disorder-normalizer

A sieve-based system for normalizing disorder mentions in biomedical data.

The disorder-normalizer tool has been written in Java and is released as free software.

You can find more explanation about our normalization system at the following webpage:
http://www.hlt.utdallas.edu/~jld082000/normalization/

Usage:

1. If using the source code available in "src" folder. First, copy the "resources" folder into the "src" folder, then run the program as below.

```
java tool.Main <training-data-dir> <test-data-dir> <terminology/ontology-file> max-sieve-level
---------------------
Sieve levels:
1 for exact match
2 for abbreviation expansion
3 for subject<->object conversion
4 for numbers replacement
5 for hyphenation
6 for affixation
7 for disorder synonyms replacement
8 for stemming
9 for composite disorder mentions
10 for partial match
---------------------
```

An example execution using the "training", "test", and ontology "TERMINOLOGY.txt" files provided in the "ncbi-data" folder is shown below.

```
C:\disorder-normalizer\src>java tool.Main ..\ncbi-data\training\ ..\ncbi-data\test\ ..\ncbi-data\TERMINOLOGY.txt 10
```

2. If using the executable jar file "disorder-normalizer.jar". First, copy the "resources" folder into the same folder as the jar file, then run the program as show in example below.

```
C:\disorder-normalizer>java -jar disorder-normalizer.jar ..\ncbi-data\training\ ..\ncbi-data\test\ ..\ncbi-data\TERMINOLOGY.txt 10
```

Output:

On executing the program, a new folder called "output" will be automatically created in the same folder as the \<test-data-dir\> to which the result from normalizing the test mentions will be written. In addition, the system performance on normalizing the test disorder mentions will be printed to the terminal.

Please Note:

- In order to run the tool on new data, please ensure that your training, test, and terminology files are in the same format as the data provided in the "ncbi-data" folder.
- The training data files are used to train the normalizer. However, since disorder-normalizer is not a learning-based system, this folder can be empty.
- The tool will attempt to normalize the mentions in the test data folder files to the terminology.
- The TERMINOLOGY file is the ontology/knowledge-base used to which the test data mentions are normalized.

Detailed explanation about our normalization system can be found at the following webpage:
http://www.hlt.utdallas.edu/~jld082000/normalization/