**MedicalNotes Classification Challenge Report**

Data Cleaning

For data cleaning, I first organized the documents into one file that organizes each document within a list object. Each sentence was tokenized into individual word. Furthermore, the stopwords and symbols were removed. In order to normalize the document, I used Snowball steamer to convert each token to its morphological root of the word. For lemmatization, I applied WordNetLemmatizer to group the different inflected words to a single item.

Feature Extraction

A distribution analysis was conducted to analyst the word count distribution for each document. There is no significant difference in the word count distribution across five categories. There are total 18 documents are empty. Among the 7 label empty documents, 4 of them were assigned as Urology (57.%). Non-negative Matrix Factorization and Latent Dirichlet allocation topic modeling was performed to discover important words (Appendix). Words such as ‘c5’ and ‘c6’ were important for identifying the orthopedic class. The word ‘tumor’ might be important for identifying radiology. Two feature extraction methods were used to convert the bag of words model into vector space: term frequency and reverse document frequency (TF-IDF) and cosine similarity. The TF-IDF model greatly outperformed the cosine similarity.

Classification

Since this is a multi-class classification problem, classification algorithms, including Support Vector Machine (SVM), Random forest, Bayes and OneVsRestClassifier were used. The machine learning package is scikit-learn and nltk. For training the model, I used stratified 10-fold cross validation method to train and evaluate each model. For parameter tuning, I used scikit GridSearch method to try 4500 combinations of all possible parameters for each model.

Results

The best model that generated based on the GridSearch default validation is SVM with rbf kernel type, gamma of 0.001 and C of 1000. The model is able to yield 0.746 accuracy. From the 5-fold stratified cross validation, the best model is SVM with rbf kenel, gamma 0.01 and C 10. The model generates the accuracy of 0.75. However, according to the submission result, the accuracy is only 0.7087. My assumption is the GridSearch model might be overfitted. However, the result from the 5-fold cross validation also under-performed the Kaggle evaluation matrix.

Future Step

In addition to the current model, there are more features can be used for this classification task. First, the header information can be a good feature to differentiate reports. Since each header group the information relates to a sub topic, potential ensemble learning methods can be used to perform the task. The K nearest neighbor algorithm might be an another alternative model. Since the goal is to classify documents into 5 different categories, the K-NN model should work well on this cluster related task.

**Appendix**

**Topics in NMF model:**

Topic #0:

histori patient medic past pain allergi review physic examin year present ani plan ill doe social famili old right knee signific extrem deni sign vital complaint mg time day today

Topic #1:

procedur patient oper anesthesia diagnosi preoper place postop left right general room incis complic steril posit supin use perform prep drape diagnos brought blood loss ml descript estim taken fractur

Topic #2:

pelvi ct contrast abdomen seen normal intraven quadrant exam lower fluid techniqu liver evid appendix cm stone pelvic kidney right abdomin obstruct appear small impress reason bowel attenu left base

Topic #3:

unremark exam left cn sensori motor hx fhx gait reflex shx pmh right med weak cc ms cours station strength coord extrem pp gen bue decreas 95 rhf respons person

Topic #4:

cervic c5 c6 c4 anterior herniat disc discectomi c7 c3 fusion cord neck pulposus spondylosi nucleus stenosi plate diagnos radiculopathi allograft instrument vertebr decompress spinal bodi pain bone osteophyt level

Topic #5:

tendon ligament tear normal medial later edema sheath joint complex collater humer bicep meniscus partial anterior plantar posterior superior cruciat extend imag minim longus patellar seri sublux patella knee extensor

Topic #6:

valv stress normal aortic perfus myocardi rest nuclear fraction mitral eject left ventricular ventricl mild test imag size wall regurgit atrial lv interpret gate function heart motion studi rate atrium

Topic #7:

colon colonoscopi polyp procedur biopsi cecum rectum colonoscop endoscop diagnosi iv rectal cancer endoscopi esophag vers mg screen fentanyl pass mcg advanc olympus normal perform esophagus explain indic consent hemorrhoid

Topic #8:

inguin hernia repair procedur right sac extern obliqu indirect diagnosi herniorrhaphi direct toler mesh floor cord taken bulg canal penros prolen left incis dissect groin fascia boy blunt descript umbil

Topic #9:

spine cervic fractur acut imag ct coron evid sagitt contrast impress thorac obtain reconstruct axial lumbar techniqu exam mild chang degen osseous align disk abnorm view sequenti demonstr structur visual

()

Fitting LDA models with tf features, n\_samples=20000 and n\_features=10000...

done in 1.434s.

**Topics in LDA model:**

Topic #0:

fetal week cm gestat normal twin fl hc ultrasound date consist pregnanc intrauterin instil ac bpd edd placenta 33 biometri 09 age 16 amniot uterus deliveri fluid day close 31

Topic #1:

tendon normal right ligament left tear distal later medial tunnel anterior carpal mild respons arteri radial ulnar muscl wrist nerv carotid median posterior minim edema imag sheath flexor joint plantar

Topic #2:

left right normal procedur size patient atrium note ventricl histori mild valv evid seen mitral function continu ventricular l4 regurgit oper assess appear hernia exam cervic umbil moder pain liver

Topic #3:

patient pain cervic histori c5 c6 anterior spine medic c4 left neck old l5 plan state year bilater includ right herniat past present extrem physic time knee lumbar spinal s1

Topic #4:

histori patient left medic year right exam unremark normal present extrem test examin old pain sign evalu ani past day bilater motor deni cours reveal physic time review neurolog allergi

Topic #5:

patient procedur left right diagnosi oper anesthesia preoper place postop normal perform use general year incis old diagnos posit histori room complic indic later note pain blood obtain skin ct

Topic #6:

patient anesthesia oper procedur histori unremark bleed room place diagnosi stabl reveal year general prep steril vital posit visual cours left exam aneurysm dr right preoper age usual tumor condit

Topic #7:

congenit histori patient feed year general medic old syndrom diseas disord heart oper anesthesia fed preoper ng bilater exam pain perform procedur hospit past cancer prostat cervic note spine diagnos

Topic #8:

normal imag stress rest perfus myocardi impress ct contrast nuclear abnorm heart techniqu obtain fraction test rate acut ventricular scan interpret indic eject evid exercis gate mci wall sinus axial

Topic #9:

tempor patient prostat hematuria procedur diagnosi tumor brain hyperplasia oper microscop preoper lobe anesthesia postop gross use benign voluntari fractur implant arter bladder histori left malleolus time dura year remov