Transformers-based Methods for Data Augmentation (DA)

This is a research project focused on DA through the use of Transformers. Our aim is to provide tools based on BERT and GPT2 that can augment text-data and thus improve classification.

Folder Information

This folder contains an example data file (train_data_file.csv) and four python3 files. Each of them with a specific method for data augmentation through the use of Transformers. Each of the files creates four different new files containing augmented data for the example data.

1. augmenting_data_single_mask_1.0.py: Utilizes BERT and Single Masking method for augmenting data.
2. augmenting_data_double_mask_1.0.py: Utilizes BERT and Double Masking method for augmenting data.
3. augmenting_data_triple_mask_1.0.py: Utilizes BERT and Triple Masking method for augmenting data.
4. augmenting_data_augmented_sentence.py: Utilizes GPT2 to augment the lenght of a sentence.

Installation Requirements

In order to run each of the documents use the package manager pip3 and install the libraries for Transformers.

```
pip3 install transformers
```

Usage

In order to use any of the files is as simple as to run them through the terminal.

```
python3 augmenting_data_double_mask_1.0.py
```

If you see the following on your terminal, it means it worked.

```
Importing the data...
Separating sentences from labels...
Augmenting data...
File augmented_data_double_masking_1.csv created.
File augmented_data_double_masking_3.csv created.
File augmented_data_double_masking_5.csv created.
File augmented_data_double_masking_10.csv created.
Finish without errors.
```

Author

José Medardo Tapia Téllez