LOINC Document Ontology

Pipeline Documentation

Feb 14, 2023

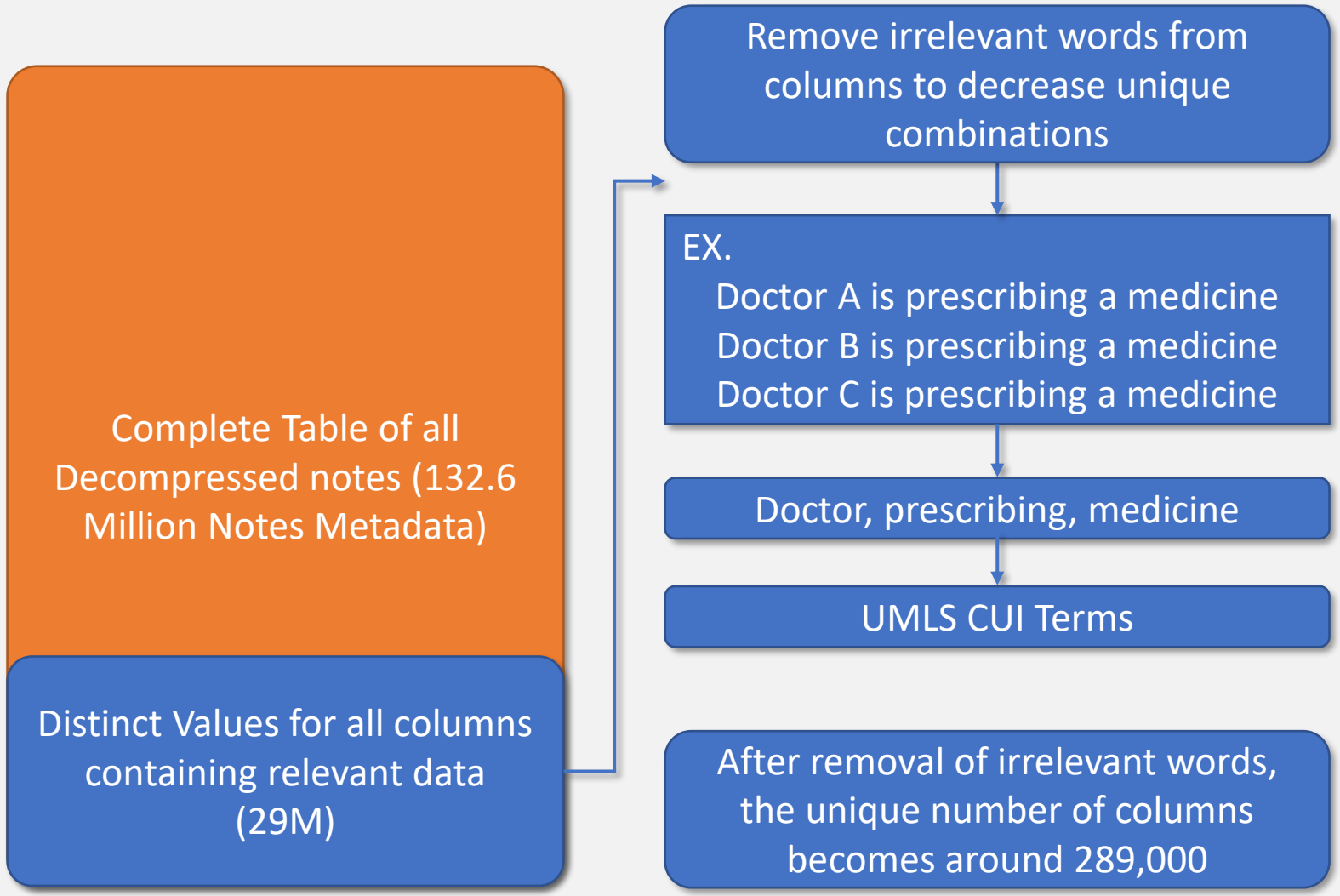
OVERVIEW

- **Metadata Pre-processing**
  - Complexity reduction from Millions to Thousands
- **Bag of Words Implementation**
  - Mapping Bag of words to LOINC Codes
  - Mapping LOINC Codes to Unique Bag of words
  - Mapping LOINC Codes to individual notes
- **Results & Analysis + Discussion**

Num of Parts	Number of Notes	%
0	7,225,678	5%
1	25,015,785	19%
2	41,216,686	31%
3	33,005,516	25%
4	21,267,870	16%
5	4,899,873	4%
Total	132,631,408	

A. Metadata Preprocessing

A.1 Notes Metadata Consolidation



Summary

- Our notes number in around **132.6 Million**.
- We can get distinct value of important columns, which decreases number to around **29M**.
- Next, we can remove irrelevant words (words that don't help us connect note to LOINC Codes).
- Final number of unique notes comes out to be around **289,000 which is iterable in under 2 hours**.

A.2 Data Table Preparation

Original Table

A	B	C	D	E	F
**	**	**	**	**	**
**		**	**	**	**
**	***				**
**	**	**	**	**	**
**	**	**	**		**
**	***	**	**	*	**
**	***	**		**	**
**	***	**		*	**
**	**	**	**	**	**

Intermediary table of Distinct Values

B	D	ID
**	**	1
	**	2
***		3
***	**	4

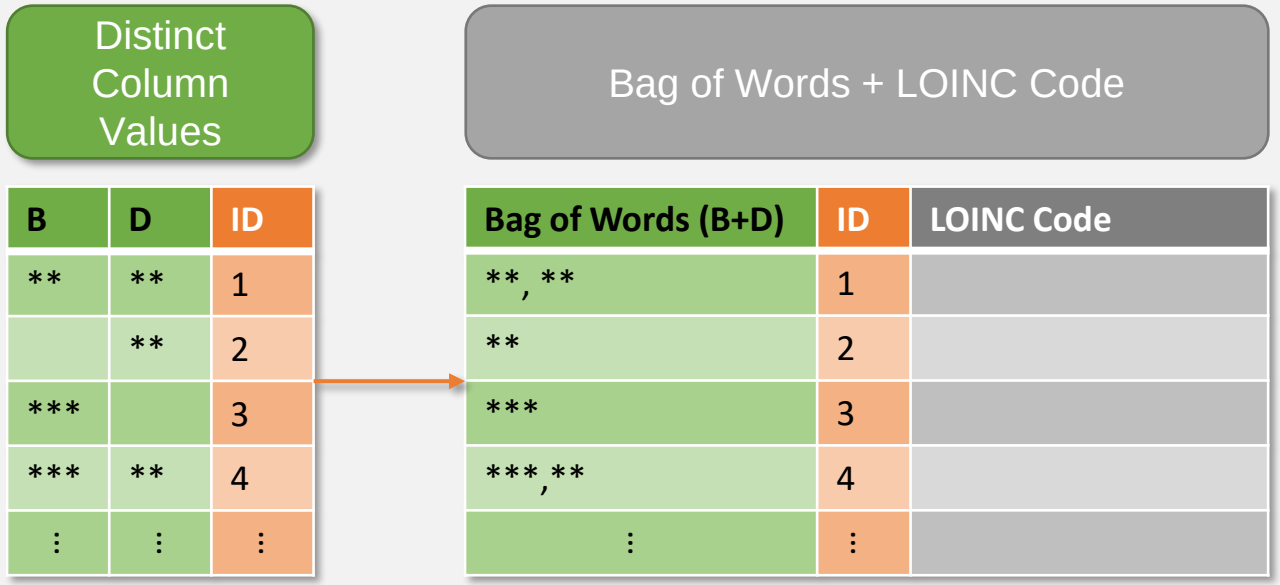
Table with ID Column

A	B	C	D	E	F	ID
**	**	**	**	**	**	1
**		**	**	**	**	2
**	***				**	3
**	**	**	**	**	**	1
**	**	**	**		**	1
**	***	**	**	*	**	4
**	***	**		**	**	3
**	***	**		*	**	3
**	**	**	**	**	**	1

Summary

- The ID field will act as the primary key, allowing for efficient joining of tables with correlated LOINC Codes.
- We will operate on the significantly smaller **Intermediary Table** and join back to the individual notes at the end.

