USING NLP TO INTERPRET MEDICAL RECORDS

Jing Lee
Metis module 5
Natural language processing
July 13, 2022

INTRODUCTION

Problem

- Healthcare records are unstructured data
- Can't do analytics directly on them
- Contains useful information that can be used in diagnosis

Goal

- NLP to extract meaningful context
- Build a recommender system

Impact

- Use a recommender system to look up records of similar patients
- Improve patient care

METHODOLOGY

1. Dataset: Kaggle
   - Medical summaries of 1,740 documents
   - Average document length: 430 words
2. Cleaning data and tokenization: regex, NLTK
3. Vectorizing: TF-IDF
4. Dimensionality reduction: Topic Modeling, NMF
5. Modeling: content-based recommender
   - return documents of similar patients

WHAT IS IN A MEDICAL SUMMARY?

- Patient information
- Complaints
- Condition
- Procedure performed
- Diagnosis
- Medications

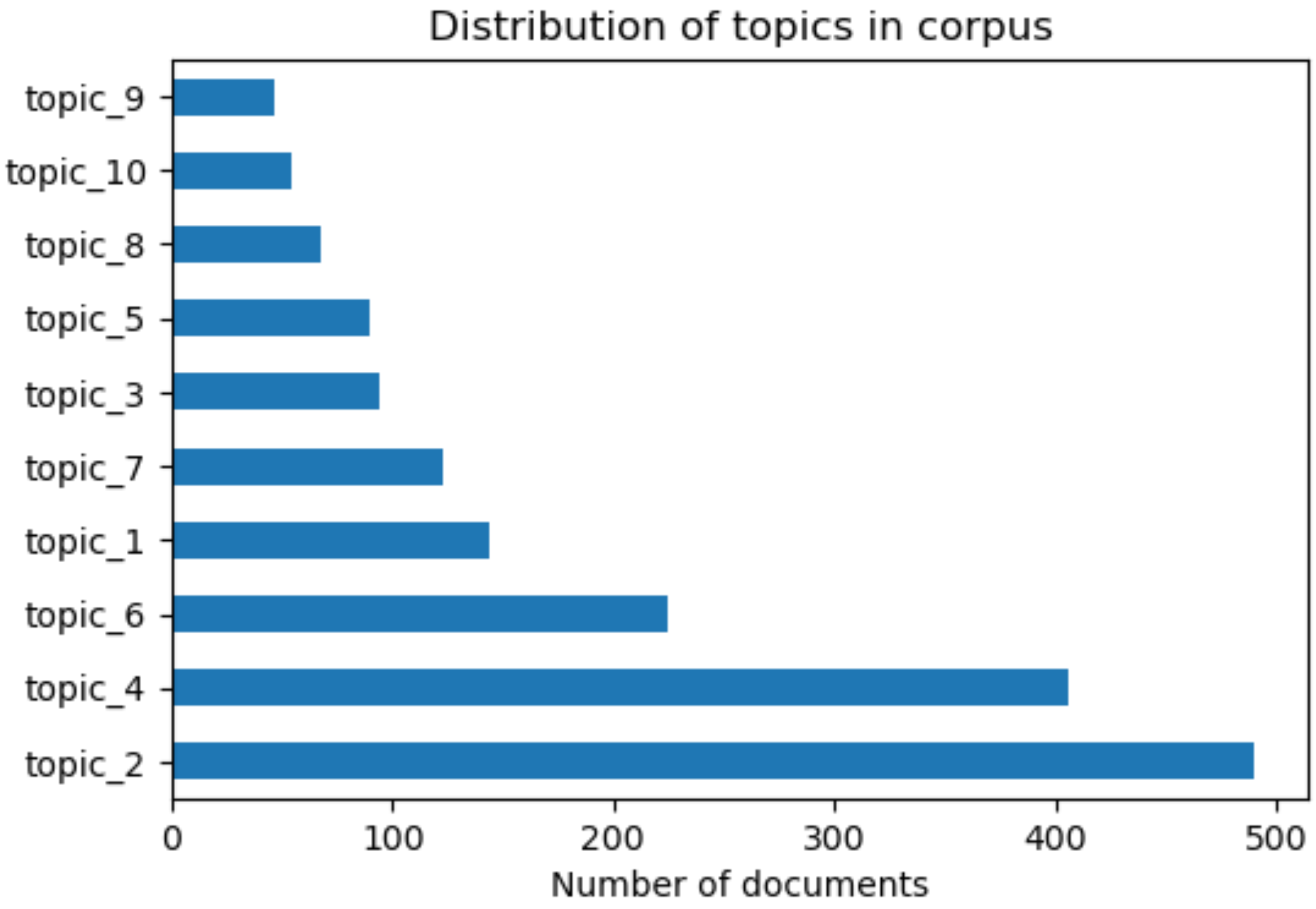
"SUBJECTIVE:, This 23-year-old white female presents with complaint of allergies. She used to have allergies when she lived in Seattle but she thinks they are worse here. In the past, she has tried Claritin, and Zyrtec. Both worked for short time but then seemed to lose effectiveness. She has used Allegra also. She used that last summer and she began using it again two weeks ago. It does not appear to be working very well. She has used over-the-counter sprays but no prescription nasal sprays. She does have asthma but doest not require daily medication for this and does not think it is flaring up."…

