Honors Report

Mahee Surya K.L. and Vishnu Tej Reddy M

Abstract— **This project deals with Automatic Cohort Detection For Understanding Patient Similarity.**

I. INTRODUCTION

The project deals with identifying the most similar documents, present in an archive, given a particular medical document as input. We have used a subset comprising of 110000 documents of MIMICIII corpus for our project.

II. PRE-PROCESSING

The data is first preprocessed and then given as input to the various models. We followed a two step approach for pre-processing and cleaning the data:

A. Removal of Filler Data

The first step of the pre-processing includes identifying the junk/filler data present in the documents and removing it. For example, documents often contain dates in the format [**2151-7-16**] and patient, doctor, hospital names in the format [**First Name4 (NamePattern1) 1775**] etc. This information is a filler information used to preserve the identity of the parties involved. Hence, we have removed this kind of fillers as part of the first phase of pre-processing.

B. Removal of empty fields

In the second part, we have removed the fields which are empty i.e., fields which do not contain any values. For example, Admission Date, Discharge Date etc., do not have any values after the first phase of pre-processing. Hence, such fields are removed.

III. APPROACHES FOLLOWED

We have explored various approaches to find the document similarity among a set of documents. One of them was Doc2Vec which was explored during the first phase of the semester. We had observed that though the similarity scores that were returned by the model were pretty high, close to 0.8, the documents turned out to be not too similar. This was found to be a flaw with the model which couldnt perform the transformation from the text space to vector space with high accuracy. This arises because all the document is represented as a vector. As the length of the document increases, the amount of information that can be held the vector decreases. This results in the high similarity scores that were observed due to the large document size. The code and outputs are located here [1].

In the second phase of the semester we explored the methods of Tf-Idf, LDA coupled with Topic Modelling and LSTM based neural network. In the final phase of the semester, we implemented the models that were explored before and obtained the corresponding outputs.

A. Tf-Idf Model

The Term frequency-Inverse document frequency method is a numerical statistic intended to reflect how important a word is to a document in a collection or corpus. The tf-idf value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some common medical terms appear more frequently in general. We used the raw tf-idf method to compute the tfs and idfs for each word. We computed the combined tf-idf score from the individual scores for every word. Now, for every word in the test document we tried to find the document which contains the maximum number of words common to the test document containing a high tf-idf score. We return the corresponding document as output. This method outperformed the Doc2Vec model. The similarity scores obtained using this method were decent in certain cases while not so good in the others, the range being 0.22-0.87.

Term Frequency: It measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (aka. the total number of terms in the document) as a way of normalization:

Tf(t,d) = (Number of times term t appears in a document d) / (Total number of terms in the document d).

$$Tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t,d}}$$

Inverse Document Frequency: It which measures how important a term is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as ”is”, ”of”, and ”that”, may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the following:

IDF(t,D) = log_e(Total number of documents / Number of documents with term t in it).

$$Idf(t,D) = log\frac{N}{|d \in D : t \in d|}$$

$$TfIdf(t,d,D) = Tf(t,d) * Idf(t,D)$$

We have used the gensim library of python to build a tf-idf based model for the training data.

The code and outputs are located here [2].

B. LDA Model

Latent Dirichlet allocation is a generative statistical model that allows sets of documents to be explained by unobserved documents that explain why some parts of the data are similar. For example, in documents it postulates that each document is a mixture of a small number of topics and that each word's presence is attributable to one of the document's topics. In LDA, each document may be viewed as a mixture of various topics where each document is considered to have a set of topics that are assigned to it via LDA.

We have used the HDP(Hierarchical Dirichlet Process) to identify the optimal number of topics that are present in the training data. We got an unusually high number of topics as output. As that wasnt computationally feasible, we set the no of topics to 500 which is a decent number for a medium sized dataset. The outputs had a good similarity score, in the range of 0.76-0.95 and outperformed the Tf-Idf model. This is due to the probability distribution approach taken by the LDA model which assumes documents to be a distribution of distributions(topics). The code and outputs are located here [3].

C. LSTM Model

Long Short-Term Memory model is built upon basic RNN model but avoiding one of the key limitations of RNN to work with long sequences due to vanishing gradients. RNN faces the same issue as was faced with Doc2Vec. The encoding vector that is obtained from the RNN contains the embedding of the whole document. As the document size increases, the ability of RNN vector to represent the whole document decreases i.e., the models faces the issue of vanishing gradient. In LSTM, a memory cell is introduced which uses gates to decide how much information needs to be forgotten or need to flow through the time steps. Using the memory cell, LSTM vectors retain most of the information that is necessary to represent the document.

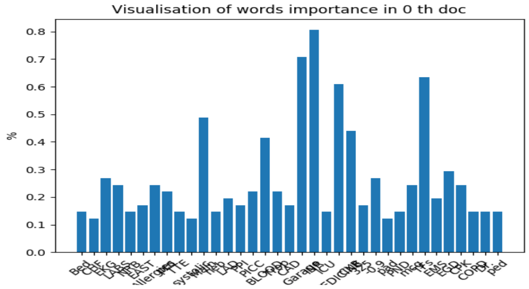
We have used the InferSent Model developed by Facebook to perform the document similarity using LSTM. We have also explored other LSTM based methods like BiLSTM, MLSTM, Siamese networks etc. These will be implemented in the future.

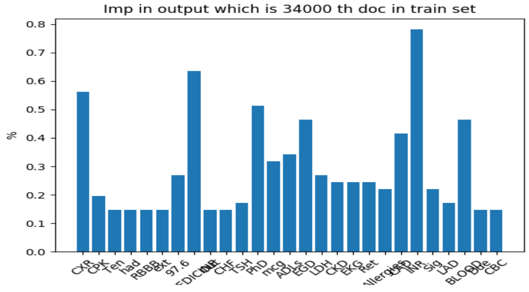
1) Infersent Model: Infersent is a sentence embedding method that provides semantic representations for English sentences. It is trained on natural language inference data and generalizes well to many different tasks. We have used infersent2.pkl model for the task of doc embedding which is trained with fastText (which have been trained on text preprocessed with the MOSES tokenizer).

Parameters: We have tried various parameters and found the optimal parameters for the document similarity task. The embedding size of each word vector is limited to 300, the encoded output vector from the LSTM model is limited to 2048 dimensions etc.

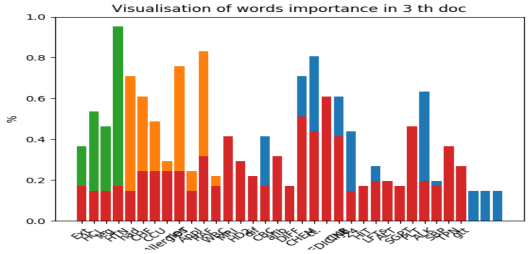
Results: To demonstrate the working of the model and visualise both inputs & outputs we ran the model on a test_set of 10 documents.

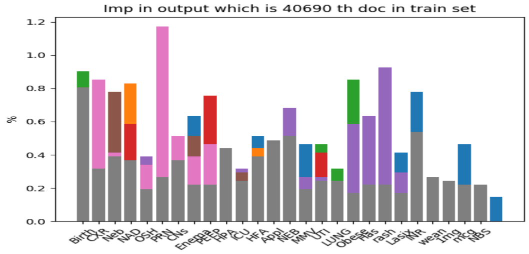
The input and output visualizations for 1st document are as follows:



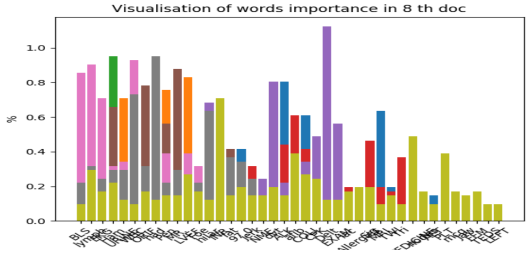


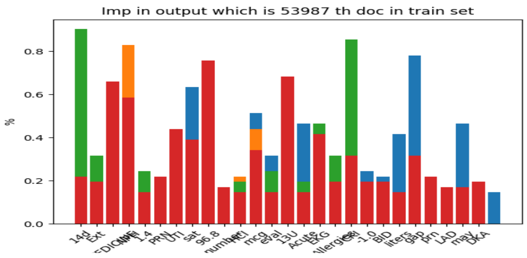
The input and output visualizations for 3rd document are as follows:





The input and output visualizations for 8th document are as follows:





cm]outtnew$_5$3987

IV. Observations

Sample Document:

Sex: M Service: Firm HISTORY OF PRESENT ILLNESS: The patient is with mental aspiration pneumonia, who was admitted to the Intensive Care in bed in respiratory distress. He was given Albuterol neb 88 percent, and he was intubated for apnea. An initial chest x-ray showed bilateral lower lobe opacities the day after admission. On presentation the patient was febrile and started on called out to the floor three days after admission. PAST MEDICAL HISTORY: Seizure disorder secondary to anoxic had increasing seizure activity from baseline according to be combative and assaultive at times). MEDICATIONS ON ADMISSION: Depakote 500 mg p.o. t.i.d., Albuterol nebs q.6 hours p.r.n., Colace 100 mg p.o. b.i.d., b.i.d., Citalopram. FAMILY HISTORY: No seizure disorder. ALLERGIES: Phenobarbital, Penicillin, Haldol. PHYSICAL EXAMINATION: Vital signs: Upon transfer to the pulse 83, ranging from 51-126, blood pressure 110/47, percent on 4 L nasal cannula. General: Examination was patient was awake and responding to voice appropriately. HEENT: Moist mucous membranes. Pupils equal, round and reactive to light. Neck: Supple. Cardiovascular: Normal S1 and S2. Regular rate and rhythm. Lungs: Decreased Abdomen: Normoactive bowel sounds. Extremities: No edema. LABORATORY DATA: White count 6.4, hematocrit 40.5; ASSESSMENT: This was a 41-year-old male with mental medication therapy, admitted with aspiration and probable Aspiration pneumonia: The patient was treated with a seven- Flagyl; however, this was discontinued before day 7 because was placed on aspiration precautions. During his hospitalization, he had multiple episodes of addition, he had multiple seizure activity, and therefore was side speech and swallow evaluation recommended ground/pureed well. In addition, the patient was placed on a quick steroid taper the rest of his admission, and his lung examination dictation. Seizure disorder: Over the past few months, the patient's his Keppra dose was decreased to 1500 b.i.d. He was The patient had multiple episodes of witnessed seizures, some to intravenous Ativan; however, on two episodes, he had than 20 min. Neurology was consulted. The patient received two EEGs seizure activity. There was generalized swelling, likely due Epilepsy Service for continuous monitoring. Psychiatric: The patient has a history of combative behavior similar code was called during this hospitalization. Psychiatry was involved and recommended p.r.n. for dose of Keppra was increased back to his home dose of mg General Medicine Team. An addendum will be dictated by the monitoring.

Tf-Idf Output: Similarity score- 0.339 Jaccard Score-0.742

Date of Birth: Sex: M Service: MEDICINE Allergies: Chief Complaint: picc line and central access 65 year-old gentleman with multiple medical problem including COPD, post tracheostomy and placement(respiratory failure presenting yet again with hypotension and altered mental status. Patient unable to give a history at this time so obtained from at for a Klebsiella UTI and hypotension. Since his the commode with assistance. On , the pt became lethargic 60 over palp and the pt was noted to be diaphoretic. He received unresponsive during this time. ABG showed 7.265/92.7/82 on an lumen was placed . Wife later arrived at the hospital and was able to provide until Friday. They were working on weaning him and he was able Friday, the pt felt mildly more SOB per his report. He was Yesterday, the pt's wife reports that he looked "very scared" eyes "rolling back in his head". He was occasionally responsive last 24 hours and his CO2 had been elevated. When they changed slightly less confused. She also notes that he was very The pt's BP has always been very low in his left arm and she a skin tear on the right. In the ED, the pt's VS were, 99.8 85 80/60-L 150/80-R 20 100% AC initially started on levophed for hypotension. However, after low in the left arm, it was checked in the right and has been Normal lactate. . 1. Squamous cell lung carcinoma, status post right 2. Prostate cancer, status post radical prostatectomy. 4. Type 2 diabetes mellitus. 6. Atrial fibrillation. 8. Gout. 10. Gastroesophageal reflux disease. 12. Hypertension. 14. Hypercholesterolemia. 16. Anxiety. 18. History of herpes zoster. and bronchitis, last in resulting in ventilator 20. vitamin B12 deficiency. 21. Diastolic heart failure. Echo : LVEF¿55% 22. bradycardia on amiodarone Social History: a 3-pack-per-day tobacco history but quit in and an overall Family History: Physical Exam: Gen-Lethargic appearing man on strecher. Will occasionally look Does not answer any questions. reactive to light. Anicteric sclera. Right subcalvian triple Cardiac- RRR. decreased breath sounds on the left. sounds. in place with no erythema or discharge. movement of his limbs. Positive clonus. Downgoing toes Pertinent Results: 09:35PM TYPE-ART TEMP-37.8 RATES-25/ TIDAL VOL-450 -ASSIST/CON INTUBATED-INTUBATED 09:35PM K+-3.8 09:17PM CK(CPK)-33* 09:17PM CK-MB-4 cTropnT-0.13* 06:45PM TYPE-ART TEMP-38.5 RATES-25/0 TIDAL VOL-450 -ASSIST/CON INTUBATED-INTUBATED 02:44PM TYPE-ART TEMP-38.3 RATES-25/ TIDAL VOL-485 AADO2-589 REQ O2-95 INTUBATED-INTUBATED VENT-CONTROLLED 11:54AM GLUCOSE-77 UREA N-28* CREAT-0.6 SODIUM-148* 11:54AM CK-MB-4 cTropnT-0.16* 09:05AM TYPE-ART RATES-/24 PO2-401* PCO2-88* PH-7.25* 07:07AM TYPE-ART O2-100 PO2-439* PCO2-107* PH-7.21* 06:20AM URINE HOURS-RANDOM 06:20AM URINE UHOLD-HOLD 06:20AM URINE COLOR-Yellow APPEAR-Clear SP -1.019 06:20AM URINE BLOOD-SM NITRITE-NEG PROTEIN-

500 LEUK-NEG 06:20AM URINE RBC-0-2 WBC-0-2 BACTERIA-NONE YEAST-NONE 05:36AM O2 SAT-84 05:34AM GLUCOSE-186* UREA N-28* CREAT-0.5 SODIUM-148* 05:34AM ALT(SGPT)-73* AST(SGOT)-50* CK(CPK)-32* ALK 05:34AM LIPASE-18 05:34AM CK-MB-NotDone cTropnT-0.08* 05:34AM ALBUMIN-3.5 CALCIUM-8.8 PHOSPHATE-3.1 05:34AM WBC-14.2* RBC-3.23* HGB-8.7* HCT-29.1* 05:34AM NEUTS-93.2* BANDS-0 LYMPHS-2.4* MONOS-4.1 05:34AM HYPOCHROM-2+ ANISOCYT-NORMAL POIKILOCY-1+ ENVELOP-2+ 05:34AM PLT COUNT-332 05:34AM PT-16.0* PTT-28.5 INR(PT)-1.7 05:30AM LACTATE-1.3 studies: changes but no major changes since previous studies. CXR- Stent projecting over the right brachiocephalic vein. Right contours toward the right consistent with previous which could represent pleural thickening or small left pleural lung which appear stable. No new left pneumo or focal . evidence of hydrocephalus. No evidence of a major CVA. US upper extremity:IMPRESSION: 2) Nonocclusive thrombus in the distal left brachial veins Transabdominal ultrasound examination was performed. The at the neck of the gallbladder. The gallbladder wall is located at the fundus of the gallbladder. No intra or duct is not dilated and measures 3 mm. Flow in the portal vein focal abnormality. IMPRESSION: Thickened gallbladder wall with a stone in the clinical setting, these findings may be consistent with cholecystitis. Gallbladder wall thickening may also be produced concern for cholecystitis, a HIDA scan may be performed for TECHNIQUE: CT images of the chest without the administration of COMPARISON: and . FINDINGS: pneumonectomy and mediastinal shift towards the right. There is There is a left pleural effusion. The patient is intubated with right brachiocephalic vein. The heart demonstrates coronary appears prominent measuring 3.5 cm. Lung window images demonstrate multifocal nodular opacities seen cavitation are identified within these nodules. Atelectasis is or pneumothorax. Septal thickening is seen throughout the left again seen, though slightly difficult to discern given the the study from , this nodule was clearly seen and appears segmental level within the left lung. Images of the upper abdomen demonstrate high-density material to sludge. A percutaneous gastrostomy tube is seen within the abdomen is unremarkable other than arterial calcifications. The throughout the thoracic spine. IMPRESSION: entire left lung. These most likely represent aspiration CT appearance. 3) Left upper lobe nodule seen on the prior study of 4) Probable sludge within the gallbladder. hypertension. 65 y/o man with PMH significant for squamous cell lung CA, type with mental status change and hypotension. ID/sepsis blood pressure, lactate 1.3. His blood pressure in the ED was blood pressure on the right was found to be normal and pressors blood culture, urine culture and cath tip culture negative on line to the right femoral. Chest CT was consistent with He remained afebrile and no pressors required throughout the His mental status improved with decreasing CO2 and also with at baseline in 70s. CT head was negative. Narcotics was taken Patient's duragesic patch was removed in the ED. anemia/coagulation and SVC clot and also atrial fibrillation for which he was on admission because he had blood oozing from his trach and foley. subclavian but the artery was puctured. His right femoral artery The 2 arterial puncture was tamponaded and there was no revealed DVT in left arm for which he was started on heparin arterial line site. Heparin drip was then stopped and he was since then. On discharge, coumadin was not restarted. It should . He presented intially with transaminitis likely from done which showed gallstone at neck of GB, no distension, tenderness respiratory: post tracheostomy. During his past admission there was concern recommended keeping the cuff pressures low with a cuff leak to foam-filled trach (tube) in the future if the cuff leak assist control ventilation. Pt with mildly elevated Na at 148. This is most needs. He recieved free water through G tube. Cardiac complain during this hospitalization. Cardiac enzymes were for atrial fibrillation. He is to avoid beta blockers and . Patient was continued on standing 8U glargine and sliding scale Patient's family reports that he is extemely anxious at anxiety 5mg hs, 2mg 8am/2pm, 1mg tid/prn, and paxil. His pain is severely worsened in the past with ativan. Would avoid further concern regarding narcotic overdose. He was on prn morphine. He had Kinair bed FEN He had picc line on discharge code meeting) 1. Xopenex 1.2 mg inhaled Q4H 3. Haldol 1 mg 0800 and 1400 5. Casec powder 2 tablespoons TID 7. Ambien 5 mg QHS 9. Lactulose 20 gm daily 1. Doxepin 3. Oxycontin 5. Ativan 11. Colace 100 mg 13. Theravite liquid 5 ml daily 15. Paxil 20 mg daily 17. Vitamin D 800 units daily 19. ASA 325 mg daily 21. Humulin SS 23. Xopenex 1.25 mg Q4H PRN 25. Haldol 1 mg Q8H PRN 27. Amiodarone 400 mg daily 1. Fluticasone Propionate 110 mcg/Actuation Aerosol : Two (2) 2. Lactulose 10 g/15 mL Syrup : Thirty (30) ML PO TID (3 3. Glycerin (Adult) 3 g Suppository : One (1) Suppository 4. Docusate Sodium 150 mg/15 mL Liquid : One Hundred (100) mg 5. Acetaminophen 325 mg Tablet : 1-2 Tablets PO Q4-6H (every 6. Bisacodyl 5 mg Tablet, Delayed Release (E.C.) : Two (2) 7. Paroxetine HCl 20 mg Tablet : One (1) Tablet PO DAILY 8. Aspirin 325 mg Tablet : One (1) Tablet PO DAILY (Daily). 9. Lansoprazole 30 mg Capsule, Delayed Release(E.C.) : One 10. Amiodarone HCl 200 mg Tablet : Two (2) Tablet PO DAILY 11. Albuterol 90 mcg/Actuation Aerosol : Six (6) Puff 12. Ipratropium Bromide 18 mcg/Actuation Aerosol : Six (6) 13. Haloperidol Lactate 2 mg/mL Concentrate : Five (5) mg PO 14. Haloperidol Lactate 2 mg/mL Concentrate : One (1) mg PO BID (2 times a day): at 8AM and 2PM. 15. Heparin Sodium (Porcine) 5,000 unit/mL Solution : One (1) 16. Zolpidem Tartrate 5 mg Tablet : Two (2) Tablet PO HS (at 17. Nitroglycerin 0.3 mg Tablet, Sublingual : One (1) Tablet, 18. Morphine Sulfate 10 mg/5 mL Solution : Five (5) mg PO Q6H 19. Piperacillin-Tazobactam 4.5 g Recon Soln : 4.5 gm 20. Insulin Glargine 100 unit/mL Solution : Eight (8) unit Discharge Disposition: Rehab Center - Discharge Diagnosis: overdose +/- aspiration pneumonia stable Discharge Instructions: more shortness of breath,

confusion, hypotension, chest pain, PLease follow up with doctors . Coumadin has been discontinued because you had significnant insertion. This should be restarted at a lower dose in days narcotic overdose.

LDA output: Similarity Score-0.82827294 Jaccard Score-0.793

Sex: M Service: 1. Attempted suicide drug overdose. 3. Herniated disk. Russian gentleman with a past medical history significant for who was subsequently brought to the Emergency Room on with a chief The patient was found in the hotel room unresponsive to Diphenhydramine 32 tab x 3, with a total 4800 mg total Room, the patient was given Narcan and was subsequently coma scale of 3 at which time the patient was times one. The patient was then toxicology screened for his was subsequently transferred to the Medical Intensive Care failure secondary to attempted overdose with Benadryl and complications and was found to be awake and responsive and to to be .... most likely secondary to aspiration cultures were sent. The patient denied any suicidal ideation precipitant for his suicide attempt. He noted that he treated with Prozac in the past which had helped his and so he was changed to Wellbutrin. He had only been on patient gives a reasonable willingness to pursue medical pain, nausea, vomiting, or diarrhea at the time. PAST MEDICAL HISTORY: Depression. Suicide attempt in time. Second suicide overdose attempt in for which he . He was hospitalized. He has a history of herniated PAST SURGICAL HISTORY: None. SOCIAL HISTORY: He is in a third marriage of five years language at . He admits to occasional overdose. FAMILY HISTORY: No known family psychiatric history to date MEDICATIONS: He was on Zantac, subcue Heparin, Levofloxacin, ALLERGIES: NO KNOWN DRUG ALLERGIES. PHYSICAL EXAMINATION: Vitals signs: He had a T-max of 101??????, respirations 25, blood pressure 108/55. General: He was in moderate to flat affect. Chest: Clear with good breath Heart: Regular, rate and rhythm. No murmurs, rubs or gallops heard. Abdomen: Soft, nontender, nondistended. Positive bowel sounds. Extremities: No clubbing, cyanosis LABORATORY DATA: He had a chest x-ray performed which showed cultures were negative to date. HOSPITAL COURSE: This was a 59-year-old man admitted to the attempted overdose with Benadryl, Tylenol 3, and OxyContin. Care Unit and medically cleared for an acute myocardial infarction by enzymes. He was transferred to the medical pneumonia and for psychiatric placement. Pulmonary: Status post extubation on , the patient need of supportive oxygen. Chest x-ray was negative for any fever was due to either aspiration pneumonia or chemical and Flagyl for a 10-day course. Infectious disease: The patient was with fevers of patient will continue on Levofloxacin and Flagyl. The Cardiovascular: The patient ruled out for myocardial cleared of any cardiac issues. Heme: The patient had a low hematocrit to 30.5 on the hematocrit of 33.5 and 32.5 respectively. The patient was hemorrhage. GI: The patient was tolerating a regular diet with no nausea Psychiatry: The patient is with a history of depression and patient was medically cleared and is now awaiting psychiatric a 1:1 sitter, and he is covered with Ativan 0.5 to 1.0 mg DISPOSITION: The patient is medically stable and cleared.

LSTM output: Similarity Score-0.977 Jaccard Score-0.844

Date of Birth: Sex: M Service: HISTORY OF PRESENT ILLNESS: This is a 77-year-old man with a was a long-time smoker who developed recent hemoptysis 2 to 3 the apical segment of the right upper lobe that was confirmed unremarkable metastatic survey, including a PET scan. He catheterized at the end of . His pulmonary was functionally quite well and was able to walk flights of was giving him obstructive symptoms. It was felt at this functional lung parenchyma. The plan - after multiple mediastinoscopy, and if the nodes were negative to proceed to before, from a cardiology perspective. He did receive except for a right apical haziness that represented the PHYSICAL EXAMINATION ON ADMISSION: The patient was noted to was in no apparent distress, and was comfortable, and was a nourished but not obese. He was normocephalic, atraumatic. was clear. His neck was supple without lymphadenopathy or no supraclavicular or axillary lymphadenopathy. His lungs display an occasional expiratory wheeze. He was in a regular abdomen was nondistended with normal active bowel sounds and hernias noted. There was no inguinal lymphadenopathy. He HOSPITAL COURSE: On the morning of , the operating room that morning where he underwent bronchoscopy, upper lobectomy. The operation proceeded without patient was extubated in the operating room and taken to the well controlled with an epidural catheter that was managed by instability. His cardiac enzymes were followed catheterization. These were all unremarkable, and his subsequent sets. On postoperative day 1, the patient was able to be out of the intensive care unit. The patient continued to water seal at midnight on the night on postoperative day 1. time on serial chest x-rays, and these chest tubes were postoperative day 3 and the anterior tube removed on There was a very, very small residual right apical but it was deemed clinically insignificant, and the patient on multiple sessions, it was deemed that he would benefit morning of discharge were 95% on 2 liters without any oxygen saturation on room air at this point - post ambulation doing well and was very comfortable at this time. His pain was deemed fit for discharge. DISCHARGE DIAGNOSES: Right upper lobe mode; status post lobectomy; coronary artery disease; status post pulmonary disease; gout; benign prostatic hyperplasia; hernia repairs in the distant past; tonsillectomy. MAJOR SURGICAL OR INVASIVE PROCEDURES: Bronchoscopy, endotracheal intubation, arterial line placement, peripheral DISCHARGE CONDITION: Good. DISCHARGE MEDICATIONS: Atorvastatin 10 mg p.o. daily, Lasix mg p.o. daily, Advair Diskus, finasteride 5 mg p.o. daily, Percocet 5/325 one to 2 tablets p.o. q.4-6h. as needed for daily, prednisone 5 mg to take 1 dose on and then to prevent constipation. DISPOSITION: The patient to be discharged to home.

V. Results —–>

The testing data has 10 documents. The scores for top five documents are shown.

VI. Conclusions

We conclude by saying that the Neural Network outperform the other models due to the transformation of text into vector space which is further encoded. Among the other models, LDA model outperformed the others.

VII. Future Work

The system built currently is a prototype and hence, scaling and using the system as a service is not implemented as of now. Gensim library used for the topic modelling supports distributed computing and one of the popular open-source project ElasticSearch can be used for tf-idf computations. This helps in building a production system which can automatically infer the parameters, build a model and get the similar documents in real-time.

References

[1] Github link to Doc2Vec code and outputs
[2] Github link to Tf-Idf code and outputs
[3] Github link to LDA code and outputs
[4] Github link to LSTM code and outputs

Document No.	Similarity Measure	SimilarityScor*e*	Jaccar*d*Score
1	Tf-Idf	0.4281	0.7651
		0.4256	0.8466
		0.4196	0.7731
		0.4177	0.7517
		0.4158	0.7816
	LDA	0.9034	0.8164
		0.8910	0.7696
		0.8907	0.7848
		0.8879	0.7695
		0.8870	0.7895
	LSTM	0.9684	0.7625
		0.9677	0.7930
		0.9667	0.7837
		0.9666	0.7638
		0.9663	0.7823
2	Tf-Idf	0.8716	0.5891
		0.8610	0.6545
		0.8595	0.6325
		0.8559	0.6365
		0.8545	0.6250
	LDA	0.9472	0.6013
		0.9224	0.6796
		0.9115	0.7012
		0.8999	0.6809
		0.8988	0.6743
	LSTM	0.9850	0.7133
		0.9833	0.5891
		0.9830	0.6529
		0.9830	0.6862
		0.9822	0.6250
3	Tf-Idf	0.8686	0.6067
		0.8582	0.6755
		0.8557	0.6124
		0.8550	0.6601
		0.8530	0.7769
	LDA	0.9679	0.6633
		0.9620	0.6397
		0.9571	0.6546
		0.9503	0.6406
		0.9490	0.6841
	LSTM	0.9725	0.7132
		0.9720	0.6067
		0.9718	0.7917
		0.9716	0.7459
		0.9713	0.7807
4	Tf-Idf	0.3126	0.7828
		0.3094	0.7949
		0.2988	0.8101
		0.2976	0.7635
		0.2975	0.7830
	LDA	0.9311	0.8129
		0.8749	0.8065
		0.8727	0.8004
		0.8708	0.7941
		0.8644	0.7992
	LSTM	0.9686	0.8498
		0.9663	0.7967
		0.9662	0.7928
		0.9661	0.8057
		0.9657	0.8226
5	Tf-Idf	0.3222	0.7599
		0.3122	0.7708
		0.3111	0.7732
		0.3106	0.7580
		0.3093	0.7555
	LDA	0.9151	0.7599
		0.9089	0.7700
		0.9028	0.7832
		0.8991	0.7905
		0.8918	0.7911
	LSTM	0.9727	0.7833
		0.9723	0.8216
		0.9717	0.8234
		0.9712	0.8136
		0.9711	0.8056