A.3 Bag of Words table with ID



Summary

- Using the **intermediary table** from last step, we will create a new Table.
- The table will have a **Bag of Words** field, storing all the unique bags of words.
- An **ID** field will carry over.
- A new **LOINC Code** field, to be programmatically populated.

B. Bag of Words Implementation

B.1 Vectorized Implementation

{'research': 0, 'exam.hypertension': 0, 'midwife': 0, 'pulmonary': 0, 'rectal': 0, 'hernia': 0, 'student': 0, 'lo
'thromboembolism': 0, 'program': 0, 'exam.feet': 0, 'sleep': 0, 'exam.extremity': 0, 'intestines': 0, 'multi-spec
'intervention': 0, 'microbiology': 0, 'encounters': 0, 'nutrition': 0, 'exam.genitourinary': 0, 'smoking': 0, 'su
'exam.hiv-related': 0, 'exam.sense': 0, 'rehabilitation': 0, 'specialist': 0, 'telephone': 0, 'encounter': 0, 'an
'hepatology': 0, 'exposure': 0, 'neurological': 0, 'breast': 0, 'obstetrics': 0, 'child': 0, 'tumor': 0, 'confere
'comprehensive': 0, 'exam.cushings': 0, 'autopsy': 0, 'labor': 0, 'exam.multiple': 0, 'surgical': 0, 'preoperativ
0, 'equipment': 0, 'reproductive': 0, 'hospital': 1, 'cardiovascular': 0, 'pharmacogenomics': 0, 'cardiopulmonary
'breastfeeding': 0, 'otolaryngology': 0, 'community': 0, 'candidate': 0, 'mellitus': 0, 'term': 0, 'restrictive':
.audio': 0, 'interdisciplinary': 0, 'gulf': 0, 'exam.skin': 0, 'vascular': 0, 'toxicology': 0, 'gaps': 0, 'perina
0, 'pharynx': 0, 'fingers': 0, 'licensed': 0, 'family': 0, 'exam.general': 0, 'colon': 0, 'aerodigestive': 0, 'an
'primary': 0, 'trauma': 0, 'exam.aid': 0, 'gall': 0, 'therapeutic': 0, 'military': 0, 'assessment': 0, 'digestive
'adhesions': 0, 'shift': 0, 'reconciliation': 0, 'exam.chronic': 0, 'registered': 0, 'nerves': 0, 'exam.mouth': 0
.obstructive': 0, 'gastroenterology': 0, 'environmental': 0, 'individual': 0, 'immunology': 0, 'exam.eye': 0, 'sy
'initial': 0, 'gynecology': 0, 'functional': 0, 'counseling': 0, 'dialysis': 0, 'bladder': 0, 'narcolepsy': 0, 'a
'recovery': 0, 'product': 0, 'orthotics': 0, 'pharmacy': 0, 'parathyroid': 0, 'cleft': 0, 'circulatory': 0, 'diso

	LoincNum	PartNumb	PartTypeName		PartName
38	100446-4	LP173051-6	Document.Setting	.	Outpatient
39	100446-4	LP420041-8	Document.SubjectMat...	.	Breastfeeding
40	100447-2	LP173418-7	Document.Kind	.	Note
41	100447-2	LP173213-2	Document.TypeOfService	.	Progress
42	100447-2	LP173051-6	Document.Setting	.	Outpatient
43	100447-2	LP268363-1	Document.SubjectMat...	.	Burn management
44	100448-0	LP173418-7	Document.Kind	.	Note
45	100448-0	LP173213-2	Document.TypeOfService	.	Progress
46	100448-0	LP173051-6	Document.Setting	.	Outpatient
47	100448-0	LP207300-7	Document.SubjectMat...	.	Cardiac surgery
48	100449-8	LP173418-7	Document.Kind	.	Note
49	100449-8	LP173213-2	Document.TypeOfService	.	Progress

```
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0]
```

Summary

- Bag of words are implemented using Vectors
- Allows for compact storage
- Allows for efficient comparison of two vectors (i.e. Bag of Words comparison)
- Vector size is constant and so doesn't depend on word count.

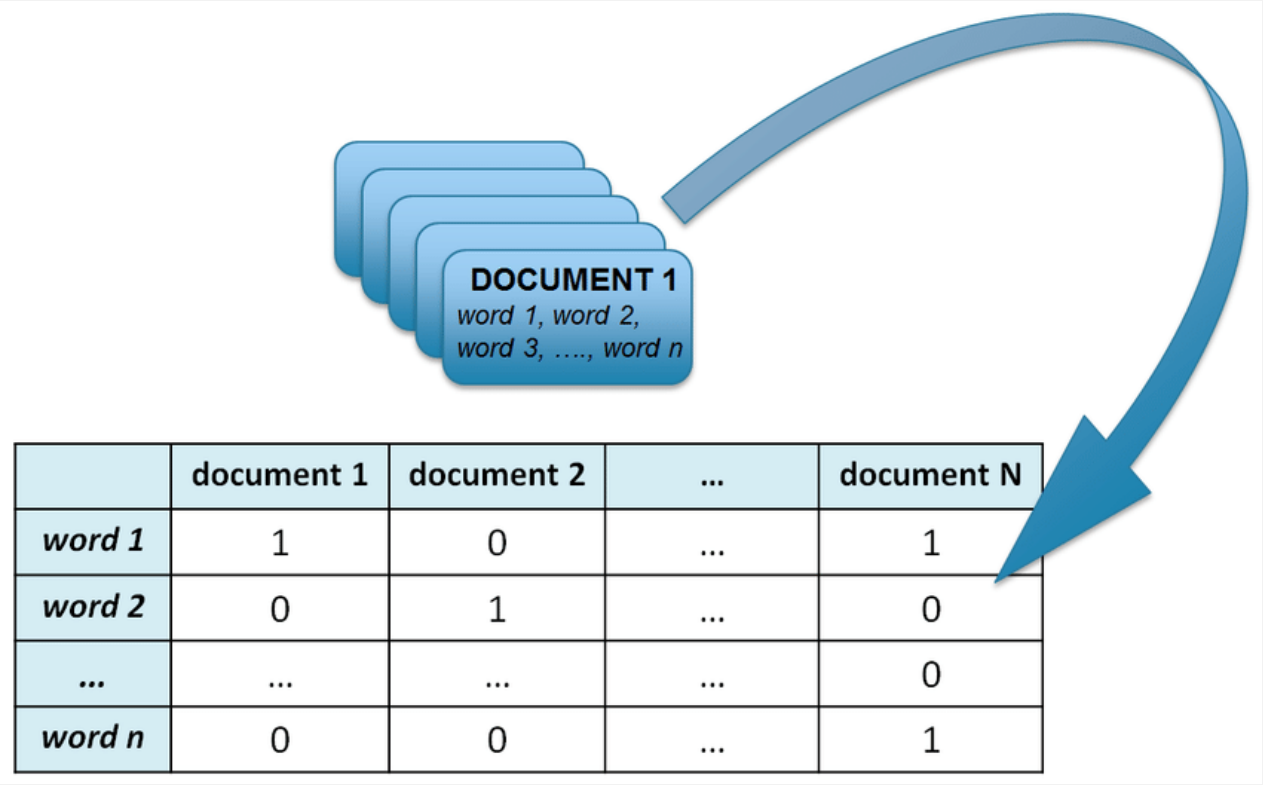
B.2 LOINC Code Word-Base

	LoincNum	PartNumb	PartTypeName		PartName
38	100446-4	LP173051-6	Document.Setting	.	Outpatient
39	100446-4	LP420041-8	Document.SubjectMat...	.	Breastfeeding
40	100447-2	LP173418-7	Document.Kind	.	Note
41	100447-2	LP173213-2	Document.TypeOfService	.	Progress
42	100447-2	LP173051-6	Document.Setting	.	Outpatient
43	100447-2	LP268363-1	Document.SubjectMat...	.	Burn management
44	100448-0	LP173418-7	Document.Kind	.	Note
45	100448-0	LP173213-2	Document.TypeOfService	.	Progress
46	100448-0	LP173051-6	Document.Setting	.	Outpatient
47	100448-0	LP207300-7	Document.SubjectMat...	.	Cardiac surgery
48	100449-8	LP173418-7	Document.Kind	.	Note
49	100449-8	LP173213-2	Document.TypeOfService	.	Progress

Summary

- Vector size is constant and dependent on the LOINC Codes.
- **DocumentOntology.csv** provides a list of LOINC Codes along with **part numbers/names**
- Using the Part names, we can create a Word-base of unique keywords (burn, note, cardiac etc.)
- The keyword count comes to **519**, which is our vector dimension.

B.3 LOINC + Document vectors



	document 1	document 2	…	document N
word 1	1	0	…	1
word 2	0	1	…	0
…	…	…	…	0
word n	0	0	…	1

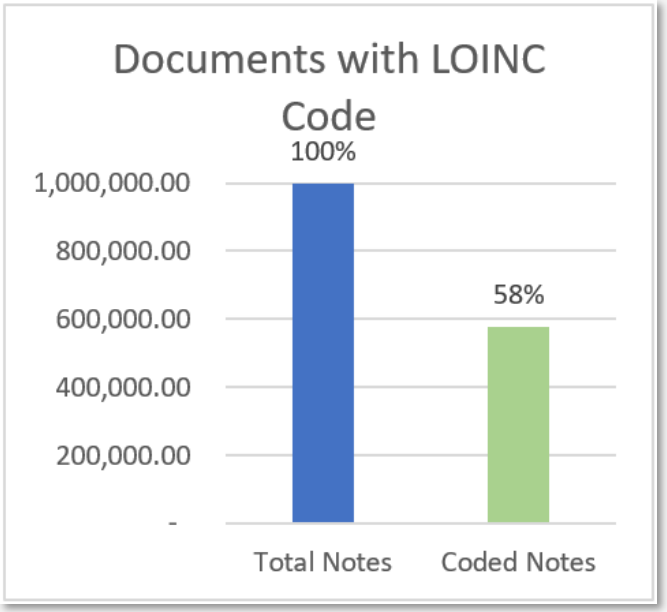
Baj-Rogowska, Anna. (2017). Agile Commerce in the light of Text Mining

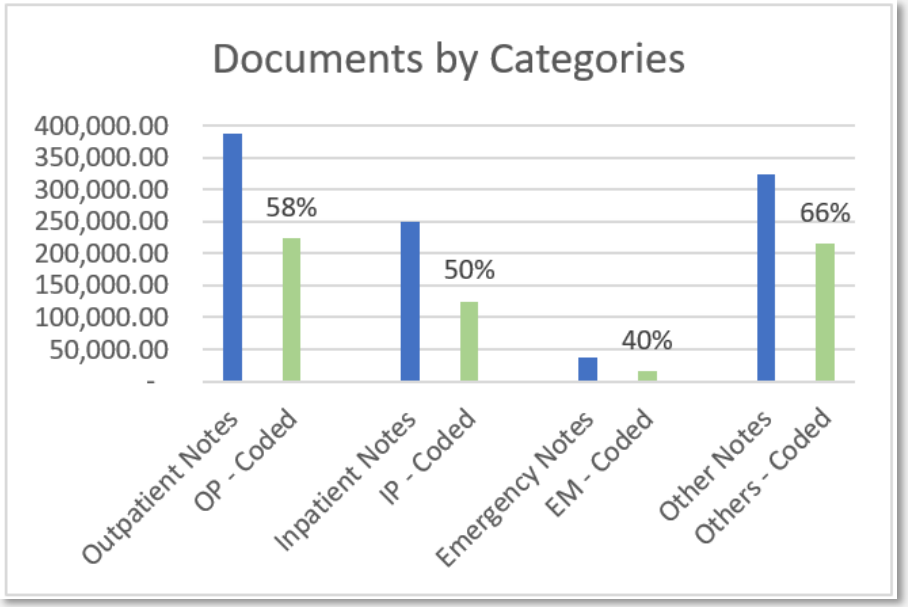
Summary

- Having defined our vector, we can create vectors for LOINC Codes
- For documents, we will only accept words in LOINC Word-base
- Takes under a minute to convert 1M metadata rows into vectors
- Compare documents to LOINC code using vectors and dot-products.
- Roughly 5-20 times faster than string comparison

C. Results & Analysis + Discussion

C.1 Results and Accuracy





Summary

- I tested the pipeline with **1M Documents.**
- Around **58%** were mapped to LOINC Codes
- Broken down into categories, the numbers range from 40-66% coverage.
- For many OP/IP documents, they are only missing **Kind of Document (KOD)** (i.e. Note, Letter, summary etc.)

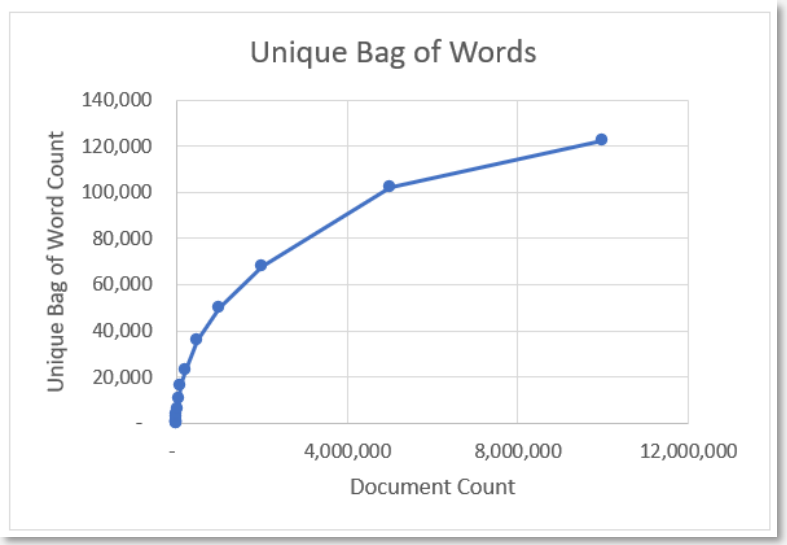
C.2 Values for Categorization

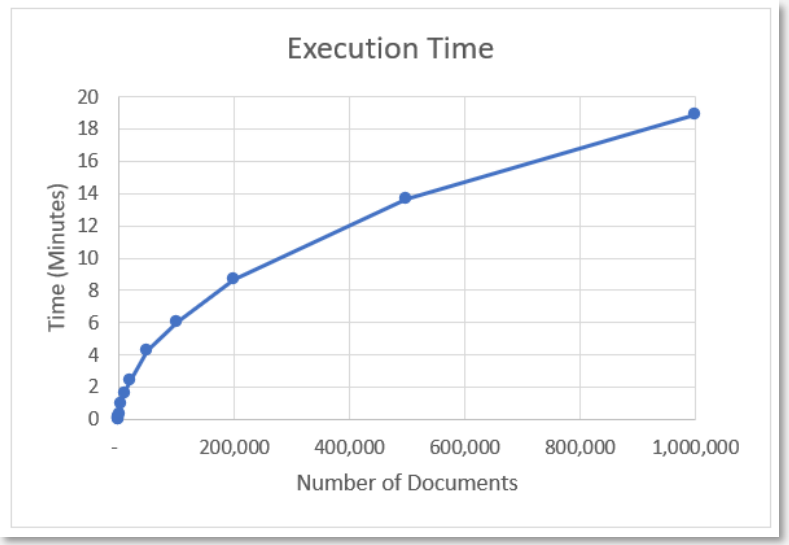
Outpatient Types	Inpatient Types	Others	Emergency
UH OUTPATIENT	UH INPATIENT	Clinic	UH EMERGENCY
WCH PHYSICIAN OP CLINIC	Inpatient	Between Visit	WCH EMERGENCY PATIENT
Outpatient	WCH INPATIENT	Historical	Emergency
MOI OUTPATIENT	MISSOURI PSYCH INPATIENT	UH NO TECHBILL	UH PSYCH ER ASSESSMENT
UH AMBULATORY SURG	WCH NEW BORN INPATIENT	UH TRAUMA	LRA ER
EF OUTPATIENT	MOI INPATIENT	UH OBSERVATION	
WCH OUTPT ANCILLARY	UH INPATIENT NBN	UH DIAGNOSTIC TESTING	
WCH SHORT STAY	WCH REHAB UNIT PATIENT	UH DIAGNOSTIC TEST	
MOI AMBULATORY PROCEDURE PATIENT	zzEF INPATIENT	Virtual Care	
Outpatient in a Bed	zzCR GERIATRIC PSYCH IP	Day Surgery	
EF SHORT STAY	LRA IN	UH THERAPY SERIES	
Non-Admit		Recurring	
MOI SHORT STAY		WCH DIAGNOSTIC TESTING	
MISSOURI PSYCH OUTPATIENT		PreClinic	
UH OUTPT CHRONIC CARE MGMT		Observation	

