1_0_Lexical_Complexity_Dataset_Functional_Test_MVP_FINAL

April 13, 2025

[ ]:
```
#@title Install Packages
```

[ ]:
```
!pip install -q transformers
!pip install -q torchinfo
!pip install -q datasets
!pip install -q evaluate
!pip install -q nltk
```

```
                        491.2/491.2 kB
5.4 MB/s eta 0:00:00
                        116.3/116.3 kB
3.2 MB/s eta 0:00:00
                        183.9/183.9 kB
10.5 MB/s eta 0:00:00
                        143.5/143.5 kB
8.0 MB/s eta 0:00:00
                        194.8/194.8 kB
11.1 MB/s eta 0:00:00
```

ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is the source of the following dependency conflicts.
gcsfs 2025.3.2 requires fsspec==2025.3.2, but you have fsspec 2024.12.0 which is incompatible.
torch 2.6.0+cu124 requires nvidia-cublas-cu12==12.4.5.8; platform_system == "Linux" and platform_machine == "x86_64", but you have nvidia-cublas-cu12 12.5.3.2 which is incompatible.
torch 2.6.0+cu124 requires nvidia-cuda-cupti-cu12==12.4.127; platform_system == "Linux" and platform_machine == "x86_64", but you have nvidia-cuda-cupti-cu12 12.5.82 which is incompatible.
torch 2.6.0+cu124 requires nvidia-cuda-nvrtc-cu12==12.4.127; platform_system == "Linux" and platform_machine == "x86_64", but you have nvidia-cuda-nvrtc-cu12 12.5.82 which is incompatible.
torch 2.6.0+cu124 requires nvidia-cuda-runtime-cu12==12.4.127; platform_system == "Linux" and platform_machine == "x86_64", but you have nvidia-cuda-runtime-cu12 12.5.82 which is incompatible.
torch 2.6.0+cu124 requires nvidia-cudnn-cu12==9.1.0.70; platform_system == "Linux" and platform_machine == "x86_64", but you have nvidia-cudnn-cu12 9.3.0.75 which is incompatible.
torch 2.6.0+cu124 requires nvidia-cufft-cu12==11.2.1.3; platform_system == "Linux" and platform_machine == "x86_64", but you have nvidia-cufft-cu12 11.2.3.61 which is incompatible.
torch 2.6.0+cu124 requires nvidia-curand-cu12==10.3.5.147; platform_system == "Linux" and platform_machine == "x86_64", but you have nvidia-curand-cu12 10.3.6.82 which is incompatible.
torch 2.6.0+cu124 requires nvidia-cusolver-cu12==11.6.1.9; platform_system == "Linux" and platform_machine == "x86_64", but you have nvidia-cusolver-cu12 11.6.3.83 which is incompatible.
torch 2.6.0+cu124 requires nvidia-cusparse-cu12==12.3.1.170; platform_system == "Linux" and platform_machine == "x86_64", but you have nvidia-cusparse-cu12 12.5.1.3 which is incompatible.
torch 2.6.0+cu124 requires nvidia-nvjitlink-cu12==12.4.127; platform_system == "Linux" and platform_machine == "x86_64", but you have nvidia-nvjitlink-cu12 12.5.82 which is incompatible.

84.0/84.0 kB

```
2.0 MB/s eta 0:00:00
```

[ ]:
```
!sudo apt-get update
! sudo apt-get install tree
```

```
Get:1 http://security.ubuntu.com/ubuntu jammy-security InRelease [129 kB]
Hit:2 http://archive.ubuntu.com/ubuntu jammy InRelease
Get:3 http://archive.ubuntu.com/ubuntu jammy-updates InRelease [128 kB]
Get:4 https://cloud.r-project.org/bin/linux/ubuntu jammy-cran40/ InRelease
[3,632 B]
Get:5 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64
InRelease [1,581 B]
Get:6 http://archive.ubuntu.com/ubuntu jammy-backports InRelease [127 kB]
Get:7 https://r2u.stat.illinois.edu/ubuntu jammy InRelease [6,555 B]
Hit:8 https://ppa.launchpadcontent.net/deadsnakes/ppa/ubuntu jammy InRelease
Hit:9 https://ppa.launchpadcontent.net/graphics-drivers/ppa/ubuntu jammy
InRelease
Hit:10 https://ppa.launchpadcontent.net/ubuntugis/ppa/ubuntu jammy InRelease
Get:11 http://security.ubuntu.com/ubuntu jammy-security/universe amd64 Packages
[1,241 kB]
Get:12 http://security.ubuntu.com/ubuntu jammy-security/restricted amd64
Packages [3,978 kB]
Get:13 http://security.ubuntu.com/ubuntu jammy-security/main amd64 Packages
[2,775 kB]
Get:14 https://cloud.r-project.org/bin/linux/ubuntu jammy-cran40/ Packages [70.9
kB]
Get:15
https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64
Packages [1,381 kB]
Get:16 http://archive.ubuntu.com/ubuntu jammy-updates/restricted amd64 Packages
[4,148 kB]
Get:17 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 Packages [3,092
kB]
Get:18 http://archive.ubuntu.com/ubuntu jammy-updates/universe amd64 Packages
[1,540 kB]
Get:19 https://r2u.stat.illinois.edu/ubuntu jammy/main amd64 Packages [2,688 kB]
Get:20 https://r2u.stat.illinois.edu/ubuntu jammy/main all Packages [8,808 kB]
Fetched 30.1 MB in 8s (3,843 kB/s)
Reading package lists… Done
W: Skipping acquire of configured file 'main/source/Sources' as repository
'https://r2u.stat.illinois.edu/ubuntu jammy InRelease' does not seem to provide
it (sources.list entry misspelt?)
Reading package lists… Done
Building dependency tree… Done
Reading state information… Done
The following NEW packages will be installed:
  tree
```

```
0 upgraded, 1 newly installed, 0 to remove and 46 not upgraded.
Need to get 47.9 kB of archives.
After this operation, 116 kB of additional disk space will be used.
Get:1 http://archive.ubuntu.com/ubuntu jammy/universe amd64 tree amd64 2.0.2-1
[47.9 kB]
Fetched 47.9 kB in 0s (356 kB/s)
debconf: unable to initialize frontend: Dialog
debconf: (No usable dialog-like program is installed, so the dialog based
frontend cannot be used. at /usr/share/perl5/Debconf/FrontEnd/Dialog.pm line 78,
<> line 1.)
debconf: falling back to frontend: Readline
debconf: unable to initialize frontend: Readline
debconf: (This frontend requires a controlling tty.)
debconf: falling back to frontend: Teletype
dpkg-preconfigure: unable to re-open stdin:
Selecting previously unselected package tree.
(Reading database … 126210 files and directories currently installed.)
Preparing to unpack …/tree_2.0.2-1_amd64.deb …
Unpacking tree (2.0.2-1) …
Setting up tree (2.0.2-1) …
Processing triggers for man-db (2.10.2-1) …
```

[ ]:
```python
#@title Imports

import transformers
import evaluate

import nltk

from datasets import load_dataset
from torchinfo import summary

from transformers import AutoTokenizer, AutoModel,␣
 ↪AutoModelForSequenceClassification
from transformers import TrainingArguments, Trainer

import os
import pandas as pd
import numpy as np
```

[ ]:
```python
# @title Mount Google Drive
```

[ ]:
```python
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

[ ]:
```
dir_root = '/content/drive/MyDrive/266-final/'
# dir_data = '/content/drive/MyDrive/266-final/data/'
dir_data = '/content/drive/MyDrive/266-final/data/se21-t1-comp-lex-master/'
dir_models = '/content/drive/MyDrive/266-final/models/'
dir_results = '/content/drive/MyDrive/266-final/results/'
```

[ ]:
```
!tree -L 2 /content/drive/MyDrive/266-final/data/se21-t1-comp-lex-master/
```

```
/content/drive/MyDrive/266-final/data/se21-t1-comp-lex-master/
   evaluate.py
   Readme.md
   test
       lcp_multi_test.tsv
       lcp_single_test.tsv
   test-labels
       lcp_multi_test.tsv
       lcp_single_test.tsv
   train
       lcp_multi_train.tsv
       lcp_single_train.tsv
   trial
        lcp_multi_trial.tsv
        lcp_single_trial.tsv

4 directories, 10 files
```

[ ]:
```
# !tree -L 4 /content/drive/MyDrive/266-final/
```

[ ]:
```
!ls -R /content/drive/MyDrive/266-final/data/se21-t1-comp-lex-master/
```

```
/content/drive/MyDrive/266-final/data/se21-t1-comp-lex-master/:
evaluate.py  Readme.md  test  test-labels  train  trial

/content/drive/MyDrive/266-final/data/se21-t1-comp-lex-master/test:
lcp_multi_test.tsv  lcp_single_test.tsv

/content/drive/MyDrive/266-final/data/se21-t1-comp-lex-master/test-labels:
lcp_multi_test.tsv  lcp_single_test.tsv

/content/drive/MyDrive/266-final/data/se21-t1-comp-lex-master/train:
lcp_multi_train.tsv  lcp_single_train.tsv

/content/drive/MyDrive/266-final/data/se21-t1-comp-lex-master/trial:
lcp_multi_trial.tsv  lcp_single_trial.tsv
```

[ ]:
```
#@title Import Data
```

[ ]:
```
single_train_df = pd.read_csv(os.path.join(dir_data, "train", "lcp_single_train.
 ↪tsv"), sep="\t")
multi_train_df = pd.read_csv(os.path.join(dir_data, "train", "lcp_multi_train.
 ↪tsv"), sep="\t")

single_test_df = pd.read_csv(os.path.join(dir_data, "test", "lcp_single_test.
 ↪tsv"), sep="\t")

try:
    multi_test_df = pd.read_csv(
        os.path.join(dir_data, "test", "lcp_multi_test.tsv"),
        sep="\t",
        on_bad_lines='skip'
    )
    print("Loaded with skipping bad lines")
except Exception as e:
    print(f"First approach failed: {e}")
    try:
        multi_test_df = pd.read_csv(
            os.path.join(dir_data, "test", "lcp_multi_test.tsv"),
            sep="\t",
            engine="python",
            quoting=3  # QUOTE_NONE
        )
        print("Loaded with Python engine")
    except Exception as e:
        print(f"Second approach failed: {e}")
        with open(os.path.join(dir_data, "test", "lcp_multi_test.tsv"), 'r') as␣
 ↪file:
            lines = file.readlines()

        import io
        good_lines = lines[:39] + lines[40:] if len(lines) >= 40 else lines
        multi_test_df = pd.read_csv(io.StringIO(''.join(good_lines)), sep="\t")
        print("Loaded by skipping problematic line manually")

print(f"\nSingle word training data: {single_train_df.shape[0]} records with␣
 ↪{single_train_df.shape[1]} columns")
print(f"Multi word training data: {multi_train_df.shape[0]} records with␣
 ↪{multi_train_df.shape[1]} columns")
print(f"Single word test data: {single_test_df.shape[0]} records with␣
 ↪{single_test_df.shape[1]} columns")
print(f"Multi word test data: {multi_test_df.shape[0]} records with␣
 ↪{multi_test_df.shape[1]} columns")

single_test_labels_df = pd.read_csv(
```

```
    os.path.join(dir_data, "test-labels", "lcp_single_test.tsv"),
    sep="\t",
    engine="python",
    quoting=3  # QUOTE_NONE
)

multi_test_labels_df = pd.read_csv(
    os.path.join(dir_data, "test-labels", "lcp_multi_test.tsv"),
    sep="\t",
    engine="python",
    quoting=3  # QUOTE_NONE
)

print(f"Single word test labels: {single_test_labels_df.shape[0]} records with␣
 ↪{single_test_labels_df.shape[1]} columns")
print(f"Multi word test labels: {multi_test_labels_df.shape[0]} records with␣
 ↪{multi_test_labels_df.shape[1]} columns")
```

```
First approach failed: Error tokenizing data. C error: EOF inside string
starting at row 40
Loaded with Python engine

Single word training data: 7232 records with 5 columns
Multi word training data: 1464 records with 5 columns
Single word test data: 808 records with 4 columns
Multi word test data: 184 records with 4 columns
Single word test labels: 917 records with 5 columns
Multi word test labels: 184 records with 5 columns
```

[ ]:

[ ]:

[ ]:

[ ]:

[ ]:

[ ]:

[ ]:

[ ]:

[ ]:

[ ]:

[ ]: