



How to USE

CLICK >

Version 0.0

C > 로컬 디스크 (C:) > research_persona > Final_project > m2snet			▼ ↺
이름	수정한 날짜	유형	
.ipynb_checkpoints	2019-11-12 오전 1...	파일 폴더	
__pycache__	2019-11-12 오후 3...	파일 폴더	
etc	2019-11-12 오후 5...	파일 폴더	
labeling	2019-11-12 오후 4...	파일 폴더	
model	2019-11-12 오후 5...	파일 폴더	
predict	2019-11-12 오후 5...	파일 폴더	
preprocess	2019-11-12 오후 3...	파일 폴더	
raw	2019-11-12 오후 3...	파일 폴더	
__init__	2019-11-12 오전 1...	Python File	
crawler	2019-11-12 오후 5...	Python File	
DevelopNOTE	2019-11-12 오후 3...	텍스트 문서	
utils	2019-11-12 오후 3...	Python File	

[raw]:

가공되지 않은 Review 데이터 적재

[preprocess]:

전처리된 데이터 적재

[labeling]:

전처리된 데이터에 감정 라벨 할당

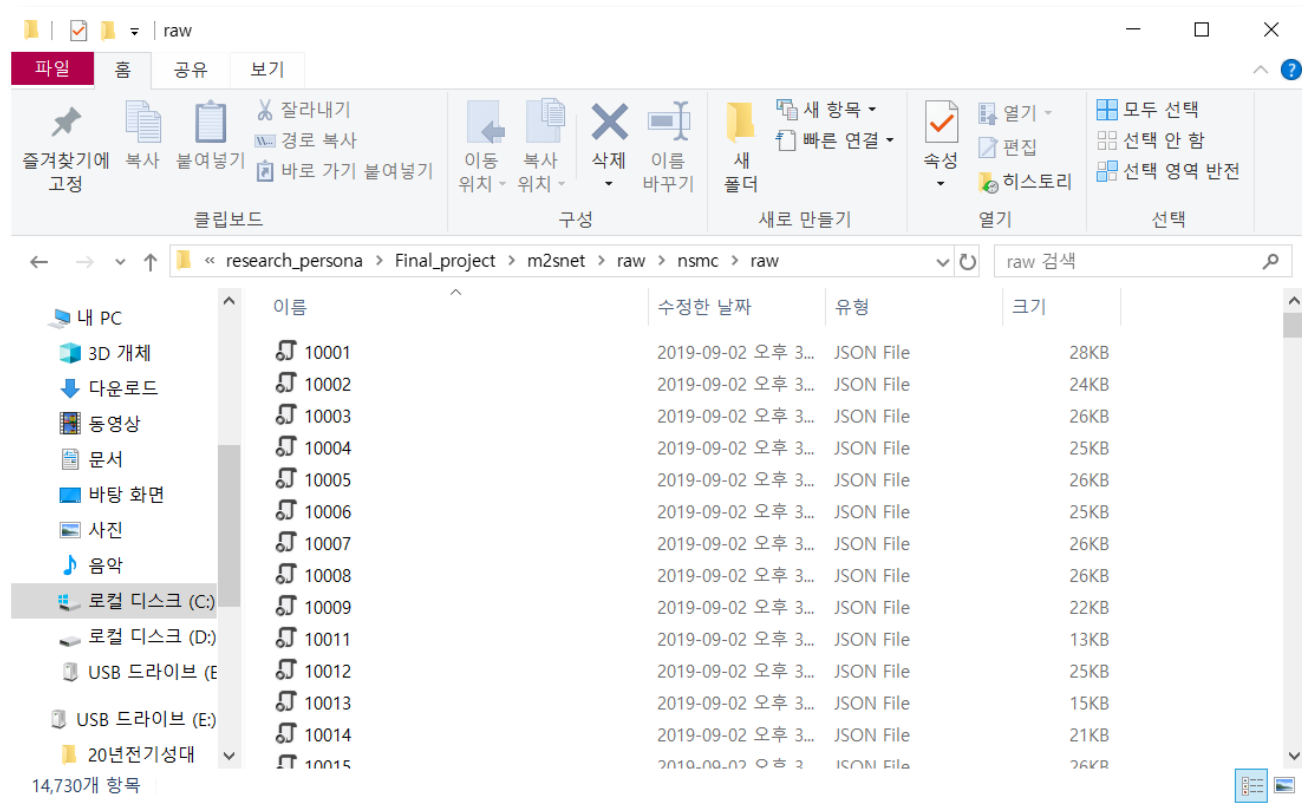
[model]:

Text를 학습시킬 수 있게 Embedding하고
감정 라벨 기반으로 지도 학습을 수행한 모델 적재

[predict]:

Input Text를 8가지 감정의 확률 값으로 mapping

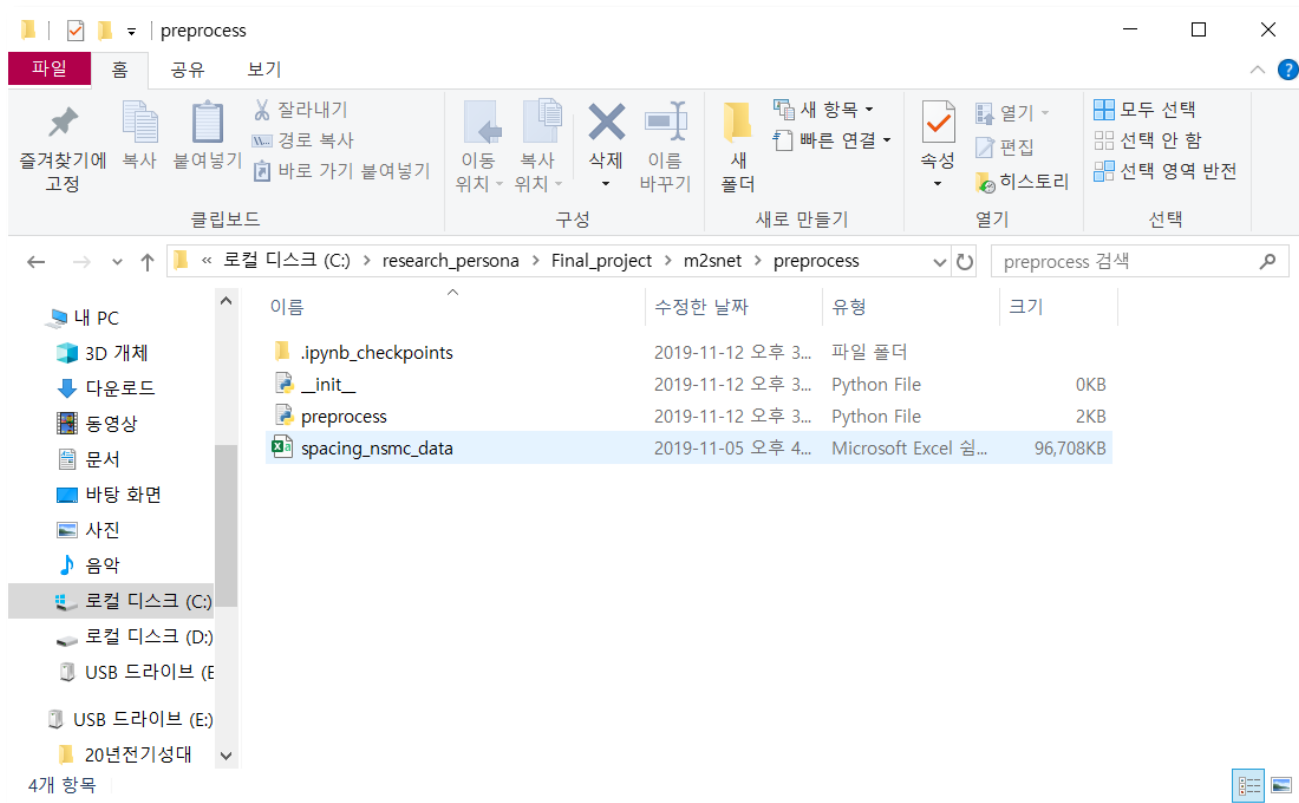
Raw



학습시킬 Raw Text file을 적재합니다.

- Version 0.0의 경우, NSMC 만을 그 대상으로 합니다.
- <https://github.com/e9t/nsmc>
- git clone으로 raw 폴더에 데이터를 적재한 후에 preprocess를 진행하세요
- 이후 version에서 다른 raw data도 handling할 수 있도록 업그레이드할 예정입니다.