VECTORIZER

- Term Frequency Inverse Document Frequency: TF-IDF
- Assign more value to the important terms

TOPIC MODELING: NMF

- Dimensionality reduction
- 1,740 features → 10 topics
- Most common topics: 2, 4, 6
- Least common topics: 8, 9, 10



3 MOST COMMON TOPICS

Topic 2

Pain

Topic 4

Too general

Topic 6

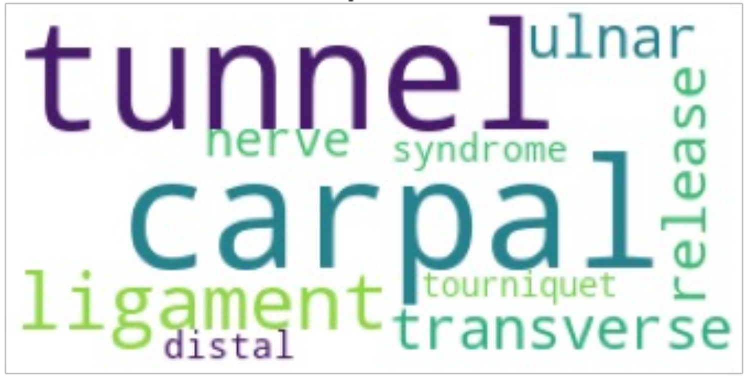
Too general

3 LEAST COMMON TOPICS

Topic 8

Heart

Topic 9



Arm injury

Topic 10

Eye

UNLABELED TOPICS

Topic 1

Topic 4

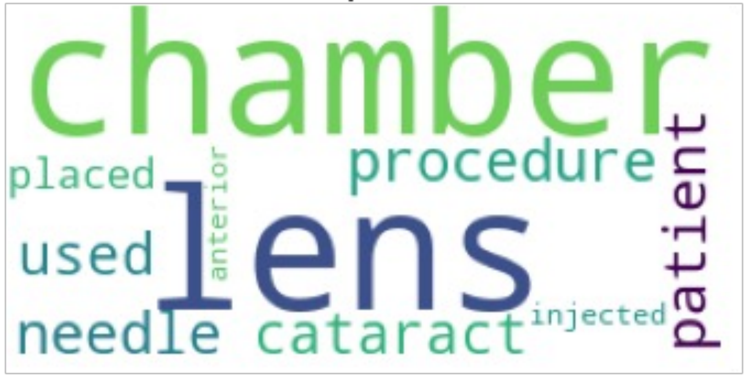
Topic 6

→ **Words are too general**

SIMILAR TOPICS: EYE

Topic 7 and 10 are about the eye

Topic 7



Topic 10

TOPIC MODELING

- Chose 10 topics for best interpretability
- 3 unlabeled topics: too general
- 2 topics about the eye
- Topics contain medical jargon

RECOMMENDER ENGINE

- Input: a document
- Recommender: 3 closest documents to view

```
Current doc: 88
The 3 recommended documents are:
doc 92: distance 0.004
doc 67: distance 0.009
doc 65: distance 0.009

Summary statistics of the distance:
Maximum distance: 0.481
Average distance is: 0.166
```

RECOMMENDER EXAMPLE

Input: doc 88

"PREOPERATIVE DIAGNOSIS:, Right inguinal hernia.,POSTOPERATIVE DIAGNOSIS:, Right inguinal hernia.,PROCEDURE:, Right inguinal hernia repair.,INDICATIONS FOR PROCEDURE: , This patient is a 9-year-old boy with a history of intermittent swelling of the right inguinal area consist" …

Recommend: doc 92

"PREOPERATIVE DIAGNOSIS:, Left inguinal hernia.,POSTOPERATIVE DIAGNOSIS:, Left direct and indirect inguinal hernia.,PROCEDURE PERFORMED:, Repair of left inguinal hernia with Prolene mesh.,ANESTHESIA: , IV sedation with local.,COMPLICATIONS:," …

CONCLUSION

- Built a baseline model for a recommender system
- It is possible to find similar patients from medical summaries
- Some topics remained unlabeled
- Require feedback from users for model performance

FUTURE

- Spacy for tokenization
  - scispaCy: package for biomedical terms
- Further tune topic modeling to achieve meaningful topics
- Have a medical professional to determine the topics
- Use feedback from users to further improve recommendations

THANK YOU!

Questions?

APPENDIX

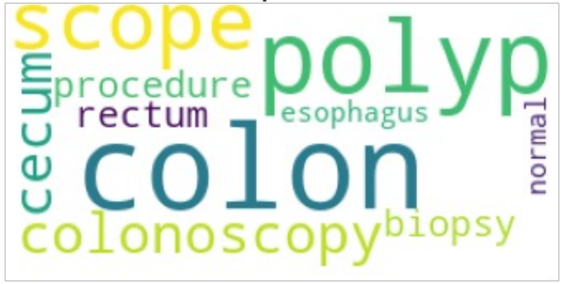
Topic 1 General

Topic 2 Pain

Topic 3 Arteries

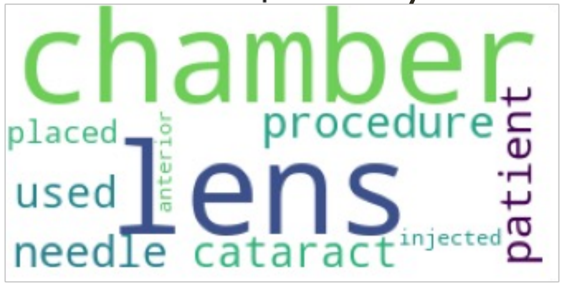
Topic 4 General

Topic 5 Colon



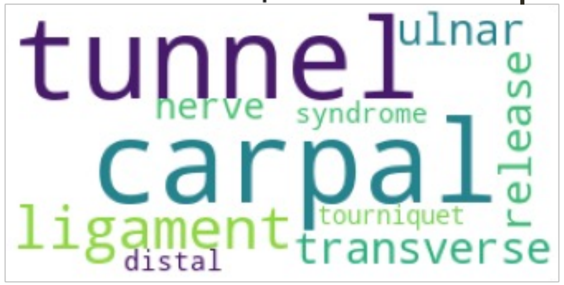
Topic 6 General

Topic 7 Eye

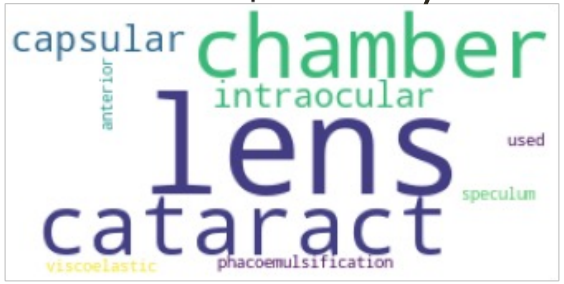


Topic 8 Heart

Topic 9 Arm injury



Topic 10 Eye



- 10 words per topic
- 3 unlabeled
- 7 labeled
- 2 similar topics (eye)