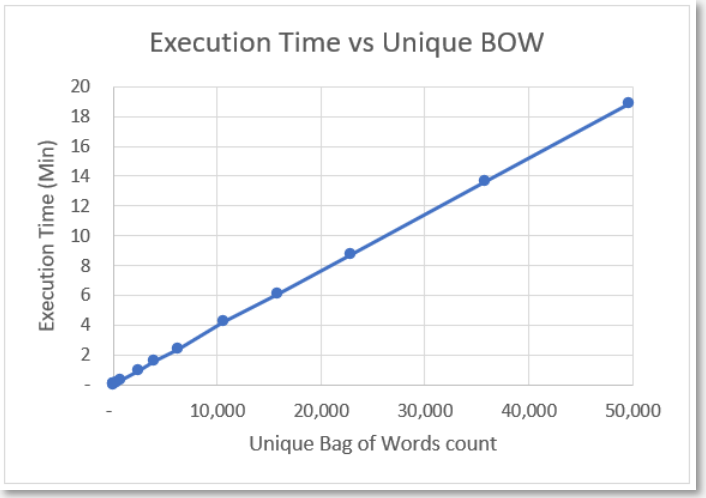
Summary

- Sample list of categories entries taken from unique list in **ENCNTR_TYPE_CD**
- Ranked based on decreasing frequency
- **Others** may include both inpatient and outpatient documents, but we can't tell based on the info we have

C.3 Complexity







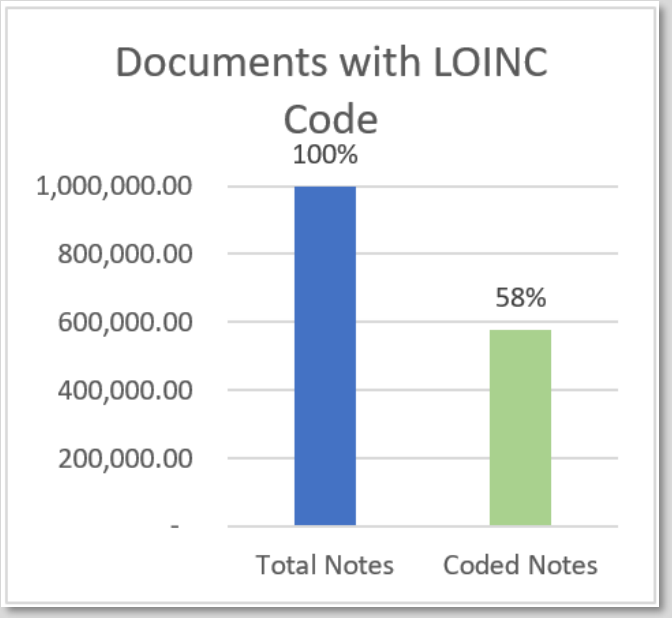
Summary

- The algorithm scales linearly with the number of unique Bag of Words, which is a fraction of the number of documents.
- The graph on the left shows the unique bag of word count as a function of document count.
- The graph on the right shows the program runtime as a function of document count.

C.4 Final Thoughts

Notes

- We might be able to significantly boost the coverage by defaulting to ‘notes’ for documents that don’t have a **KOD** key-word.
- There are LOINC Codes that are ‘Discouraged’ or ‘Deprecated’
  - Very few but identifying and removing them may lower our numbers by a percentage or two



D. END