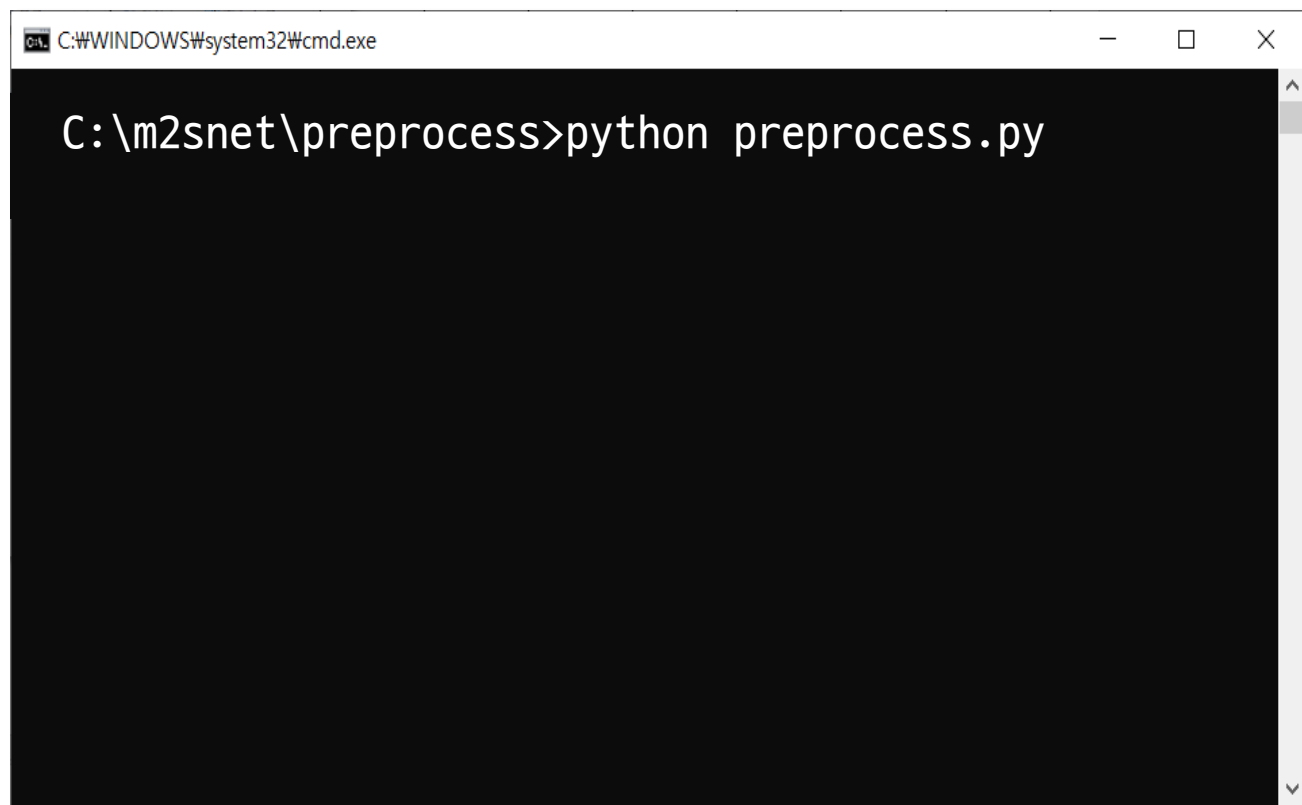
Preprocess



Labeling전 띄어 쓰기 전처리를 실시합니다.

- 작업을 수행하면 좌측 그림과 같이 'spacing_nsmc_data'가 생성됩니다.
- Version 0.0에서는 NSMC 데이터만 다루기 때문에 이렇게 결과가 저장되며 추후 업데이트 예정입니다.
- 현재 Labeling 단에서 숫자, 외국어, 불용어 전처리를 수행 중인데 이는 다음 버전에서 preprocess 단에서 처리될 예정입니다.

Preprocess



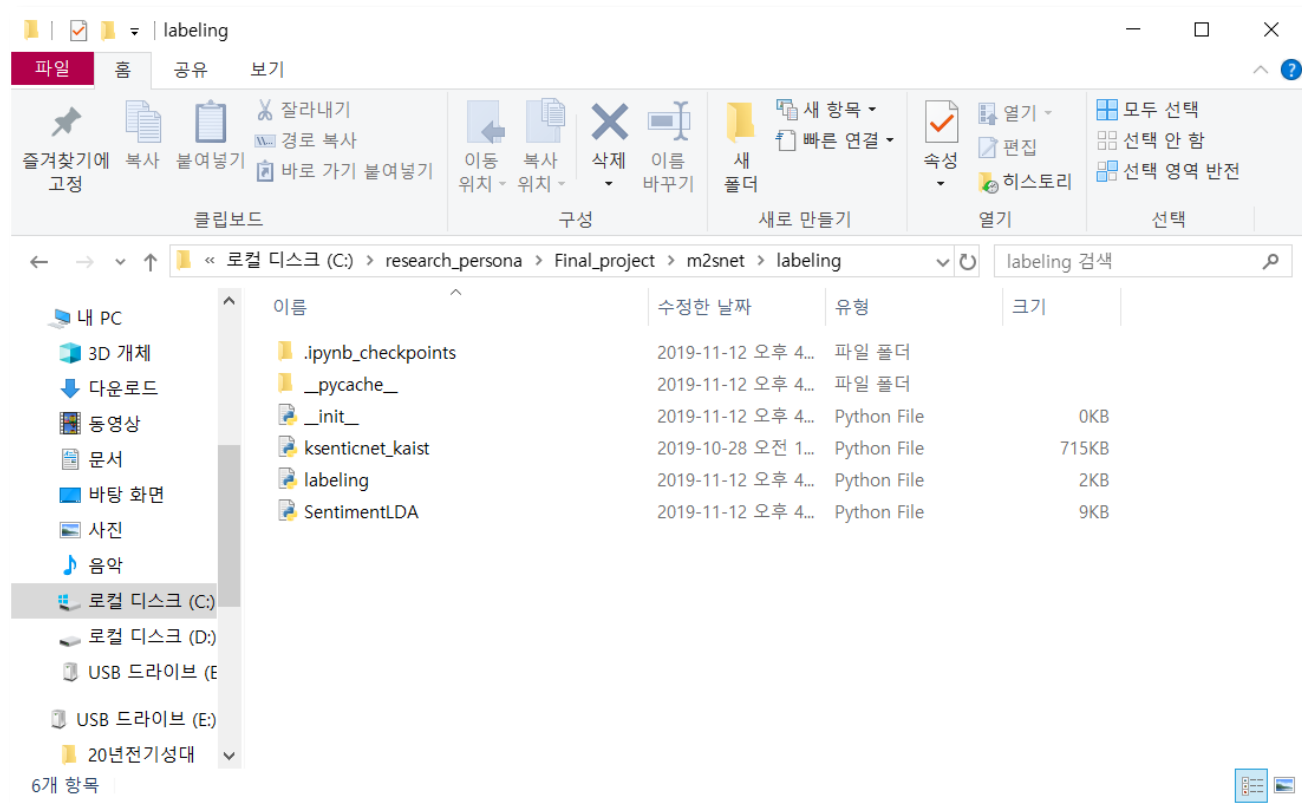
A screenshot of a Windows command prompt window. The title bar at the top reads "C:\WINDOWS\system32\cmd.exe". The command prompt shows the directory "C:\m2snet\preprocess" and the command "python preprocess.py" entered at the prompt.

```
C:\WINDOWS\system32\cmd.exe  
  
C:\m2snet\preprocess>python preprocess.py
```

Labeling전 띄어 쓰기 전처리를 실시합니다.

- Window + R 키를 눌러 실행창을 열고 command line interface를 동작합니다.
- m2snet/preprocess 폴더로 이동한 다음,
- `python preprocess.py` 명령어로 전처리 작업을 실시합니다.
- **%주의: 70만 Text 기준 3시간 소요**

Labeling



학습 데이터 셋에 Sentiment Label을 할당합니다.

- 한글 이외의 문자, Stopwords 전처리를 실시한 후 JST 감정 매핑을 실시합니다. (이 후 사라질 기능)
- 향후 Gibbs Sampling을 AutoEncoder로 대체할 예정이며 현재는 CBOW로 계산하고 있습니다.

