SMOKING CESSATION AND NLP

Derrick McCray

ELECTRONIC HEALTH RECORDS

- Promises to improve patient care and streamline operations
- Challenges aplenty including: cost, training and workflow concerns
- The challenge of unstructured data
- Meaningful Use
- Various notes with unstructured data

782836641 DH
9369592
01111
974771
3/15/2002 12:00:00 AM
OB Discharge Summary
Signed
DIS
Report Status :
Signed OB EMR L and D
DISCHARGE SUMMARY
NAME :
XIEACASS BETHCONRI BALLOON
UNIT NUMBER :
242-36-71
ADMISSION DATE :
20020315
DISCHARGE DATE :
20020318
PRINCIPAL DISCHARGE DIAGNOSIS :
Vaginal Delivery With First Degree Laceration
ASSOCIATED DIAGNOSES :
Advanced Maternal Age ; Depression , history of ; Hepatitis C Antibody Positive ; Polyneuropathy , history of ; Problems With Abuse , history of ; Rh Nonsensitization ; Stopped Smoking This Pregnancy , history of
PRINCIPAL PROCEDURE OR OPERATION :
Spontaneous Vertex Vaginal Delivery
ASSOCIATED PROCEDURES OR OPERATIONS :
POSTPARTUM DIAGNOSTIC PROCEDURES :
None
POSTPARTUM THERAPEUTIC PROCEDURES :
None
HISTORY AND REASON FOR HOSPITALIZATION :
Active Labor
PHYSICAL EXAMINATION :
HEIGHT NORMAL 66 HEENT NORMAL MOUTH NORMAL NECK NORMAL BREASTS NORMAL NIPPLES NORMAL CHEST NORMAL COR NORMAL ABDOMEN NORMAL EXTREM NORMAL SKIN mottled on both lower extremities .
NODES NORMAL VULVA NORMAL VAGINA NORMAL CERVIX NORMAL OS NORMAL ADNEXAE NORMAL UTERUS NORMAL UTERINE SIZE IN WEEKS 15 HOSPITAL COURSE (include complications if any) :
This 42 year old Gravida 3 Para 2002 was admitted to the Naliheall County Memorial Hospital Obstetrical service on 03/15/2002 at 10:04 pm for the indication (s) :
active labor .
She delivered a 2809 gram male infant on 03/16/2002 at 02:50 am with apgar scores of 7 and 9 at one and five minutes respectively at 38.0 weeks gestation via spontaneous vertex vaginal delivery .
During her labor she encountered the following complication (s) :
none .
During her delivery she encountered the following complication (s) :
. Postpartum she encountered the following complication (s) :
depression .
She was discharged on 03/18/2002 at 12:45 pm in good condition .
DISCHARGE ORDERS (medications instructions to patient , follow-up care) :
DISCHARGE ACTIVITY :
No Restrictions
DISCHARGE DIET :
No Restrictions
POSTPARTUM DISPOSITION :
Home Under Care Of Shingle Geabell Hospital
POSTPARTUM CARE SITE :
Dh Ob
POSTPARTUM RETURN APPOINTMENT (DAYS) :
42
BREAST FEEDING AT DISCHARGE :z
No
POSTPARTUM RH IMMUNE GLOBULIN :
Given
POSTPARTUM MEASLES / MUMPS/RUBELLA VACCINE :
Not Indicated
MEDICATION (S) ON DISCHARGE :
Multivitamins And Folate (Stuart Prenatal With Folate) ; Fluoxetine (Prozac)
Electronically Signed :
Polle , Nella O 03/18/2002 9:01:18 PM
[report_end]

NATURAL LANGUAGE PROCESSING

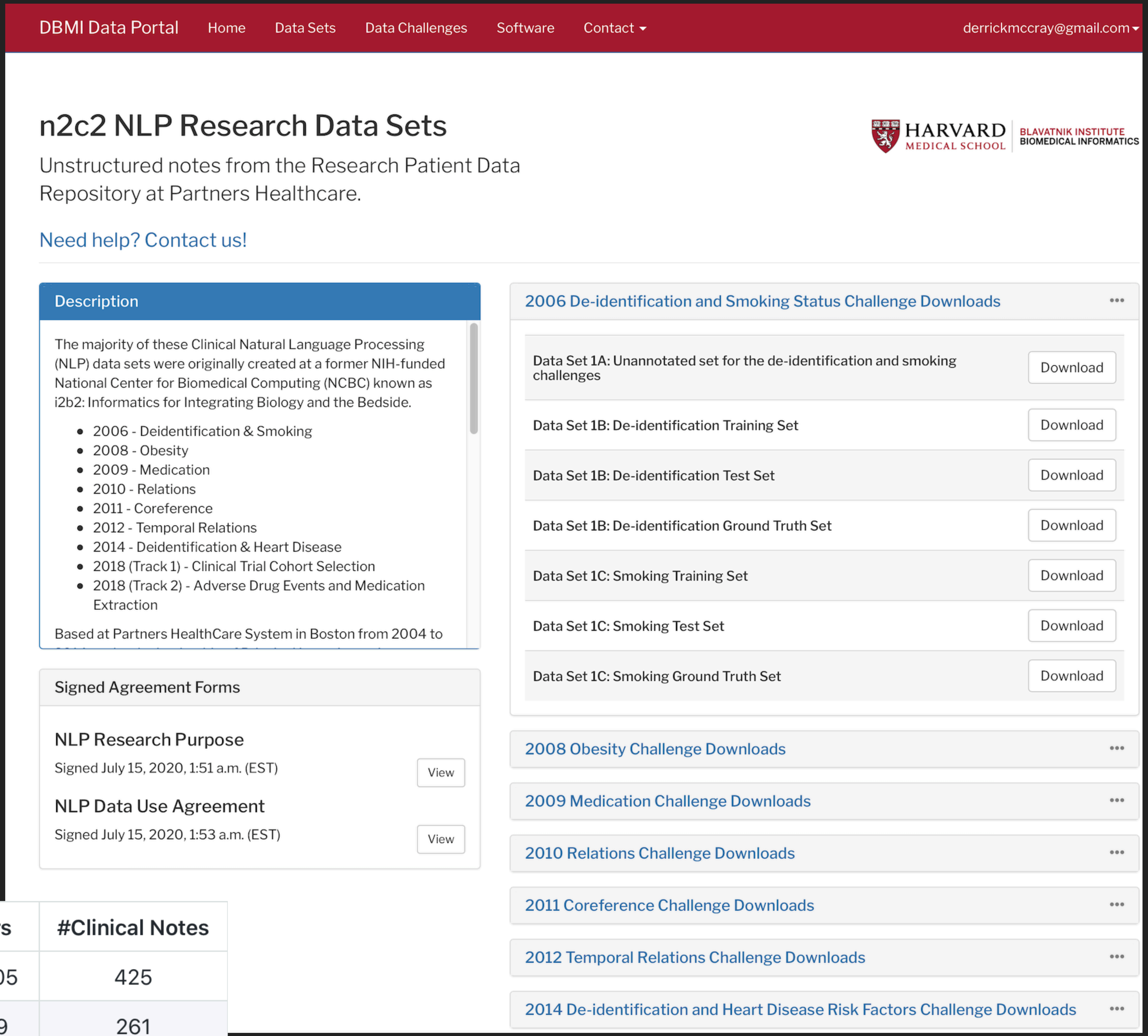
- NLP allows a computer to read and understand human language.
- This promises to be a perfect fit in identifying clinical information from completely unstructured notes
- Cutting edge research

SMOKING CESSATION PROJECT

- Pilot for a proof of concept for NLP in healthcare
- Scope
- Goal? Identify smokers vs non-smokers
- What are the various categories (i.e. current, past, quit longer than a year)
- Why? Smoking cause wholly preventable diseases.

DATASET

- i2b2
- n2c2
  - deidentification
  - structure
- emrQA
  - process
  - structure



Datasets	QA pairs	QL pairs	#Clinical Notes
i2b2 relations (concepts, relations, assertions)	1,322,789	1,008,205	425
i2b2 medications	226,128	190,169	261
i2b2 heart disease risk	49,897	35,777	119
i2b2 smoking	4,518	14	502
i2b2 obesity	354,503	336	1,118
emrQA (total)	**1,957,835**	**1,225,369**	**2,425**

EXPLORATORY DATA ANALYSIS

- Python, Pandas, Numpy, Jupyter, Scikit Learn, Anaconda
    - powerful stack
    - Duplicate Checking
    - JSON conversion
    - Unkown Status
    - Record Counts
    - Time intensive

ANNOTATION

- Annotation
  - using emrQA classification
  - own annotation process

```
In [117]: # get distinct values of the emrQA_class from the data
          clin_notes.groupby(by='emrQA_class').count()['emrQA_smoker']

Out[117]: emrQA_class
          CURRENT SMOKER     46
          NON-SMOKER         82
          PAST SMOKER        47
          SMOKER             12
          UNKNOWN           315
          Name: emrQA_smoker, dtype: int64
```

```
In [47]: # the actual sentence from the discharge note that was wrongly classified by the emrQA project's classification
         # fyi this function is defined later in this notebook
         find_smoking_specific_text(cna['note_text'], 660)

Out[47]: ['He is a heavy smoker and drinks 2-3 shots per day at times .']
```

note_id	note_text	emrQA_class	emrQA_smoker	emrQA_class_num	emrQA_smoker_num	DM_SMOKER_YN	DM_SMOKER_CLASS	SMOKER
660	156406283 HLGMC 7213645 64723/51cy 5/28/1993 12:00:00 AM Discharge Summary Unsigned DIS Report Status : Unsigned ADMISSION DATE : 5-28-93 DISCHARGE DATE : 6-4-93 HISTORY OF PRESENT ILLNESS : The patient is a 58 year old right hand dominant white male with a long history of hypertension , changed his medic The patient has a history of adult onset diabetes mellitus , ankylosing spondylitis , status post myocardial infarction Briefly , he was talking to a friend at 5:30 p.m. the day prior to admission , when he had to grab his locker and sit d His voice became slurred and he had a mild central dull headache . He was unable to move the left side of his body and felt numb on that side . He was taken to Wayskemedcalltown Talmi and transferred to Heaonboburg Linpack Grant Medical Center with a c His blood pressure was 220/110 there . He denies any visual symptoms or cortical-type symptoms . He is a heavy smoker and drinks 2-3 shots per day at times . MEDICATIONS ON ADMISSION : Vasotec 40 mg q.day , Soma 1 tablet q.day , Demerolprn , Clonidine . ALLERGIES : The patient has no known drug allergies . PAST MEDICAL HISTORY : As described above . FAMILY HISTORY : The family history is positive for diabetes mellitus , positive for cancer . SOCIAL HISTORY : The patient lives with two people in Cinglendda . PHYSICAL EXAMINATION : On physical examination , patient is in no acute distress , afebrile , blood pressure 134/80 , heart rate 80 and regula Cardiovascular exam : regular rate and rhythm with a I/VI systolic ejection murmur . His lungs were clear to auscultation and percussion	PAST SMOKER	NON-SMOKER	2.0	0.0	Y	CURRENT	1

MODEL DEVELOPMENT

- Discharge Summary Note Text
- Bag of words approach
- Logistic Regression vs Naive Bayes
- Accuracy
  - Null Accuracy to beat of 70%

```
In [33]: # Calculate null accuracy.
         print('Percent Current Smoker:', y_test.mean())
         print('Percent Not Current Smoker:', 1 - y_test.mean())

         Percent Current Smoker: 0.2978723404255319
         Percent Not Current Smoker: 0.7021276595744681
```

```
In [35]: #choosing the best model and best ngram parameter
         for x in range(1,11):
             v = CountVectorizer(ngram_range=(x,10), stop_words='english')
             print(f'ngram {x} - 10')
             print(model_accuracy_test(v, X_train, X_test, y_train, y_test))
             print('')

         ngram 1 - 10
         ('Features: ', 657879)
         {'Naive Bayes': 0.7446808510638298, 'Logistic Regression': 0.7446808510638298, 'KNN': 0.7446808510638298}

         ngram 2 - 10
         ('Features: ', 648157)
         {'Naive Bayes': 0.7446808510638298, 'Logistic Regression': 0.723404255319149, 'KNN': 0.723404255319149}

         ngram 3 - 10
         ('Features: ', 598657)
         {'Naive Bayes': 0.7659574468085106, 'Logistic Regression': 0.723404255319149, 'KNN': 0.723404255319149}

         ngram 4 - 10
         ('Features: ', 532374)
         {'Naive Bayes': 0.7021276595744681, 'Logistic Regression': 0.723404255319149, 'KNN': 0.723404255319149}

         ngram 5 - 10
         ('Features: ', 459569)
         {'Naive Bayes': 0.574468085106383, 'Logistic Regression': 0.7021276595744681, 'KNN': 0.7021276595744681}

         ngram 6 - 10
         ('Features: ', 384261)
         {'Naive Bayes': 0.5957446808510638, 'Logistic Regression': 0.7021276595744681, 'KNN': 0.7021276595744681}

         ngram 7 - 10
         ('Features: ', 307961)
         {'Naive Bayes': 0.7021276595744681, 'Logistic Regression': 0.7021276595744681, 'KNN': 0.7021276595744681}

         ngram 8 - 10
         ('Features: ', 231171)
         {'Naive Bayes': 0.7021276595744681, 'Logistic Regression': 0.7021276595744681, 'KNN': 0.7021276595744681}

         ngram 9 - 10
         ('Features: ', 154159)
         {'Naive Bayes': 0.7021276595744681, 'Logistic Regression': 0.7021276595744681, 'KNN': 0.7021276595744681}

         ngram 10 - 10
         ('Features: ', 77081)
         {'Naive Bayes': 0.723404255319149, 'Logistic Regression': 0.7021276595744681, 'KNN': 0.7021276595744681}
```

- First approach with basic untuned models 74-76% accuracy

NAMED ENTITY RECOGNITION

- Thought the model could be improved be being more targeted
- enter spaCy, scispaCy and med7
- medication list
- problem list

MEDICATION LIST & PROBLEM LIST

- scispaCy biomedical model
- med7 drug model
- was able to parse the note to get a medication list and problem list per patient discharge note

DISCHARGE DIAGNOSIS :

Basilar artery stenosis DISEASE with basilar thrombosis DISEASE and " top of the basilar " syndrome DISEASE .

Includes recent infarct DISEASE to pons , mid brain , left thalamus , bilateral temporal lobes and left visual cortex , on Coumadin CHEMICAL .

The patient is do not resuscitate at family 's request .

Left lower lobe pneumonia DISEASE , resolving now on oral antibiotics .

Gallstones DISEASE , thought to be inactive .

Improving liver enzyme elevation .

Anemia DISEASE , discharge hematocrit 28 , not iron CHEMICAL deficient .

Low potassium CHEMICAL , thought due to Gentamicin CHEMICAL .

any problems from this .

click to unscroll output; double click to hide

Her sodiums were never below 130 .

She does well with tube feeds .

Her tube feeds orders are as follow :

full strength Replete with fiber at 70 cc DOSAGE . per hour .

In addition , the patient gets 250 cc. of juice (not water) three times a day FREQUENCY .

She also gets Lactinex granules DRUG three packages in each bottle of tube feeds .

She also gets Metamucil DRUG one DOSAGE teaspoon with the first bolus DOSAGE of juice each day FREQUENCY .

Please note that evaluation by the swallowing therapist , showed that the patient is aspirating at this time , but there is great hope from the nature of her deficit and the good movements of her tongue that normal swallowing should return soon .

FURTHER REFINEMENTS AND ADDITIONAL FEATURES

- Smoking related phrases from the notes
  - TextBlob
- Alcohol sentiment
- problem list
  - lung disease, cancer, copd, asthma
- medication list
  - vanceril
  - inhaler

```
In [46]: def find_smoking_specific_text(d_frame, note_id):
             smoke_list = ['SMOK','TOBACCO','CIG']
             smoke_sents = []

             for s in TextBlob(d_frame.loc[note_id]).sentences:
                 for t in smoke_list:
                     if t in str(s).upper():
                         smoke_sents.append(str(s))

             return list(set(smoke_sents))

In [45]: #An example smoking related phrase
         find_smoking_specific_text(X,515)

Out[45]: ['She cut down dramatically on smoking two years ago , but has continued to smoke , although , very recently , she
         admits to only " two puffs " per day .']
```

note_id	note_text	emrQA_class	emrQA_smoker	emrQA_class_num	emrQA_smoker_num	DM_SMOKER_YN	DM_SMOKER_CLASS	smoker	smoking_text	med_list	prob_list	alcohol_sentimer	lung_disease	copd	cancer	asthma	vanceril	inhaler
660	156406283 HLGMC 7213645 64723/51cy 5/28/1993 12:00:00 AM Discharge Summary Unsigned DIS Report Status : Unsigned ADMISSION DATE : 5-28-93 DISCHARGE DATE : 6-4-93 HISTORY OF PRESENT ILL The patient is a 58 year old The patient has a history of Briefly , he was talking to a His voice became slurred an He was unable to move the He was taken to Wayskeme His blood pressure was 220 He denies any visual sympt He is a heavy smoker and d MEDICATIONS ON ADMISS Vasotec 40 mg q.day , Som ALLERGIES : The patient has no known d PAST MEDICAL HISTORY : As described above . FAMILY HISTORY : The family history is positive SOCIAL HISTORY : The patient lives with two pe PHYSICAL EXAMINATION : On physical examination , p Cardiovascular exam : regular rate and rhythm with His lungs were clear to ausc The abdomen was soft and Back and neck were stiff an On neurological examinatio Able to describe two routes The cranial nerves showed Motor examination showed Left leg could flex 2/5 and w On sensory examination , h The finger-to-nose was okay On reflex examination , 2 on Three on the left knee , 4 on	PAST SMOKER	NON-SMOKER	2.0	0.0	Y	CURRENT	1	He is a heavy smoker and drinks 2-3 shots per day at times .	Percodan;Soma;Flexeril;Nifedipine;Micronase;Demerolprn;Clonidine;Vasotec;Percocet;Valium;Aldomet	Soma;Sore;Inferior Myocardial Infarction;Ankylosing Spondylitis;Finger-To-Nose;Cancer;Numb;Muscle Spasms;Non-Tender;Hypertension;Diabetes Mellitus;Drug Allergies;Primary Hemisensory Loss;Central Dull Headache;Myocardial Infarction;Right Thalamic Hemorrhage;Bruits;Td;Allergic;Distortion	-0.2	0	0	1	0	0	0

ACCURACY

- Features: Smoking Text, Lung Disease, COPD, Cancer, Asthma, Inhaler
- Ngram Range tuning
- Logistic Regression eventually predicted the most accurately.
- 89% Accuracy

In [82]:

```python
text_cols = t_feats

X = cna[feats]
y = cna.smoker

# Train, test split
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=99)

#Count Vectorizer
vect = CountVectorizer(ngram_range=(1, 7), stop_words='english')
X_train_smoking_dtm = vect.fit_transform(X_train.smoking_text)
X_test_smoking_dtm = vect.transform(X_test.smoking_text)

#Add features for train sparse matrix
extra = sp.sparse.csr_matrix(X_train.drop(text_cols, axis=1).astype(float))
extra.shape
X_train_dtm_extra = sp.sparse.hstack((X_train_smoking_dtm, extra))

#Add features for test sparse matrix
extra = sp.sparse.csr_matrix(X_test.drop(text_cols, axis=1).astype(float))
X_test_dtm_extra = sp.sparse.hstack((X_test_smoking_dtm, extra))

#Combine sparse matrices
X_train_dtm = sp.sparse.hstack((X_train_smoking_dtm, X_train_dtm_extra))
X_test_dtm = sp.sparse.hstack((X_test_smoking_dtm, X_test_dtm_extra))

# Use Logistic Regression to predict smoking status
logr = LogisticRegression()
logr.fit(X_train_dtm, y_train)
y_pred_class = logr.predict(X_test_dtm)

print(metrics.accuracy_score(y_test, y_pred_class))
```

```
0.8936170212765957
```

TAKEWAYS AND IMPROVEMENTS

- Unstructured Text can certainly be used to improve patient care and quality
- This particular model did stumble on texts like the following

```
In [83]: cna.loc[540]['smoking_text']
Out[83]: "SOCIAL HISTORY :\nThe patient 's social history is notable for a heavy smoking history ..She is interested in quit
         ting smoking ."
```

- Feature correlations for NLP?
- Clinical Datasets

QUESTIONS?