DEPEND SYSTEM

Introduction

This document includes the instructions to run the DEPEND aspect extraction system and classify new patient comments. The system consists three sub-systems: System A, System B and a script to combine two systems. The document also discuss the processes to setup necessary environment run the system.

Environment setup

This system was developed and tested on a Unix system (Ubuntu 16.04, 64-bit), so it is recommended to use a unix system for the best performance.

Install Python

Install python 2.7 to run the system. After installing python, nevigate to the project root directory and run the following commands to install the required python modules.

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

NLTK data

System A requires NLTK stopwords corpus to run. If you do not have have nltk python module installed, run the following command to install:

```
$ pip install nltk
```

To download nltk corpus, run the commands below after starting the python command prompt:

```
> import NLTK
> nltk.download('stopwords')
```

- In the Python command line type “import NLTK” and press Enter
- After that type “nltk.download('stopwords')”

Install R

System B was developed in R. It requires R version 3.2.3. Apart from the built in packages, the system requires the following R packages:

- plyr 1.8.4
- dplyr 0.7.4
- NLP 0.1-1d
- tm 0.7-1

- RWeka 0.4-34
- e1071 1.6-8
- caret 6.0-77
- RTextTools 1.4.2
- Glmnet 2.0-13
- kernlab 0.9-25
- Mlr 2.11

To install a R package, start the R command prompt and run:

```
> install.packages("plyr")
```

Run System

Run the following comamnd to make prediction:

```
./run_prediction.sh <path/to/data/file.csv> <dataset type>
```

`<path/to/data/file.csv>` should be replaced by the csv file path that contains comments patient comments. An example datafile is given at 'sample_data_file.csv'. Moreover, `<dataset type>` should be replaved by either `MMHSCT` or `SRFT` based on the type of the data source. By default, the outputs will be saved in the root directory of the system. There are three types of outputs:

1. `output.csv`: this file contains prediction per comment.
2. `top_comments_system_a.csv`: Top five comments predicted by the System A
3. `top_comments_system_b.csv`: top five comments predicted by the System B