Homework # 5

November 28, 2020

1 POS tagging and Document Classification

Use the data set of 5000 medical cases "**medicaltranscriptions.csv**". Build a document classification model.

(1) Peform POS-tag on each medical transcript and create a diciontary of nouns only for all documents (i.e. NN, NNP, NNS and NNPS)

(2) Create a document vector representation using word count.

(3) Convert the "description" of each case to a vector (using the pretrained word to vector dictionary "PubMed-and-PMC-w2v.bin"). You may use the average of word vectors as the document vector representation.

(4) Build a multi-class classification SVM model to classify a document into one of the "medical_ specialty" categories. Use the document vector as your predictors and use "medical_ specialty" as your target variable.

(5) What are the in-sample recall rates for each document types?

(6) Which "medical_ specialty" has the highest recall rate?

(7) Validate your SVM model by training your model on 3000 cases and apply the model to the rest 1999 cases. Create the confusion matrix and calculate the recall rates for each document type.

(8) How are the test recall rates compare to the in sample recall rates?