Labeling



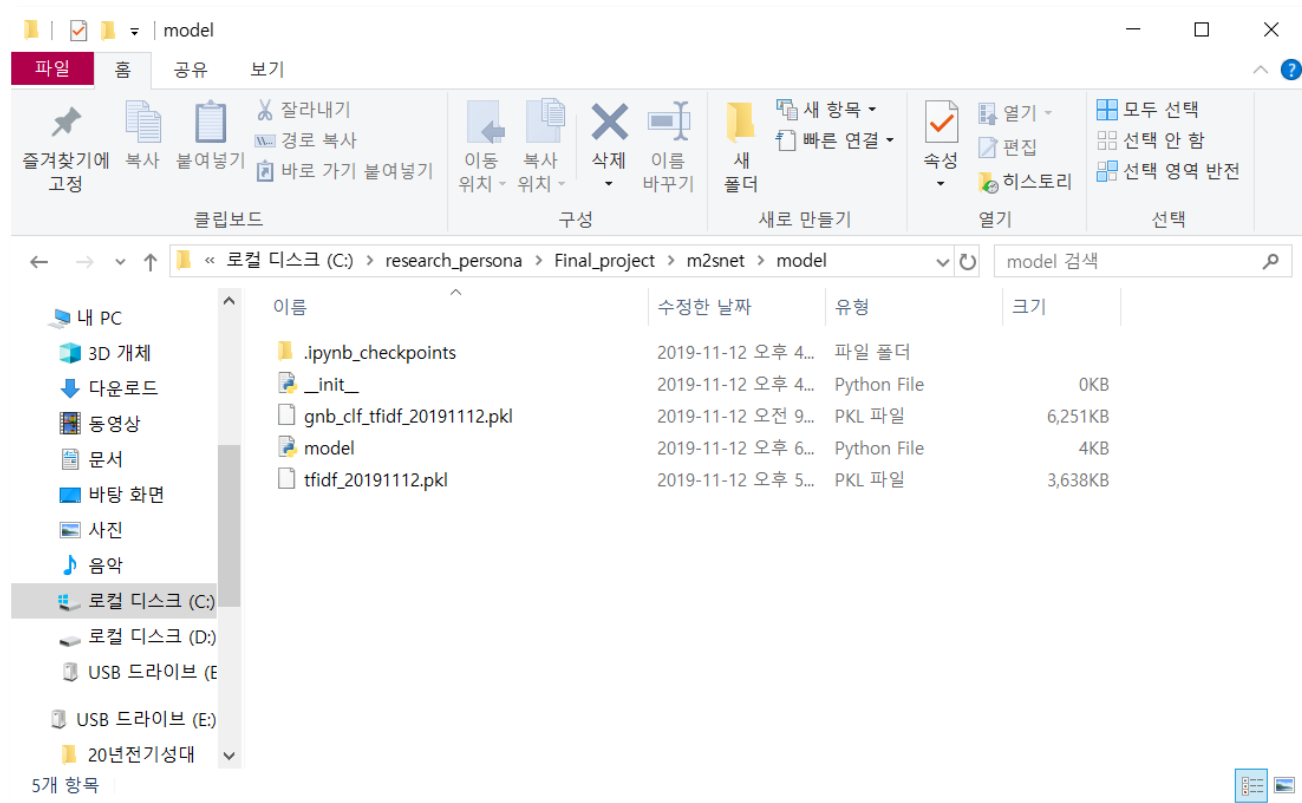
A screenshot of a Windows command prompt window. The title bar at the top reads "C:\WINDOWS\system32\cmd.exe". The command prompt shows the current directory as "C:\m2snet\labeling" and the command "python labeling.py" has been entered. The window has standard Windows window controls (minimize, maximize, close) in the top right corner.

```
C:\WINDOWS\system32\cmd.exe
C:\m2snet\labeling>python labeling.py
```

학습 데이터 셋에 Sentiment Label을 할당합니다.

- Window + R 키를 눌러 실행창을 열고 command line interface를 동작합니다.
- m2snet/labeling 폴더로 이동한 다음,
- `python labeling.py` 명령어로 전처리 작업을 실시합니다.
- **%주의: 70만 Text 기준 7시간 소요**

