bert_sim

January 18, 2021

[2]:
```
!pwd
```

/home/Trial_notebook/bert_word_sim

[2]:
```
!git clone https://github.com/sonoisa/sentence-transformers
!cd sentence-transformers; pip install -r requirements.txt
```

```
Cloning into 'sentence-transformers'…
remote: Enumerating objects: 710, done.
remote: Total 710 (delta 0), reused 0 (delta 0), pack-reused 710
Receiving objects: 100% (710/710), 197.81 KiB | 228.00 KiB/s, done.
Resolving deltas: 100% (476/476), done.
Collecting transformers==2.3.0
  Downloading transformers-2.3.0-py3-none-any.whl (447 kB)
     |                            | 447 kB 7.4 MB/s eta 0:00:01
Requirement already satisfied: tqdm in /opt/conda/lib/python3.7/site-
packages (from -r requirements.txt (line 2)) (4.45.0)
Requirement already satisfied: torch>=1.0.1 in /opt/conda/lib/python3.7/site-
packages (from -r requirements.txt (line 3)) (1.5.0)
Requirement already satisfied: numpy in /opt/conda/lib/python3.7/site-packages
(from -r requirements.txt (line 4)) (1.18.1)
Requirement already satisfied: scikit-learn in /opt/conda/lib/python3.7/site-
packages (from -r requirements.txt (line 5)) (0.23.1)
Requirement already satisfied: scipy in /opt/conda/lib/python3.7/site-packages
(from -r requirements.txt (line 6)) (1.4.1)
Requirement already satisfied: nltk in /opt/conda/lib/python3.7/site-packages
(from -r requirements.txt (line 7)) (3.2.4)
Requirement already satisfied: mecab-python3==0.996.5 in
/opt/conda/lib/python3.7/site-packages (from -r requirements.txt (line 8))
(0.996.5)
Requirement already satisfied: sacremoses in /opt/conda/lib/python3.7/site-
packages (from transformers==2.3.0->-r requirements.txt (line 1)) (0.0.43)
Requirement already satisfied: boto3 in /opt/conda/lib/python3.7/site-packages
(from transformers==2.3.0->-r requirements.txt (line 1)) (1.14.2)
Requirement already satisfied: regex!=2019.12.17 in
/opt/conda/lib/python3.7/site-packages (from transformers==2.3.0->-r
requirements.txt (line 1)) (2020.4.4)
Requirement already satisfied: requests in /opt/conda/lib/python3.7/site-
```

1

packages (from transformers==2.3.0->-r requirements.txt (line 1)) (2.23.0)
Requirement already satisfied: sentencepiece in /opt/conda/lib/python3.7/site-packages (from transformers==2.3.0->-r requirements.txt (line 1)) (0.1.91)
Requirement already satisfied: future in /opt/conda/lib/python3.7/site-packages (from torch>=1.0.1->-r requirements.txt (line 3)) (0.18.2)
Requirement already satisfied: threadpoolctl>=2.0.0 in /opt/conda/lib/python3.7/site-packages (from scikit-learn->-r requirements.txt (line 5)) (2.1.0)
Requirement already satisfied: joblib>=0.11 in /opt/conda/lib/python3.7/site-packages (from scikit-learn->-r requirements.txt (line 5)) (0.14.1)
Requirement already satisfied: six in /opt/conda/lib/python3.7/site-packages (from nltk->-r requirements.txt (line 7)) (1.14.0)
Requirement already satisfied: click in /opt/conda/lib/python3.7/site-packages (from sacremoses->transformers==2.3.0->-r requirements.txt (line 1)) (7.1.1)
Requirement already satisfied: s3transfer<0.4.0,>=0.3.0 in /opt/conda/lib/python3.7/site-packages (from boto3->transformers==2.3.0->-r requirements.txt (line 1)) (0.3.3)
Requirement already satisfied: jmespath<1.0.0,>=0.7.1 in /opt/conda/lib/python3.7/site-packages (from boto3->transformers==2.3.0->-r requirements.txt (line 1)) (0.10.0)
Requirement already satisfied: botocore<1.18.0,>=1.17.2 in /opt/conda/lib/python3.7/site-packages (from boto3->transformers==2.3.0->-r requirements.txt (line 1)) (1.17.2)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in /opt/conda/lib/python3.7/site-packages (from requests->transformers==2.3.0->-r requirements.txt (line 1)) (1.24.3)
Requirement already satisfied: chardet<4,>=3.0.2 in /opt/conda/lib/python3.7/site-packages (from requests->transformers==2.3.0->-r requirements.txt (line 1)) (3.0.4)
Requirement already satisfied: idna<3,>=2.5 in /opt/conda/lib/python3.7/site-packages (from requests->transformers==2.3.0->-r requirements.txt (line 1)) (2.9)
Requirement already satisfied: certifi>=2017.4.17 in /opt/conda/lib/python3.7/site-packages (from requests->transformers==2.3.0->-r requirements.txt (line 1)) (2020.11.8)
Requirement already satisfied: python-dateutil<3.0.0,>=2.1 in /opt/conda/lib/python3.7/site-packages (from botocore<1.18.0,>=1.17.2->boto3->transformers==2.3.0->-r requirements.txt (line 1)) (2.8.1)
Requirement already satisfied: docutils<0.16,>=0.10 in /opt/conda/lib/python3.7/site-packages (from botocore<1.18.0,>=1.17.2->boto3->transformers==2.3.0->-r requirements.txt (line 1)) (0.15.2)
Installing collected packages: transformers
  Attempting uninstall: transformers
    Found existing installation: transformers 2.11.0
    Uninstalling transformers-2.11.0:
      Successfully uninstalled transformers-2.11.0

```
Successfully installed transformers-2.3.0
WARNING: You are using pip version 20.1.1; however, version 20.3.3 is
available.
You should consider upgrading via the '/opt/conda/bin/python3.7 -m pip install
--upgrade pip' command.
```

