

Assignment # 5

Dihui Lai
dlai@wustl.edu

Washington University in St Louis — June 23, 2019

1 NLP on Medical Transcripts

Understand medical notes is a challenging NLP problem. Lots of good application can be made if a machine can read doctors' notes and interpret the underlying medical conditions and severity. In this exercise, you are presented a simple data of 5000 medical cases "**medicaltranscriptions.csv**". Each case has the transcript and the associated medical specialty. Please

- (1) Do a tf-idf analysis on the terms that appear in all transcripts, identify the terms that has the highest tf-idf score.
- (2) Build a classification model and use the transcripts to predict the medical specialty by considering of the followings:
 - Find a way to convert each transcript into structured vector
 - Use log-loss as your error measure
 - Build a multi-class classification model using random forest
 - Build a multi-class classification model using glm
- (3) Compare the model performance use corss-validation methods. Which model would you recommend?