



Learn Git and GitHub without any code!

Using the Hello World guide, you'll start a branch, write comments, and open a pull request.

[Read the guide](#)

 [castorini](#) / [hedwig](#)

PyTorch deep learning models for document classification

[#pytorch](#) [#deep-learning](#) [#document-classification](#)

 229 commits

 1 branch

 0 packages

 0 releases

 14 contributors

 Apache-2.0

Branch: master ▾

[New pull request](#)

[Create new file](#)

[Upload files](#)

[Find file](#)

[Clone or download ▾](#)



achyudh Merge pull request [#62](#) from castorini/achyudh-update-readme ...

Latest commit c4c0545 8 days ago

 [common](#)

Fix parameter order for bert forward call

5 months ago

 [datasets](#)

Migrate relevance transfer to pytorch-transformers

7 months ago

 [docs](#)

Integrate BERT into Hedwig ([#29](#)) ([#11](#))

14 months ago

models	Fix parameter order for bert forward call	5 months ago
tasks	Fix parameter order for bert forward call	5 months ago
utils	Migrate BERT and HBERT to pytorch-transformers	7 months ago
.gitignore	Make Kim CNN ONNX-exportable (#136)	2 years ago
LICENSE	Initial commit	15 months ago
README.md	Update instructions for setting up hedwig-data	9 days ago
__init__.py	Fix package imports	15 months ago
requirements.txt	Migrate BERT and HBERT to pytorch-transformers	7 months ago
setup.py	Fix package imports	15 months ago

README.md



This repo contains PyTorch deep learning models for document classification, implemented by the Data Systems Group at the University of Waterloo.

Models

- [DocBERT](#) : DocBERT: BERT for Document Classification ([Adhikari et al., 2019](#))
- [Reg-LSTM](#): Regularized LSTM for document classification ([Adhikari et al., NAACL 2019](#))

- [XML-CNN](#): CNNs for extreme multi-label text classification ([Liu et al., SIGIR 2017](#))
- [HAN](#): Hierarchical Attention Networks ([Zichao et al., NAACL 2016](#))
- [Char-CNN](#): Character-level Convolutional Network ([Zhang et al., NIPS 2015](#))
- [Kim CNN](#): CNNs for sentence classification ([Kim, EMNLP 2014](#))

Each model directory has a `README.md` with further details.

Setting up PyTorch

Hedwig is designed for Python 3.6 and [PyTorch](#) 0.4. PyTorch recommends [Anaconda](#) for managing your environment. We'd recommend creating a custom environment as follows:

```
$ conda create --name castor python=3.6
$ source activate castor
```

And installing PyTorch as follows:

```
$ conda install pytorch=0.4.1 cuda92 -c pytorch
```

Other Python packages we use can be installed via pip:

```
$ pip install -r requirements.txt
```

Code depends on data from NLTK (e.g., stopwords) so you'll have to download them. Run the Python interpreter and type the commands:

```
>>> import nltk
>>> nltk.download()
```

Datasets

There are two ways to download the Reuters, AAPD, and IMDB datasets, along with word2vec embeddings:

Option 1. Our [Wasabi](#)-hosted mirror:

```
$ wget http://nlp.rocks/hedwig -O hedwig-data.zip
$ unzip hedwig-data.zip
```

Option 2. Our school-hosted repository, [hedwig-data](#) :

```
$ git clone https://github.com/castorini/hedwig.git
$ git clone https://git.uwaterloo.ca/jimmylin/hedwig-data.git
```

Next, organize your directory structure as follows:

```
.
├── hedwig
└── hedwig-data
```

After cloning the hedwig-data repo, you need to unzip the embeddings and run the preprocessing script:

```
cd hedwig-data/embeddings/word2vec
tar -xvzf GoogleNews-vectors-negative300.tgz
```

If you are an internal Hedwig contributor using the machines in the lab, follow the instructions [here](#).