데이터를 학습합니다. 분류 기준이 세 가지 이상인 경우 (예, 기쁨, 슬픔, 분노) (Learn data in a multi-class classification manner. In the case of three or more classification criteria (e.g., happiness, sadness, anger))
- 문장 요약**: 다양한 텍스트를 사용자가 원하는 기준에 따라 자동으로 요약합니다. (Summarize various texts automatically according to the criteria you want.)
- 문장 생성**: 사용자 데이터에 기반하여, 주제에 적합한 다양한 텍스트를 생성합니다. (Generate various texts suitable for the topic based on user data.)
- 문체 전환**: 다양한 텍스트를 사용자가 원하는 형식 및 문체로 전환합니다. (Convert various texts into the format and style you want.)

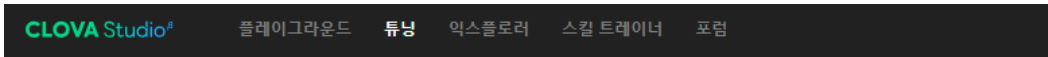
At the bottom right, there is a '대화' (Conversation) job type card. It is highlighted with a red box. The '모델 엔진' (Model Engine) dropdown menu is also highlighted with a red box and set to 'LK-D2'. Below the card, there are '취소' (Cancel) and '생성' (Generate) buttons, with the '생성' button also highlighted with a red box.

튜닝 - 작업 화면

유니코드(UTF-8) 파일 이어야 한다.

Excel을 그냥 저장하면 유니코드(UTF-8) 파일이 아니다.

Text 와 Completion 형식이다. <https://brunch.co.kr/@topasvga/3764>
최소 400건 이상 만들어야 한다.



작업 이름을 정확히 입력해주세요.

데이터셋 업로드

데이터셋 규격 및 주의사항에 유의하여 데이터셋을 업로드해주세요. [포맷 다운로드](#)

TIP

데이터셋 규격

- 데이터셋 확장자는 `.csv` 혹은 `.jsonl` 만 지원합니다. 인코딩은 UTF-8 형식을 준수해주세요.
- **파일은 최소 400건 이상의 데이터를 권장합니다.**
- `.csv` 파일의 경우 데이터셋 포맷 파일 **활용**을 권장하며, 포맷 파일과 상이한 규격의 파일은 업로드 불가능합니다.

주의사항

- 파일의 한 행(Text, Completion 쌍)당 공백 포함 총 4,000 자 이하여야 합니다. 초과 시, 데이터셋의 일부만 업로드 됩니다.
- Text에는 입력값을, Completion에는 원하는 결과값을 입력해주세요.
- 문서 분류 작업의 경우, 각 분류 범주 당 최소 200건 이상의 데이터를 권장합니다.
- 다중 분류(Multi-class classification)의 경우, 분류 라벨을 최대 16개까지 입력 가능합니다.
- 또한, 분류 라벨은 띄어쓰기와 특수 문자를 제외한 단어 한 개로 구성하기를 권장합니다.
- 개인정보가 포함된 데이터셋을 업로드하여 발생하는 모든 문제와, 이로 인하여 발생한 결과에 대한 모든 책임은 사용자에게 있습니다.

데이터셋 업로드

 **파일 업로드**
파일을 클릭해서 업로드해주세요.

튜닝 - 작업 화면

유니코드(UTF-8) 파일 이어야 한다.

Excel을 그냥 저장하면 유니코드(UTF-8) 파일이 아니다.

Text 와 Completion 형식이다. <https://brunch.co.kr/@topasvga/3764>

최소 400건 이상 만들어야 한다.

토큰 산정

×

작업 이름 튜닝1

학습 데이터셋 튜닝테스트파일_conv (3).csv 60.6KB

예상 사용 토큰

12075 토큰 × 8 epoch

요청하신 작업 수행에는 위와 같은 토큰이 사용될 예정이며 진행한 학습 epoch 만큼 배수되어 과금됩니다.
예상치이므로, 실제 사용되는 토큰 수와 약간의 차이가 있을 수 있습니다.

취소

학습

튜닝 - 작업 화면

유니코드(UTF-8) 파일 이어야 한다.

Excel을 그냥 저장하면 유니코드(UTF-8) 파일이 아니다.

Text 와 Completion 형식이다. <https://brunch.co.kr/@topasvga/3764>

최소 400건 이상 만들어야 한다.

오른쪽위 내 계정 이름 클릭 > 내작업에서 확인 가능하다.

튜닝 시작과 종료시 메일 계정으로 시작과 종료 메일이 온다.

위 예제는 20분 걸린다.

내 작업

플레이그라운드 튜닝 익스플로러

작업 상태	작업 이름	작업 종류	클라이언트	모델 엔진 ①
학습 대기 중	튜닝1*	대화	웹	LK-D2
학습 완료	test12	생성	API	HGX-003
학습 완료	222	대화	웹	LK-D2

튜닝 - 테스트 예시

오른쪽위 > 내작업 > 튜닝에서 확인한다. 학습 완료 클릭
튜닝 데이터셋과 '동일한 포맷으로 질의' 작성 테스트
예)

어르신 : 독감주사 맞아야하는데 어디로 가야하나?

실행

참고

<https://brunch.co.kr/@topasvga/3764>

튜닝1

Workflow ID	tuning-1823-240805-152920-zfpml
Problem Type	대화
Model Engine	LK-D2
Dataset	튜닝테스트파일_conv (3).csv, 12075 of tokens
Train Loss	0.8876
Tokens Used	94,920

Inference Test

플레이그라운드에서는 파라미터를 조절하여 테스트할 수 있습니다. [플레이그라운드](#)를 이용하여 Inference test를 진행해주세요.

Input

어르신 : 독감주사 맞아야하는데 어디로 가야하나?

실행

Output

AI: 가까운 보건소로 가시면 됩니다.

튜닝 - 테스트 앱 생성

오른쪽위 > 내작업 > 튜닝에서 확인한다. 학습 완료 클릭
튜닝 데이터셋과 '동일한 포맷으로 질의' 작성 테스트

예)

어르신 : 독감주사 맞아야하는데 어디로 가야하나?

실행

참고

<https://brunch.co.kr/@topasvga/3764>

튜닝1

대화 | LK-D2

생성 일시 2024.08.05. 15:29

Workflow ID	tuning-1823-240805-152920-zfpm1
Problem Type	대화
Model Engine	LK-D2
Dataset	튜닝테스트파일_conv (3).csv, 12075 of tokens
Train Loss	0.6976
Tokens Used	94,920

Inference Test

플래그라운드에서는 파라미터를 조절하여 테스트할 수 있습니다. 플래그라운드로 이동하여 Inference test를 진행해주세요.

Input

어르신 : 독감주사 맞아야하는데 어디로 가야하나?

실행

Output

이전

공유

테스트 앱 생성

튜닝 - API요청

사용자 데이터셋을 기반으로 학습이 완료되면, API 요청을 위한 코드 보기 기능 제공

- Curl과 python 코드 제공

테스트 앱 생성

테스트 앱 이름

튜닝1

3/30

생성

서비스 앱은 [여기](#)에서 신청해주세요.

히스토리

테스트 앱

생성 일시

생성된 테스트 앱이 없습니다.

코드 보기

API

다음 코드를 복사하여 사용자 환경에 적용할 수 있습니다.

코드 타입

curl

python

```
curl --location --request POST 'https://clovastudio.apigw.ntruss.com/testap
--header 'X-NCP-NT5ALY79F49Ft60iBw6XU4vs+' \
--header 'X-NCP-APIGW-API-KEY: ' \
--header 'X-NCP-CLOVASTUDIO-REQUEST-ID: ' \
--header 'Content-Type: application/json' \
--data '{
  "text": "input text",
  "start": "",
  "..."
}'
```

복사

API URL

https://clovastudio.apigw.ntruss.com/testapp/v1/tasks/x571ge1y/completions

복사

API Gateway Key

OOJ

재발급

복사

AI Filter

네이버 AI 오픈 존치를 기반으로 AI Filter를 적용합니다. [가이드 보기](#)

사용 사용 안 함

튜닝 - 챗모드, 불러오기

The screenshot shows the CLOVA Studio interface. On the left, the 'Mode' dropdown is set to '챗 모드'. The 'Engine' is set to 'HCX-003'. The 'Top P' slider is at 0.8, 'Top K' is at 0, 'Maximum tokens' is at 256, 'Temperature' is at 0.5, 'Repetition penalty' is at 5, 'Stop sequences' is '시퀀스 입력 후 Tab', and 'Seed' is '랜덤'. The main area shows a '시스템' (System) section with a prompt input field. A '불러오기' (Load) dialog is open, showing a table of jobs. The first row, '튜닝1', is highlighted with a red box.

작업 이름	작업 종류	모델 버전	생성 일시
튜닝1	WEB	LK-D2	8월 5일 15:29
test12	API	HCX-003	4월 29일 09:28
222	WEB	LK-D2	4월 27일 22:25

튜닝 - 챗모드, 불러오기

The screenshot shows the CLOVA Studio interface. On the left, the 'Mode' dropdown is set to '챗 모드'. The 'Engine' is set to 'HCX-003'. The 'Top P' slider is at 0.8, 'Top K' is at 0, 'Maximum tokens' is at 256, 'Temperature' is at 0.5, 'Repetition penalty' is at 5, 'Stop sequences' is '시퀀스 입력 후 Tab', and 'Seed' is '랜덤'. The main area shows a '시스템' (System) section with a prompt input field. A '불러오기' (Load) dialog box is open, showing a table of jobs. The first row, '튜닝1', is highlighted with a red box.

작업 이름	작업 종류	모델 버전	생성 일시
튜닝1	WEB	LK-D2	8월 5일 15:29
test12	API	HCX-003	4월 29일 09:28
222	WEB	LK-D2	4월 27일 22:25

튜닝 - 챗모드, 불러오기

Engine을 튜닝된 Engine으로 변경, 튜닝1로 변경하자
어르신 : 독감주사 맞아야하는데 어디로 가야하나? 로 질문 하자. > 실행
AI 가 답을 준다.

The screenshot shows the CLOVA Studio interface. On the left, the 'Mode' dropdown is set to '일반 모드' and the 'Engine' dropdown is set to '튜닝1'. Below these are sliders for 'Top P' (0.8), 'Top K' (0), 'Maximum tokens' (100), and 'Temperature' (0.5). On the right, the 'Generation type' is set to 'Rolling'. The main chat area shows a prompt: '어르신 : 독감주사 맞아야하는데 어디로 가야하나?' and a response: '어르신, 가까운 보건소에 가시면 돼요.' The '실행' button is highlighted in red.

감사합니다.