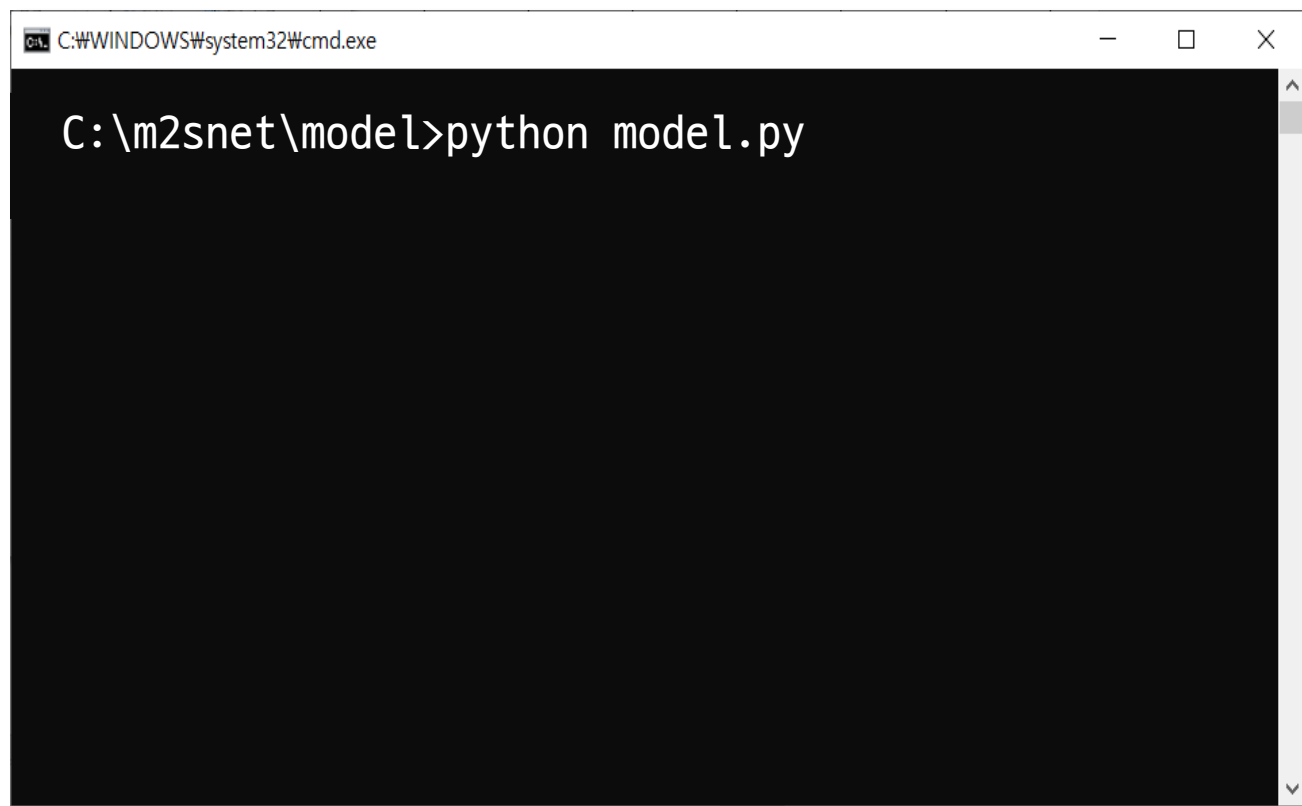
Model



Labeled된 데이터를 기반으로 Embedding, 지도 학습을 실시합니다.

- 현재 Embedding은 성능 관계로 TF-IDF만 지원합니다.
- 모델도 Multi Gaussian Naïve Bayes 모델만 사용합니다.
- Label Language Model 등으로 확장되면 Embedding과 Modeling도 변형된 상태로 모듈을 제공할 예정입니다.

