SynthNotes User Guide

Christopher Gropp, Kris Brown

March 2020

1 What is SynthNotes?

SynthNotes is a system to produce synthetic medical notes using information extracted from real notes, such as those in the MIMIC-III dataset. The resulting notes are intended to be used for developing and testing machine learning systems.

2 Installation

2.1 Requirements

SynthNotes relies on the following software;

- Python 3.0 or higher
- `tqdm`
- `numpy`
- `pandas`
- `lxml`
- `pyarrow`
- `fastparquet`
- `stringdist`
- `scikit-learn`
- `scipy`

Note that if you are installing using PIP, these dependencies should be resolved automatically.

2.2 PIP

2.3 Manual Installation

Clone the repository from `https://github.com/ebegoli/SynthNotes.git`. Then, navigate into that directory and execute

```
python setup.py install
```

or

```
pip install -e .
```

3 Usage

Once SynthNotes is installed, there is a short sequence of commands to create the necessary files used to generate new synthetic notes. In order, the full process is as follows;

1. parse
2. preprocess
3. cluster
4. generate

Information on these commands is available using

```
    synthnotes --help
```

and more detailed information on each command, including its arguments, can be viewed with

```
    synthnotes parse --help
```

and similar (replace “parse” with the appropriate command name).

3.1 Key Features

- Reproducibility - each stage of the SynthNotes pipeline has an optional random seed argument that can be used to control the sampling processes used. Additionally, this feature can be used with the document generation stage to ensure that separate processes produce distinct notes.