IDS 560 – Final Project Report

DuPage Medical Group

WE CARE FOR YOU

By,
Koffi Agbavo
Sree Pranavi Kanduri
Vivek Reddy Devi
Jonathan Nichols

Report Outline

Business Motivation:

Problem Statement:

DuPage Medical group (DMG) is one of the largest and most successful independent multi-specialty physician groups in Illinois. Working with large files of medical records, DMG faces issues with mapping the ICD-10 codes (International Classification of Diseases) with the respective medical condition. This results in a potential loss of revenue for the company. DMG makes money by treating its patients, billing their visits and getting reimbursed by the insurance company, the government, or the patients themselves for the service provided. Hence, it is vital for DMG to own a solution that helps identify patients' actual medical conditions they are being treated for and bill them accurately. Pertaining to our project, we have worked on the medical condition: Atherosclerosis.

Research Questions:

❏ How is atherosclerosis identified?

❏ What does the unstructured clinical data comprise of?

❏ What are the target words to identify the signs or symptoms of atherosclerosis?

❏ What is the impact of our solution on the client's business?

Overall Approach:

An NLP solution is used to parse the information to map the medical conditions with the right ICD- 10 code. This could be identified by examining the context of the clinical notes which show the presence or absence of the disease. The project deliverable is the outcome of parsing our data through a PyContextNLP library built with Python, which is integrated into Dupage's SQL database for different stakeholders to query according to their business requirement.

Key Risks:

❏ Misclassification of the different atherosclerosis keywords

❏ Not having comprehensive keyword library (target words) for a greater classification

❏ Integration of our output onto DMG's environment

❏ Confidentiality of data

Project Plan Overview:

Project Sponsor: DuPage Medical Group (DMG) – Ayis Pyrros

Ayis Pyrros is the representative of DMG. He is responsible for defining the business problem, guiding the team, direct supervision, clarifying questions through the project's timeline.

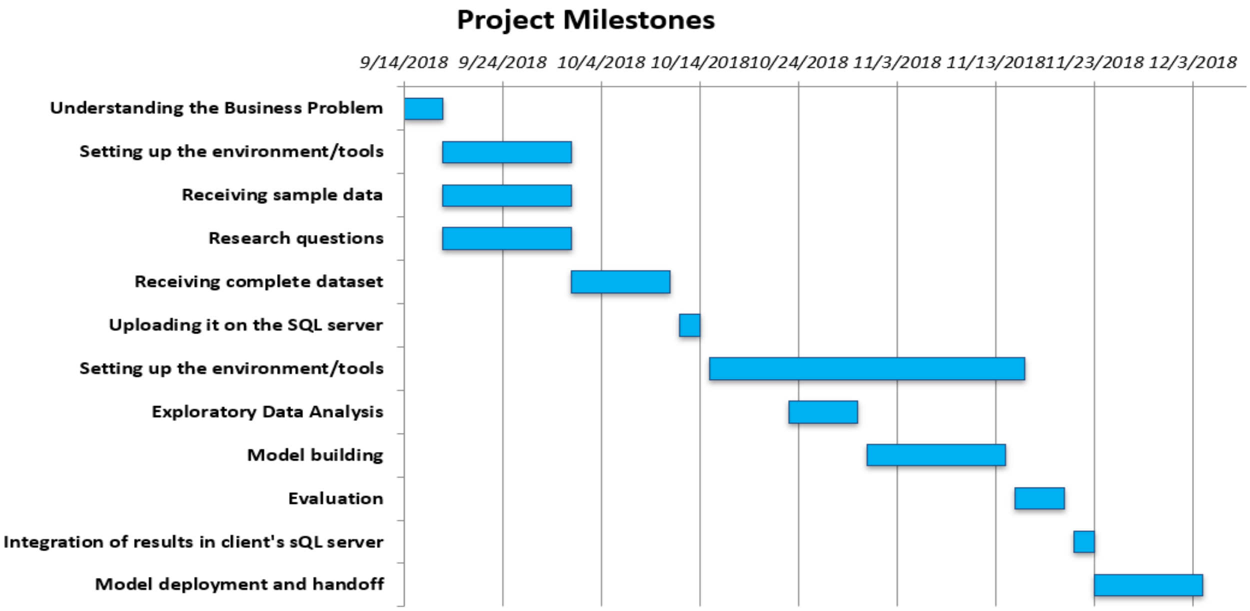
Team Members:

Sree Pranavi Kanduri	Point of Contact, Data Modeling and Testing
Koffi Agbavo	Data Acquisition and preparation
Vivek Reddy Devi	Data Modeling and Testing
Jonathan Nichols	Deployment and Hand-off

Point of Contact responsibilities:

Coordinating team efforts to ensure the goals of the project are met, team communications and scheduling meetings with the project sponsor and course instructor.

Project team responsibilities and Milestones:



**All project and subsidiary management plans were reviewed and approved by the project sponsor.*

Analytic Resources:

Data: CSV and Tab delimited files containing electronic Health Records from Ayis Pyrros
Tools: SQL Server, Jupyter Notebook (Python 3)
Other Applications: Viscosity (VPN), Remote desktop connection
Libraries used in Python: pandas, numpy, pyodbc, pyContextNLP, pyContextNLP.helpers, pyContextNLP.itemData

Data Preparation:

<u>What does our data consist of ?</u>

- ❏ MRN - Medical Record Number
- ❏ Order Proc ID - Patient ID
- ❏ Date - No significance
- ❏ Examination - Area where the x-ray was taken
- ❏ Line - Indexes of rows associated with each individual patient
- ❏ Narrative - Radiologist notes

<u>Steps for data preparation:</u>

- ❏ The TSV files were uploaded onto the SQL server
- ❏ An ODBC connection was established between the SQL server and Python
- ❏ Merged the narrative column by rolling up to (MRN- Patient ID) level
- ❏ Eliminated the duplicate columns (i.e, Same patient IDs)
- ❏ Cleaned the narrative column in all the records to simple plain text
- ❏ Raw text from Narrative column is cleaned to create a Markup function to tag the target and modifiers terms

Descriptive Analysis:

- ❏ Total Records: 3 million rows
- ❏ Number of unique occurrences with “Artery”: 18,086 rows
- ❏ Number of unique occurrences with “Atherosclerosis”: 26,537 rows

Technical Approach:

To improve the identification of the assigned medical condition and categorize them into: Negation, Experiencer, and Temporal Status within the health records, the Python library “PyContextNLP” (an extension of a ConText Algorithm with user-defined modifiers) is used. This helps us decide the presence or absence of atherosclerosis based on the Narrative column from our data. PyContextNLP methodology was chosen as it can derive useful insight from unstructured clinical text data.

The ‘context algorithm’ uses regular expressions to search for trigger words preceding or following the indexed medical condition. A list of targets (medical condition) and modifiers(words that define the context of the medical condition). The final PyConText output consists of the trigger terms with the following categories: negated, hypothetical, historical, experienced.

Example: (Narrative notes) The intracranial segments of the internal carotid arteries as well as the anterior and middle cerebral arteries are patent; minimal ***calcified plaque*** ***is present*** along the pericavernous portion of the internal carotid arteries

Explanation: Here, “calcified plaque” is our target term and “is present” is one of the modifiers in the forward direction which confirms the existence of the disease.

Output: evidence_of_calcplaq, probable_existence

Step-wise Explanation of the Process:

❏ Import the TSV data files on SQL server and conduct the basic exploratory data analysis *Note: CSV files were comma delimited which divided our entire narrative columns into multiple columns hence we asked our sponsor for TSV files*

❏ An ODBC connection from the SQL server to Python was established and an option to input the data directly with a CSV or TSV file was made in the code

❏ Assigned labels to the columns and creating a data frame

- The narrative columns were merged into a single cell at an MRN-Patient ID level, so the context of the entire sentence was not lost
- Only the relevant columns were retrieved as date, line and description were not necessary for our process and for efficient execution of the algorithm
- The narrative column was in an unstructured format and hence we needed to convert into simple plain text which is, ignoring the punctuations, double spaces etc.,
- Two CSV files were generated with the Targets and Modifiers to append to our code *PFA the two files along with the document*
- Created target and modifier items in our code using the github link for the above files
- Convert Modifiers and Targets CSV files into items (Lex, type, regular expressions)
    - Lex or Literal : Context within the notes
    - Type : definite existence, probable negated existence, probable existence, definite negated existence, ambivalent existence
    - Regular Expression (Regex): Nomenclature or different ways the words of interest can be written to search in the text

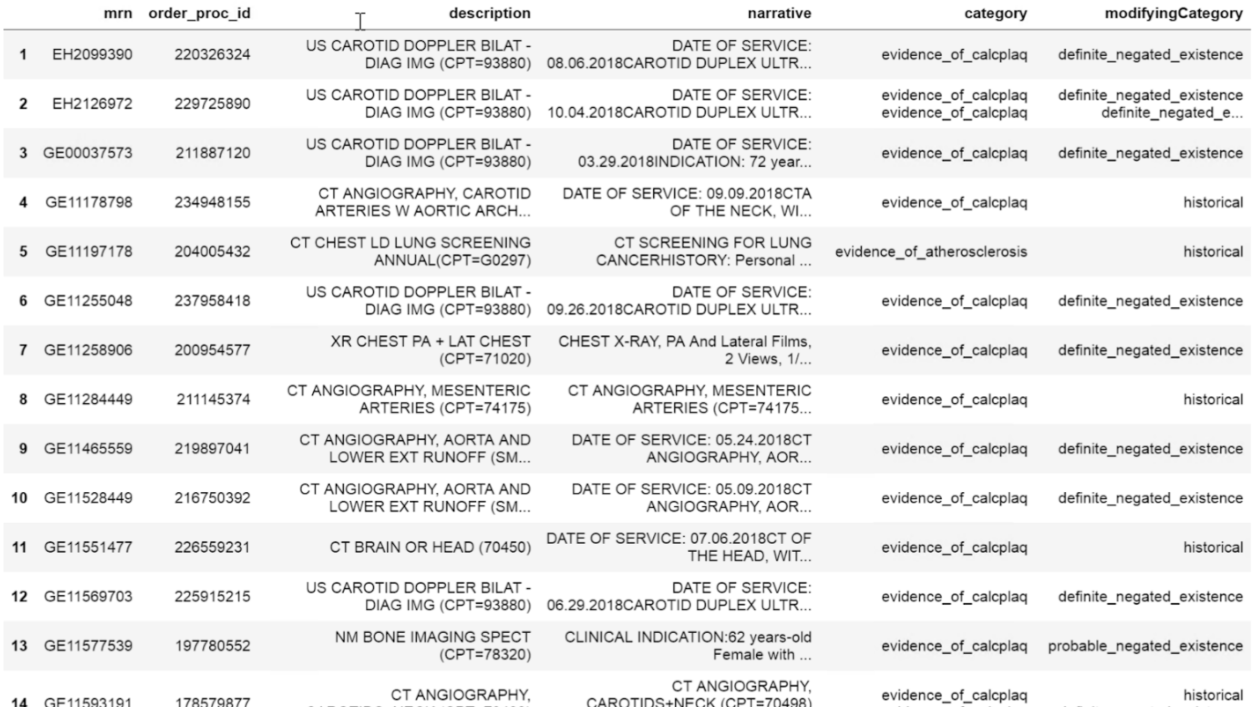
<u>List of Targets:</u>

<table>
<tr><th>Lex</th><th>Type</th><th>Regex</th></tr>
<tr><td>athero</td><td>evidence_of_atherosclerosis</td><td>(\batherosclerosis\b|\batheroclerotic\b)</td></tr>
<tr><td>vascular calc</td><td>evidence_of_vascalc</td><td>\svascular\scalc</td></tr>
<tr><td>arteriosclerosis</td><td>evidence_of_atherosclerosis</td><td>(\barteriosclerosis\b|\barterioclerotic\b)</td></tr>
<tr><td>arterial plaque</td><td>evidence_of_artplaque</td><td>\sarterial\splaque</td></tr>
<tr><td>calcified plaque</td><td>evidence_of_calplaque</td><td>\s(\bcalcific\b|\bcalcified\b)\splaque</td></tr>
</table>

<u>Example List of modifiers:</u>

<table>
<tr><th>Direction</th><th>Lex</th><th>Regex</th><th>Type</th></tr>
<tr><td>backward</td><td>Is in the differential</td><td>Is\sin\sthe\sdifferential</td><td>Ambivalent_existence</td></tr>
<tr><td>bidirectional</td><td>Ruled out</td><td>‘ ‘</td><td>Definite_negated_existence</td></tr>
<tr><td>forward</td><td>Is negative</td><td>(is|was) negative</td><td>Definite_negated_existence</td></tr>
</table>

- Parsed the XML format text and marked up the sentences with modifiers and targets. The “.xml” files are parsed to show the presence or absence of the medical condition using the modifiers and targets
- Two new columns were created to show:
    - *Category:* Evidence of athero, calcified plaque, vascular calc, arteriosclerosis
    - *Modifying category:* The modifer target category such as definite_negated_existence etc.,
- When the target words are identified, the modifiers are assigned to specific phrases with the respective tags
- The result obtained is a “.xml” format where the targets and modifiers are appended for each record
- The result consists of all the previous columns that were present in the original dataset: MRN, order_proc_id, description, narrative, category, modifying category
- The final output is retrieved in a CSV file
- The snapshot of the output:



	mrn	order_proc_id	description	narrative	category	modifyingCategory
1	EH2099390	220326324	US CAROTID DOPPLER BILAT - DIAG IMG (CPT=93880)	DATE OF SERVICE: 08.06.2018CAROTID DUPLEX ULTR...	evidence_of_calcplaq	definite_negated_existence
2	EH2126972	229725890	US CAROTID DOPPLER BILAT - DIAG IMG (CPT=93880)	DATE OF SERVICE: 10.04.2018CAROTID DUPLEX ULTR...	evidence_of_calcplaq evidence_of_calcplaq	definite_negated_existence definite_negated_e...
3	GE00037573	211887120	US CAROTID DOPPLER BILAT - DIAG IMG (CPT=93880)	DATE OF SERVICE: 03.29.2018INDICATION: 72 year...	evidence_of_calcplaq	definite_negated_existence
4	GE11178798	234948155	CT ANGIOGRAPHY, CAROTID ARTERIES W AORTIC ARCH...	DATE OF SERVICE: 09.09.2018CTA OF THE NECK, WI...	evidence_of_calcplaq	historical
5	GE11197178	204005432	CT CHEST LD LUNG SCREENING ANNUAL(CPT=G0297)	CT SCREENING FOR LUNG CANCERHISTORY: Personal ...	evidence_of_atherosclerosis	historical
6	GE11255048	237958418	US CAROTID DOPPLER BILAT - DIAG IMG (CPT=93880)	DATE OF SERVICE: 09.26.2018CAROTID DUPLEX ULTR...	evidence_of_calcplaq	definite_negated_existence
7	GE11258906	200954577	XR CHEST PA + LAT CHEST (CPT=71020)	CHEST X-RAY, PA And Lateral Films, 2 Views, 1/...	evidence_of_calcplaq	definite_negated_existence
8	GE11284449	211145374	CT ANGIOGRAPHY, MESENTERIC ARTERIES (CPT=74175)	CT ANGIOGRAPHY, MESENTERIC ARTERIES (CPT=74175...	evidence_of_calcplaq	historical
9	GE11465559	219897041	CT ANGIOGRAPHY, AORTA AND LOWER EXT RUNOFF (SM...	DATE OF SERVICE: 05.24.2018CT ANGIOGRAPHY, AOR...	evidence_of_calcplaq	definite_negated_existence
10	GE11528449	216750392	CT ANGIOGRAPHY, AORTA AND LOWER EXT RUNOFF (SM...	DATE OF SERVICE: 05.09.2018CT ANGIOGRAPHY, AOR...	evidence_of_calcplaq	definite_negated_existence
11	GE11551477	226559231	CT BRAIN OR HEAD (70450)	DATE OF SERVICE: 07.06.2018CT OF THE HEAD, WIT...	evidence_of_calcplaq	historical
12	GE11569703	225915215	US CAROTID DOPPLER BILAT - DIAG IMG (CPT=93880)	DATE OF SERVICE: 06.29.2018CAROTID DUPLEX ULTR...	evidence_of_calcplaq	definite_negated_existence
13	GE11577539	197780552	NM BONE IMAGING SPECT (CPT=78320)	CLINICAL INDICATION:62 years-old Female with ...	evidence_of_calcplaq	probable_negated_existence
14	GE11593191	178579877	CT ANGIOGRAPHY,	CT ANGIOGRAPHY, CAROTIDS+NECK (CPT=70498)	evidence_of_calcplaq	historical

- *Validation:* Given our own input sentences with the medical condition present and absent to check for all the categories

Conclusion

The main motive of developing this NLP solution is to identify the medical condition “atherosclerosis” so that the right ICD- 10 code can be tagged to all the patients that have this condition. By tagging the right ICD-10 code, the right insurance is claimed. This solution will help DMG by not losing money because of misclassification of the diseases.

Future scope: This code can be used for other medical conditions by editing creating the required set of target items and modifiers.