Model

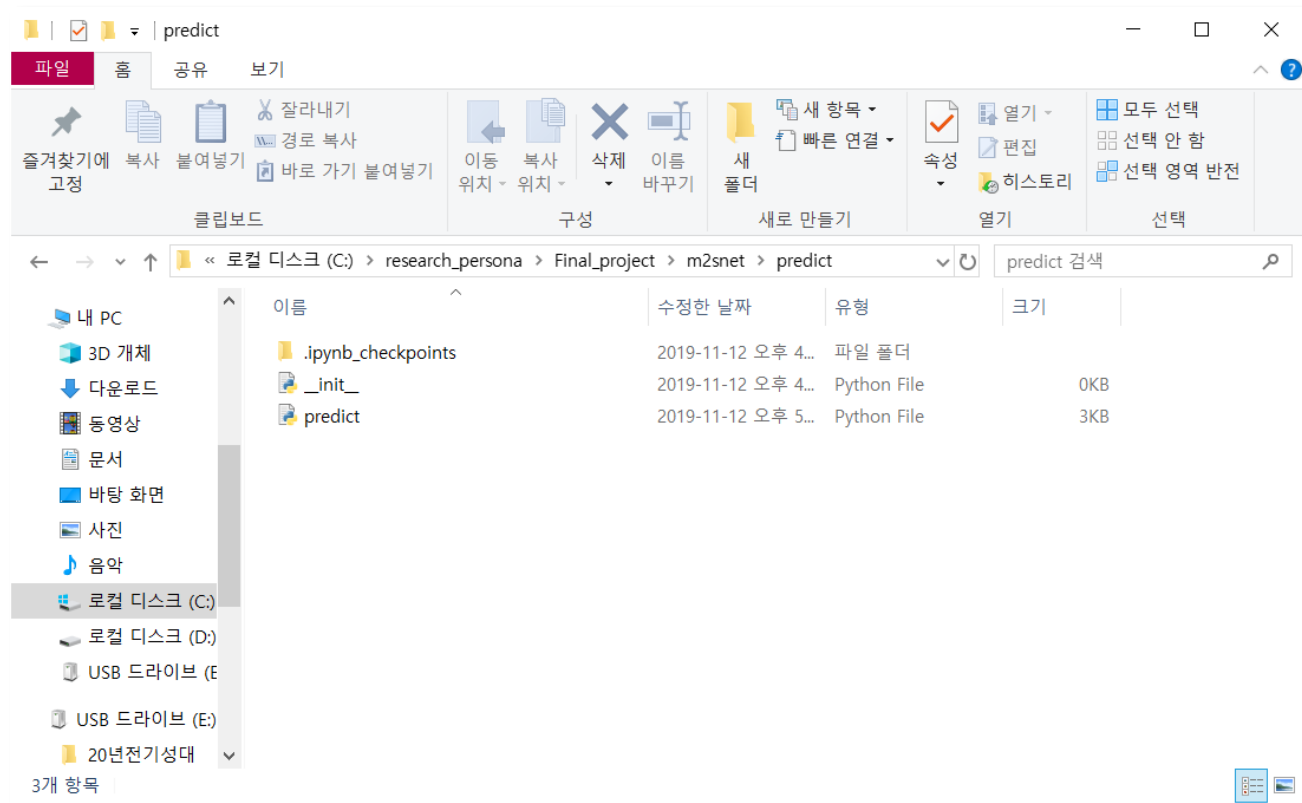


A screenshot of a Windows command prompt window. The title bar at the top reads "C:\WINDOWS\system32\cmd.exe". The command prompt shows the current directory as "C:\m2snet\model" and the command "python model.py" has been entered. The background of the command prompt is black with white text.

Labeled된 데이터를 기반으로 Embedding, 지도 학습을 실시합니다.

- Window + R 키를 눌러 실행창을 열고 command line interface를 동작합니다.
- m2snet/model 폴더로 이동한 다음,
- `python model.py` 명령어로 전처리 작업을 실시합니다.
- %주의: 70만 Text 기준 10분 소요 (Embedding)

Predict



들어오는 Text에 대하여 어떤 감정 범주를 가질 지 확률 값을 도출합니다.

- Text를 입력하면 Main Sentiment와 각 감정 범주에 대한 확률 값을 출력하고 이를 시각화 합니다.
- 다른 API와 접목시키기 위해선 코드 수준 수정이 필요합니다.

Predict

C:\WINDOWS\system32\cmd.exe - python predict.py

(keras) C:\research_persona\Final_project\m2snet\predict>python predict.py
C:\ProgramData\Anaconda3\envs\keras\lib\site-packages\jpype\core.py:210: User

Deprecated: convertStrings was not specified when starting the JVM. The default behavior in JPype will be False starting in JPype 0.8. The recommended setting for new code is convertStrings=False. The legacy value of True was assumed for this session. If you are a user of an application that reported this warning, please file a ticket with the developer.

```

"""
Loading JIT Compiled ChatSpace Model
Input 'embeddingname': tfidf_20191112.pkl
Input 'modelname': gnb_clf_tfidf_20191112.pkl

```

감정을 알고싶은 Text를 입력해주세요. 중국집에서 짜장면을 시켰는데 배달원이 불친절했고 무엇보다 맛이 없었다.

Probability Distribution:

```

[0.11456478 0.20982618 0.21680885 0.05903336 0.08222858 0.04560184
 0.21198296 0.05995346]

```

ANGER