[4]:
```
!wget -O sonobe-datasets-sentence-transformers-model.tar "https://www.floydhub.
 ↪com/api/v1/resources/JLTtbaaK5dprnxoJtUbBbi?
 ↪content=true&download=true&rename=sonobe-datasets-sentence-transformers-model-2"
!tar -xvf sonobe-datasets-sentence-transformers-model.tar
```

```
--2021-01-08 18:47:31--  https://www.floydhub.com/api/v1/resources/JLTtbaaK5dprn
xoJtUbBbi?content=true&download=true&rename=sonobe-datasets-sentence-
transformers-model-2
Resolving www.floydhub.com (www.floydhub.com)… 172.67.72.144, 104.26.0.30,
104.26.1.30, …
Connecting to www.floydhub.com (www.floydhub.com)|172.67.72.144|:443…
connected.
HTTP request sent, awaiting response… 200 OK
Length: unspecified [application/tar]
Saving to: 'sonobe-datasets-sentence-transformers-model.tar'

sonobe-datasets-sen     [         <=>        ] 422.28M  20.8MB/s    in 19s

2021-01-08 18:47:50 (22.6 MB/s) - 'sonobe-datasets-sentence-transformers-
model.tar' saved [442788352]

./
./training_bert_japanese/
./training_bert_japanese/0_BERTJapanese/
./training_bert_japanese/0_BERTJapanese/added_tokens.json
./training_bert_japanese/0_BERTJapanese/config.json
./training_bert_japanese/0_BERTJapanese/pytorch_model.bin
./training_bert_japanese/0_BERTJapanese/sentence_bert_config.json
./training_bert_japanese/0_BERTJapanese/special_tokens_map.json
./training_bert_japanese/0_BERTJapanese/tokenizer_config.json
./training_bert_japanese/0_BERTJapanese/vocab.txt
./training_bert_japanese/1_Pooling/
./training_bert_japanese/1_Pooling/config.json
./training_bert_japanese/config.json
./training_bert_japanese/modules.json
```

[4]:
```
%cd sentence-transformers
```

```
/home/Trial_notebook/bert_word_sim/sentence-transformers
```

```python
[5]: from sentence_transformers import SentenceTransformer
     import numpy as np
```

```python
[7]: model_path = "../training_bert_japanese"
     model = SentenceTransformer(model_path, show_progress_bar=False)
```

```python
[38]: sentences = ["    "," "," "," "," "]
```

```python
[39]: sentence_vectors = model.encode(sentences)
```

```python
[40]: import scipy.spatial

      queries = ["    "," "]
      query_embeddings = model.encode(queries)

      closest_n = 5
      for query, query_embedding in zip(queries, query_embeddings):
          distances = scipy.spatial.distance.cdist([query_embedding],␣
       ↪sentence_vectors, metric="cosine")[0]

          results = zip(range(len(distances)), distances)
          results = sorted(results, key=lambda x: x[1])

          print("\n\n=====================\n\n")
          print("Query:", query)
          print("\nTop 5 most similar sentences in corpus:")

          for idx, distance in results[0:closest_n]:
              print(sentences[idx].strip(), "(Score: %.4f)" % (distance / 2))
```

```
=====================


Query:

Top 5 most similar sentences in corpus:
     (Score: 0.0000)
   (Score: 0.2832)
     (Score: 0.3014)
   (Score: 0.3500)
   (Score: 0.3716)


=====================
```

```
Query:

Top 5 most similar sentences in corpus:
    (Score: 0.0000)
     (Score: 0.1230)
   (Score: 0.2388)
     (Score: 0.2832)
   (Score: 0.3482)
```

[ ]:

[ ]: