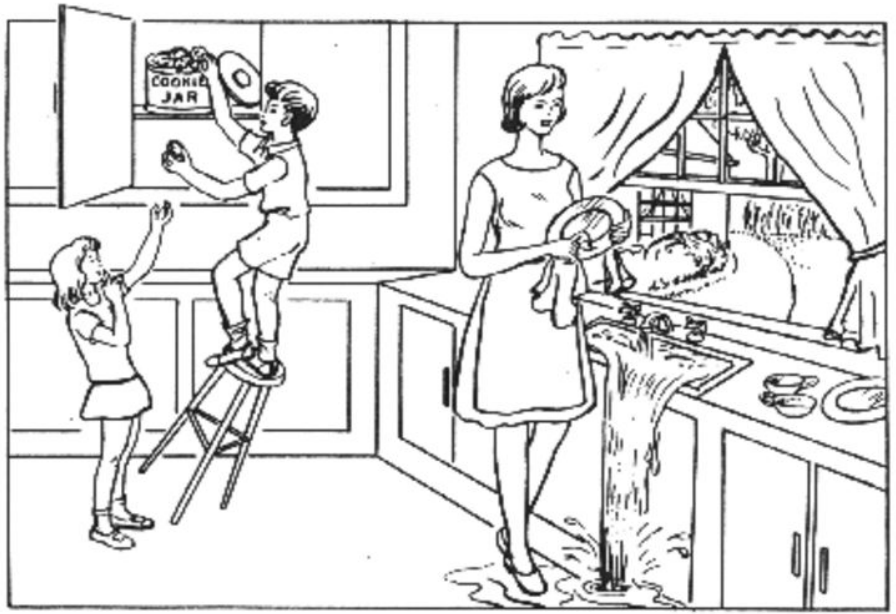
Applied NLP in healthcare

STA2453 - February 23, 2021

About Me

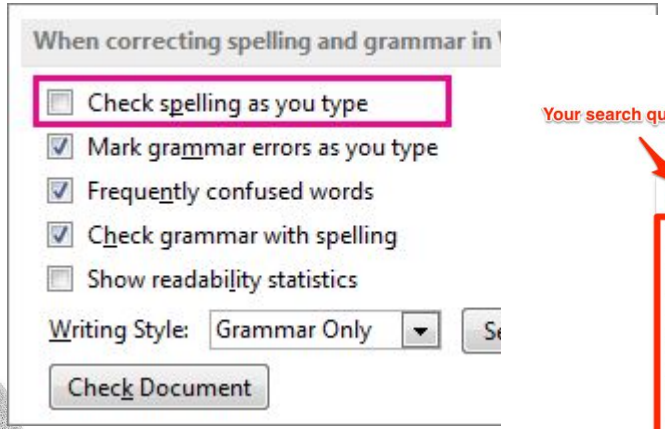
- 2016 - 2018: M.Sc. Computer Science @ University of Toronto
- 2018 - now: Data Scientist @ Unity Health Toronto (St. Michael's Hospital)
  - **D**ata **S**cience & **A**dvanced **A**nalytics
  - Improve patient outcomes
  - Improve hospital efficiency



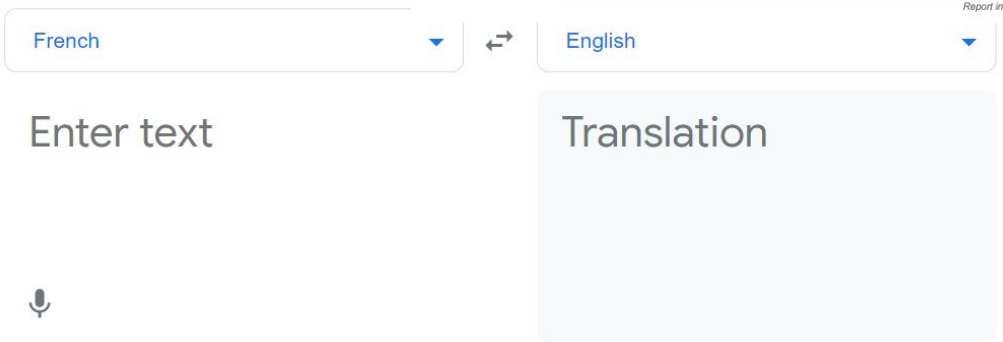
Some of the things I work on...

- Building and evaluating machine learning models with healthcare data
  - Predicting patient deterioration in General Internal Medicine
  - Extracting information from Multiple Sclerosis consult notes
  - Predicting admission to the hospital from past hemodialysis session data
- Deploying and monitoring machine learning models in the hospital
- Developing dashboards
- Lots and lots and lots of data wrangling

Why NLP?


When correcting spelling and grammar in
Check spelling as you type
Mark grammar errors as you type
Frequently confused words
Check grammar with spelling
Show readability statistics
Writing Style:
Grammar Only
Check Document


Your search query
Google
Options to "Remove" searches based on your browser history
skateboards for
skateboards for sale
skateboards for cheap
skateboards for kids
skateboards for beginners
skateboards for girls
skateboards for toddlers
skateboards for sale near me
skateboards for adults
skateboards for 8 year olds
skateboards for dogs
Remove
Remove
Autocomplete suggestions
Google Search
I'm Feeling Lucky
Report inappropriate predictions


French
English
Enter text
Translation

Examples of NLP in healthcare

https://clinical-nlp.github.io/2020/program.html

- Various Approaches for **Predicting Stroke Prognosis** using Magnetic Resonance Imaging Text Records
- **Multiple Sclerosis Severity Classification** From Clinical Text
- BERT-XML: Large Scale Automated **ICD Coding** Using BERT Pretraining
- Incorporating Risk Factor Embeddings in Pre-trained Transformers Improves **Sentiment Prediction** in Psychiatric Discharge Summaries
- **Information Extraction** from Swedish Medical Prescriptions with Sig-Transformer Encoder
- Evaluation of Transfer Learning for **Adverse Drug Event** (ADE) and Medication Entity Extraction
- A Portuguese Neural Language Model for **Clinical Named Entity Recognition**

NLP moves fast...

Agenda

- Pre-process! Preprocess! Pre process!

- **Use case 1:** Extracting information from Tuberculosis Clinic consult notes through rule-based NLP

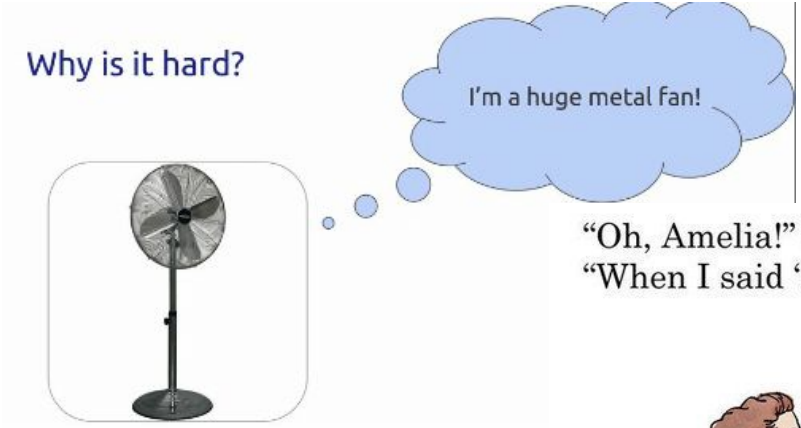
- **Use case 2:** Predicting patient outcomes from nurse notes through topic modeling

- **Use case 3:** Extracting information from Multiple Sclerosis consult notes through rule-based NLP and word embeddings

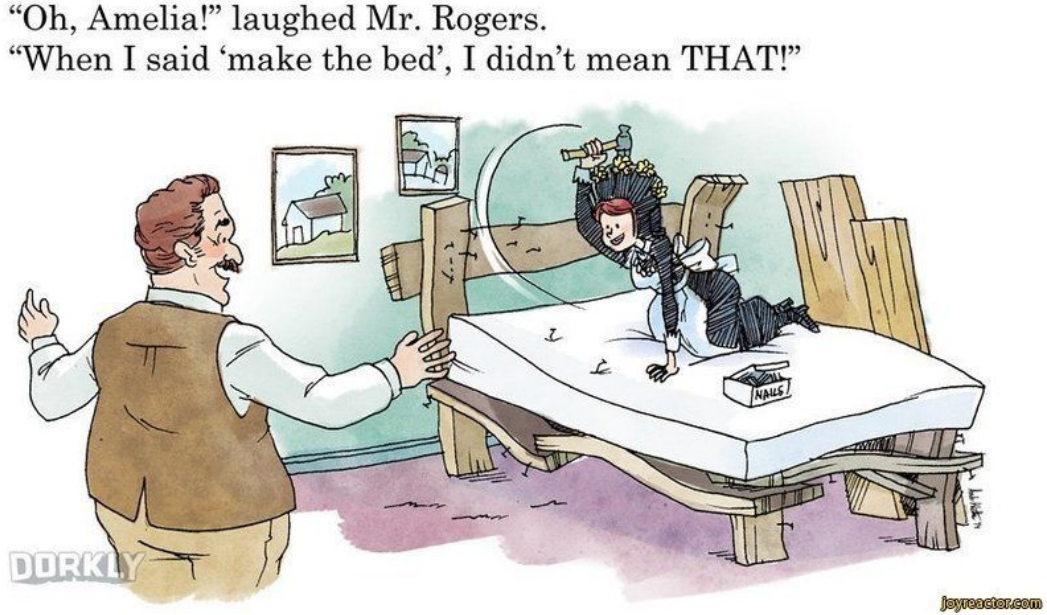
- Clinical dictation
- Diagnosis coding
- Predicting LOS from admission notes
- Radiology notes
- Chatbots
- Audio data???
- Handwritten notes!?!?!

Preprocessing

Language is... ambiguous

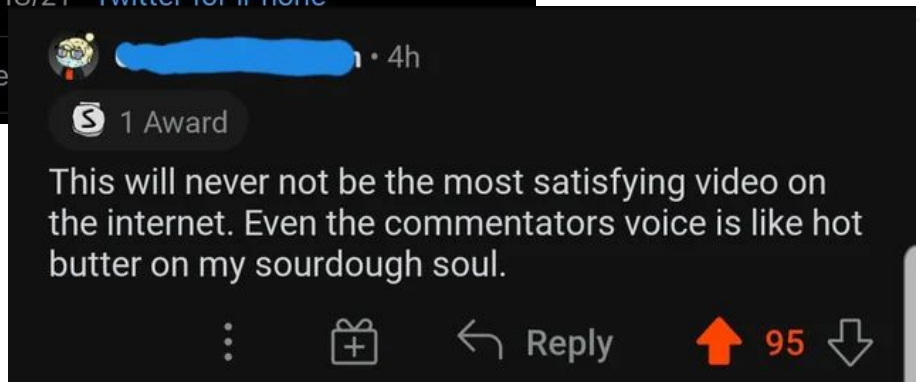


NLP is hard.



Language is… constantly changing

r/BrandNewSentence



Home / News & Opinion

Dogs Are Teaching Machines to Sniff Out Cancer

Hard to work with healthcare text data

- Flexible formatting:

  "Height: (*in*) **75** Weight (*lb*): **245** BSA (*m2*): **2.39** *m2* BP (*mm Hg*): **92/52** HR (*bpm*): **120**"

- Atypical grammar: "Ultrasound showed no evidence of [a] pseudoaneurysm"
- Language specific to medical domain: "recent tension **PTX** at **OSH**"
  - Domain-specific acronyms: "T2"
- Misspellings: "ventilator dependent **respiratoy** failure"

* Examples from Leaman et al., "Challenges in clinical natural language processing for automated disorder normalization", 2015.

Common Preprocessing Steps

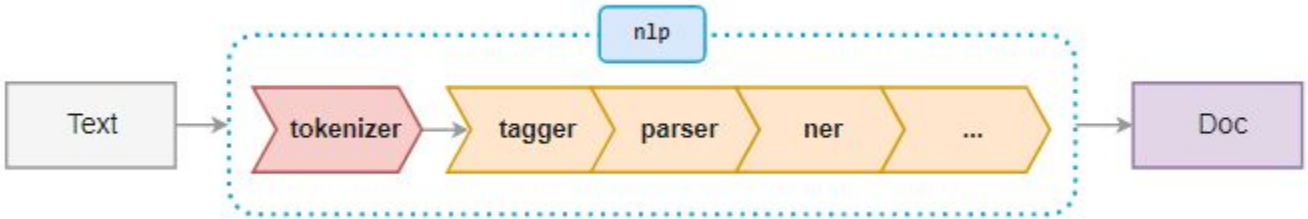


Image Source: https://medium.com/analytics-vidhya/nlp-preprocessing-pipeline-what-when-why-2fc808899d1f

Common Preprocessing Steps

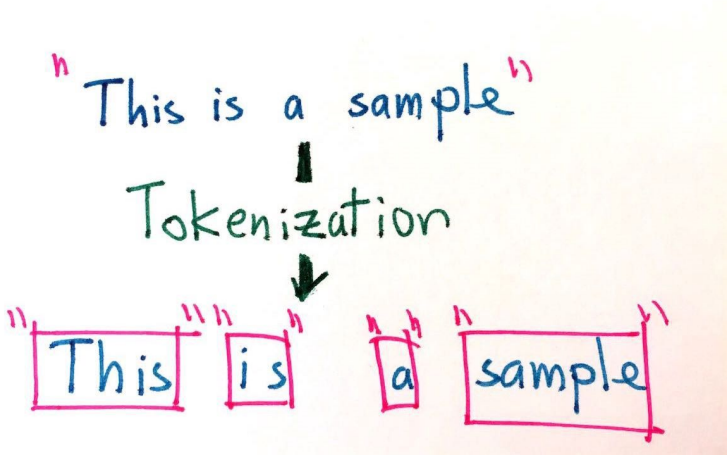
- Stop word removal
  - Common words in the English language (e.g., “a”, “the”)
- Removing extraneous characters
  - White space
  - Punctuation?
  - HTML?
  - Numbers?
- Lemmatization

	original_word	lemmatized_word
0	trouble	trouble
1	troubling	trouble
2	troubled	trouble
3	troubles	trouble

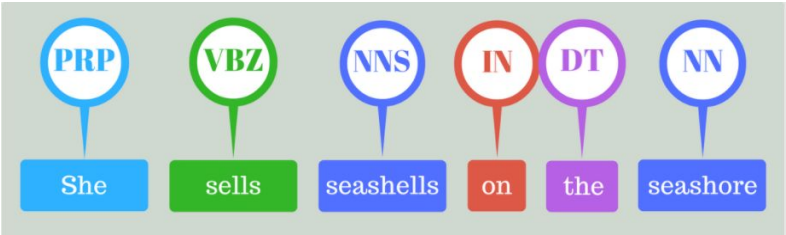
Source for image: https://www.topbots.com/text-preprocessing-for-machine-learning-nlp/

Common Preprocessing Steps

- Tokenization
- Remove frequent/infrequent words
- Tagging
  - Parts-of-speech: a category of words that have similar grammatical properties (e.g., nouns, adjectives, verbs, etc.)

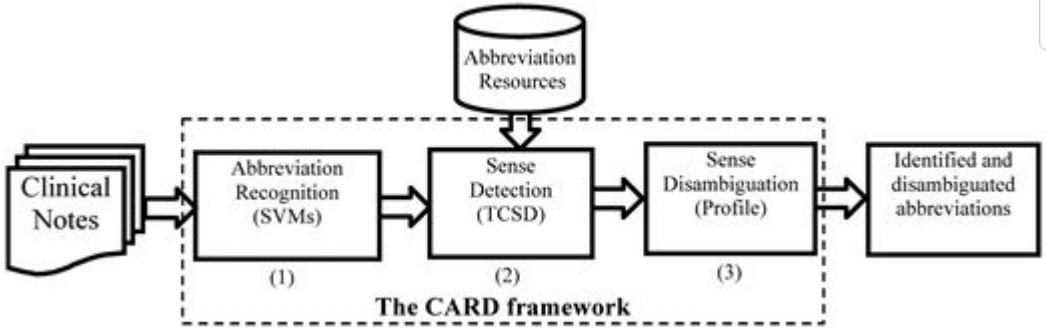


Sources for images:
https://medium.com/data-science-in-your-pocket/tokenization-algorithms-in-natural-language-processing-nlp-1fceab8454af



Common Preprocessing Steps

- Correct spelling mistakes.
  - Hunspell R package: https://cran.r-project.org/web/packages/hunspell/vignettes/intro.html
- Abbreviation expansion
  - e.g., “pt” → patient
  - Clinical Abbreviation Recognition and Disambiguation (Wu et al., 2017)




```r
library(hunspell)

# Check individual words
words <- c("beer", "wiskey", "wine")
correct <- hunspell_check(words)
print(correct)
```

```
[1]  TRUE FALSE  TRUE
```

```r
# Find suggestions for incorrect words
hunspell_suggest(words[!correct])
```

```
[[1]]
[1] "whiskey"  "whiskery"
```

Common Preprocessing Steps

- Normalize names of drugs → RxNorm
  - Drugs are assigned a RxCUI → Concept Unique Identifier
  - API

getAllProperties	/rxcui/{*rxcui*}/allProperties	Return all properties for a concept	Active
getAllRelatedInfo	/rxcui/{*rxcui*}/allrelated	Return all related concept information	Active
getApproximateMatch	/approximateTerm	Approximate match search to find closest strings	Active/Current
getDisplayTerms	/displaynames	Return the auto suggestion names	Active
getDrugs	/drugs	Return the related drugs	Active
getIdTypes	/idtypes	Return the available identifier types	Active

https://rxnav.nlm.nih.gov/REST/approximateTerm?term=zocor%2010%20mg&maxEntries=4

(returns)

```
<rxnormdata>
  <approximateGroup>
    <inputTerm>zocor 10 mg</inputTerm>
    <maxEntries>4</maxEntries>
    <comment>Trying zocor as drug; </comment>
    <candidate>
      <rxcui>563653</rxcui>
      <rxaui>2278986</rxaui>
      <score>75</score>
      <rank>1</rank>
    </candidate>
    <candidate>
      <rxcui>104490</rxcui>
      <rxaui>6362167</rxaui>
      <score>60</score>
      <rank>2</rank>
    </candidate>
    <candidate>
      <rxcui>104490</rxcui>
      <rxaui>786234</rxaui>
      <score>60</score>
      <rank>2</rank>
    </candidate>
  </approximateGroup>
</rxnormdata>
```

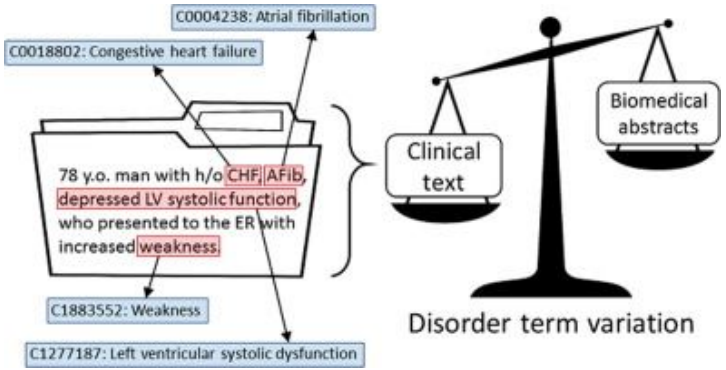
https://rxnav.nlm.nih.gov/REST/rxclass/class/byRxcui.json?rxcui=7052&relaSource=MEDRT&relas=may_treat

(returns)


```
{
    "userInput": {
        "relaSource": "MEDRT",
        "relas": "may_treat",
        "rxcui": "7052"
    },
    "rxclassDrugInfoList": {
        "rxclassDrugInfo": [
            {
                "minConcept": {
                    "rxcui": "30236",
                    "name": "Morphine Sulfate",
                    "tty": "PIN"
                },
                "rxclassMinConceptItem": {
                    "classId": "D004417",
                    "className": "Dyspnea",
                    "classType": "DISEASE"
                },
                "rela": "may_treat",
                "relaSource": "MEDRT"
            },
            {
                "minConcept": {
                    "rxcui": "30236",
                    "name": "Morphine Sulfate",
                    "tty": "PIN"
                },
                "rxclassMinConceptItem": {
                    "classId": "D010148",
                    "className": "Pain, Intractable",
```

Common Preprocessing Steps

- Normalize clinical concepts
  - UMLS (Unified Medical Language System) → set of files and software that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems.
  - SNOMED CT → systematically organized computer processable collection of medical terms providing codes, terms, synonyms and definitions used in clinical documentation and reporting



Source for image: Leaman et al., 2015

Models take as input numbers... How can we represent text as numbers?

- Count the number of text matches (Use case #1)
- Look at the topics in each text (Use case #2)
- Map each word to a vector representation (Use case #3)
- Bag-of-words (BoW)
- Term frequency-inverse document frequency (TF-IDF)
- Linguistic features

Use Case 1: NLP + TB

Tuberculosis Clinic

- The Tuberculosis (TB) Program at the hospital provides specialized and comprehensive care for patients with suspected or diagnosed tuberculosis and is one of the largest TB clinics in Canada.
- Many key performance metrics for the TB clinic are only available in unstructured text notes. Previously, the only way to get at this data was through manual chart abstraction, which is a time and labour intensive process.
    - E.g., What is the patient's diagnosis? Active TB? Latent TB infection? No diagnosis?
    - E.g., What was the duration of a previous treatment? 4 months? 12 months?

Tuberculosis Clinic

- Primary goal: Develop an inventory of NLP rulesets that can be used to rapidly extract data variables relevant to the TB clinic from digital text notes.

- Secondary goals:
  - Develop a research database of TB clinical variables.
  - Develop a visualization tool for the extracted data elements.

Tuberculosis Clinic

- In order to improve hospital operations, patient attributes and risk factors need to be extracted from free-form text.

- Text data: text documents from ~200 patients within the TB clinic
  - Transcriptions of dictated consult notes of a tuberculosis clinic.
  - Radiology reports.

Variables of interest

- Medications → Did the patient ever start medication of interest?
- Drug sensitivity → Was the tuberculosis resistant to antibiotic?
- Treatment duration
- Sites of active tuberculosis
- Adverse drug reactions
- Was the patient previously treated for tuberculosis?
- What was the chest x-ray result?
- Smoking status
- Diagnosis → Active TB or latent TB infection?

Rule-based NLP

- Approach: **weighted keyword search**.
- Use **regular expressions** to match patterns of text.
- Regular expression rules assign a score to each sentence.
- The cumulative score determines the final classification.
- Intuition: There are patterns/keywords/phrases in sentences, and some are more important than others.
    - “…. was diagnosed with…”
    - “… has a clear diagnosis of…”
    - “… probable diagnosis of…”

Rule-based NLP

Leverage existing patterns in the text

Current smoker:

- … has been smoking 6 cigarettes daily for the past 3 years. No history of smoking any other drug substances …
- … smokes tobacco from a pipe twice daily, and has done so for the past 30 years…

Non-smoker

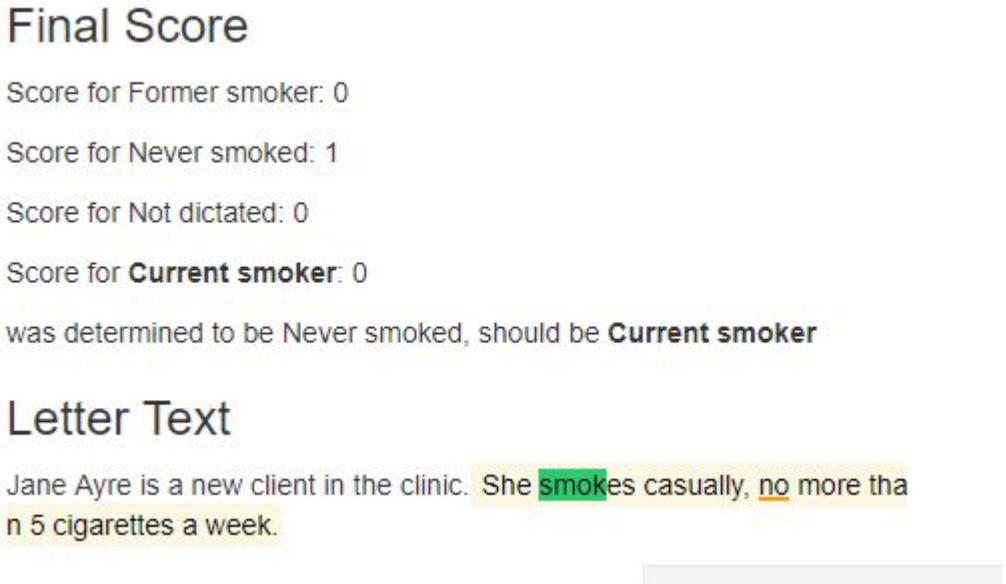
- … smokes crack cocaine on occasion, but does not smoke tobacco…
- … he does not smoke tobacco and loves dogs…
- … she has no history of smoking…

UI

- Developed interface for developing and refining rules.

UI

- Interface highlights instances where rules fail → end-user can then refine the rule accordingly.



Rule-based NLP

✔️ Easy to explain and maintain

✔️ Easy to incorporate domain knowledge

✔️ Easy to debug

✔️ Doesn’t need a lot of data

❌Manual effort

❌ What if you don’t have access to an expert to build the rules?

❓ Generalizes well?

Tools used in this project

- Python → regular expressions
- Javascript, CSS, Python → develop interface that allowed to interactively build rules with experts

Use Case 2: NLP + GIM

Predicting risk deterioration in GIM patients

- 7% of General Internal Medicine (GIM) patients die or are transferred to an Intensive Care Unit (ICU)
- Most ICU transfers are unexpected; 4/5 ICU transfers occur with less than 3 hours warning
- #1 root cause of unplanned ICU transfer is failure of monitoring (46%)
- Goal: Develop an **early warning system**.
    - Reduced mortality
    - Less ICU use and shorter LOS
    - Better communication with patients and families

Predicting risk deterioration in GIM patients

- Outcomes:
  - Death in GIM
  - Transfer from GIM to ICU
  - Transfer from GIM to palliative care
- Data:
  - Routinely-measured labs and vitals
  - Clinical orders (e.g., “NPO order”)
  - Demographics
  - Medication orders
  - **Nurse notes**

Predicting risk deterioration in GIM patients

- Processing:
  - Create patient timeseries of 6-hour intervals
  - Multiple measures in same interval → average
  - Compute indicator variables (1 if measured, 0 if not measured)
  - Last observation carried forward

- Output:
  - Categorize patients as High risk vs Medium risk vs Low risk
  - Alerts sent to pager
  - Emails sent to teams

Nurse notes

- ~800,000 notes
- Information on patient status/disposition, changes in monitoring, changes in orders, etc.
    - Some of this information will also be documented in a more structured way in clinical orders, vitals, etc.
- Note length: ~ 60 words/note
- Preprocessing:
    - Lowercase
    - Remove punctuation, numbers, stop words, most frequent words
    - **Abbreviation expansion**

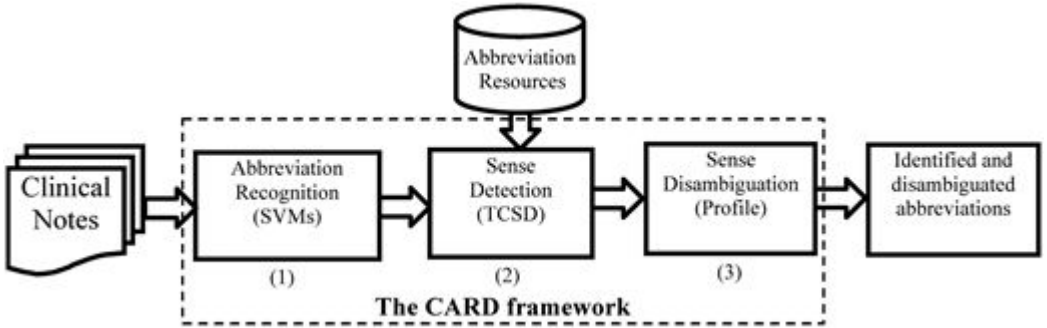
Abbreviation Expansion

Example on note from MIMIC-III dataset

please eval for ptx, chf, effusions, infiltrates final report indications : hypertension, v-fib rest, s/p right subclavian line placement. portable chest : comparison is made to previous films from four hours prior. findings : there has been placement of a right sided subclavian catheter, with the tip in the proximal svc..

please evaluation for pneumothorax , congestive heart failure , effusions , infiltrates final report indications : hypertension , 5 - fib rest , soft / posterior right subclavian line placement .portable chest : comparison is made to previous films from four hours prior .findings : there has been placement of a right sided subclavian catheter , with the tip in the proximal service

Source: Brenna Li, Jienan Yao, Stephen Gou, Yuyang Liu.



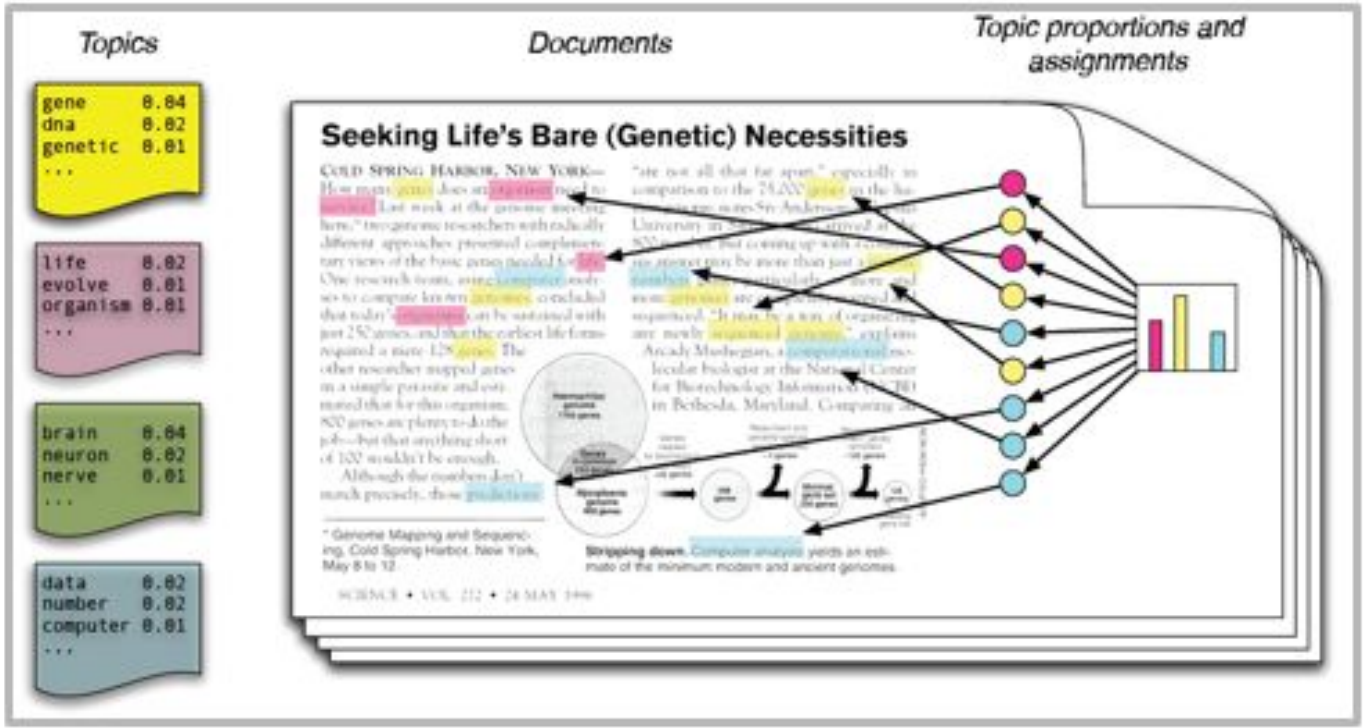
Topic Modeling

We can represent each text note as topic probabilities.

note_id	*topic_1*	*topic_2*	*topic_3*	*topic_4*	*topic_5*
1	0.1	0.1	0.1	0.1	**0.6**
2	0.05	0.05	**0.5**	0.25	0.25

How do we get these topics?

Topic Modeling: LDA



Topic Modeling

- Extract topics from each note using Latent Dirichlet Allocation (LDA).

- LDA automatically identifies topics present in a text and derives hidden patterns.

- Topics are a repeating pattern of co-occurring terms in text.

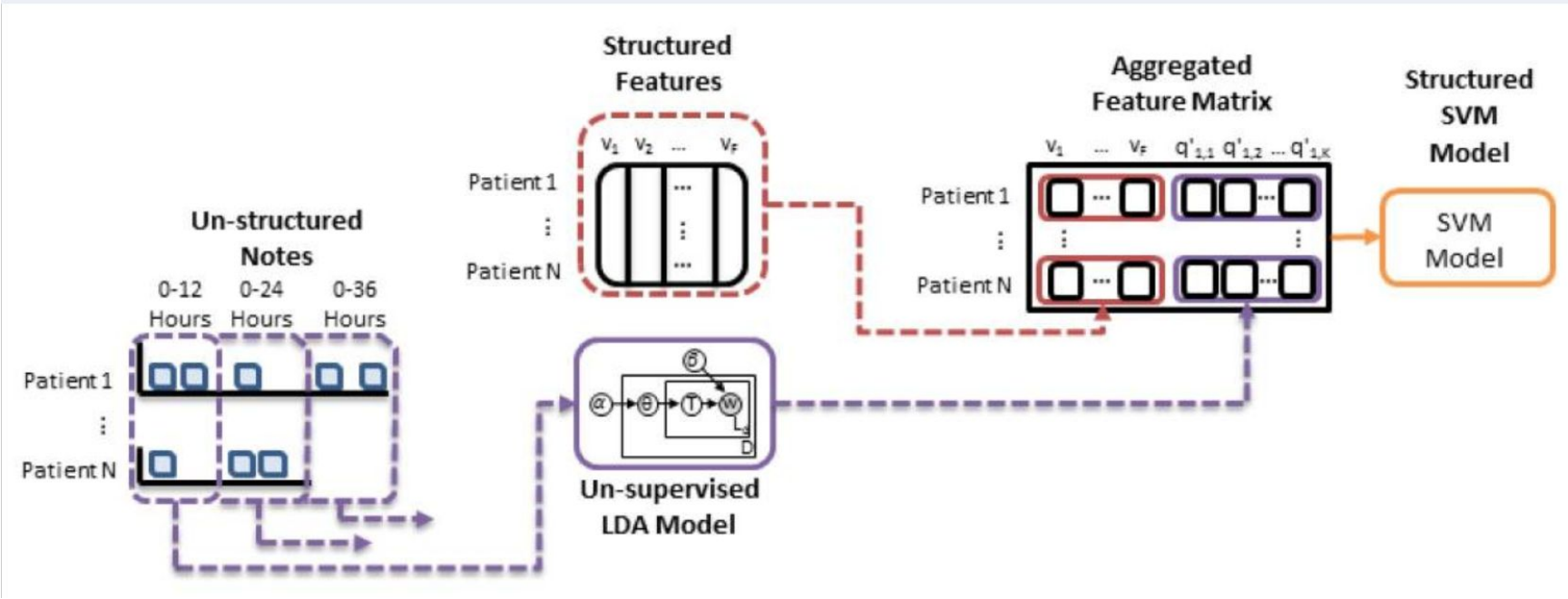
What does it look like?

Table 1. Example of Enriched Topics

	Topic	Top Ten Words	Possible Topic
In-Hospital Mortalities	41	family pt neuro name status eyes care remains movement pupils	End-of-life care requested
	28	vent pt abg intubated peep remains secretions resp ac sedated	Respiratory failure with intubated patient
	44	gtt insulin map remains line heparin levo ptt fluid levophed	Severe hypotension by sepsis-induced systemic vasodilation
	37	pt noted cc hr bp cchr am micu urine bs	Many physiological parameters noted, corresponding to critical state

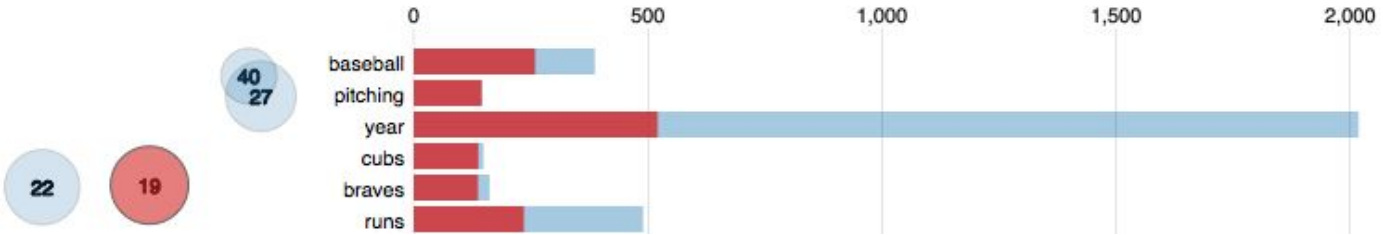
Source: Ghassemi et al., “Unfolding Physiological State: Mortality Modelling in Intensive Care Units”, 2014.

Combining LDA features with tabular data



Tools used in this project

- Gensim: https://radimrehurek.com/gensim/ → build LDA topic model in Python
- Tidytext: https://www.tidytextmining.com/topicmodeling.html → build LDA topic model with R
- Pyldavis: https://github.com/bmabey/pyLDAvis → visualization of topic models
- CARD: https://sbmi.uth.edu/ccb/resources/abbreviation.htm → abbreviation expansion



Tools used in this project

- tidyr, dplyr, lubridate → R data processing
- pandas → Python data processing
- pytorch → deep learning models with Python
- earth → MARS models with R

Use Case 3: NLP + MS

Multiple Sclerosis Clinic

- Multiple Sclerosis (MS) is a chronic neurological disease affecting the central nervous system (CNS) and is the leading cause of neurological disability in young adults.
- Canada has one of the highest rates of multiple sclerosis (MS) in the world.
- The MS clinic at St. Michael's Hospital is among the largest in the world.
    - For each patient visit, a consult note is dictated by the treating physician → details on patient history and the neurological examination.
- The MS clinic has built its own database from the consult notes.
    - Time-intensive, labor-intensive

Multiple Sclerosis Clinic

- Goal 1: Automate/help the current data extraction process.
- Goal 2: Build risk prediction models based on the extracted variables.

- Text data: neurology consult notes
  - patient's entire health history (e.g., demographics, family history, relapse history, etc.)
  - results from MRI test and neurology exam (e.g., tests about coordination, reflexes, etc.)
  - current disease modifying treatments and suggested treatments

Expanded Disability Scale Status

- Score ranging from 0 to 10 (high scores represent greater level of disability)
- Typically found in each dictated consult note → track disability progression over time
- Will sometimes be stated in the note… but not always...
  - E.g., “patient has an EDSS score of 4.” “their EDSS is 2.”
- Chart abstractors assign a score for every visit.

Neurology consult note

- 1 note/visit → ~11,000 notes
- Preprocessing:
  - Remove header/footer → patient names, doctor names, addresses, etc.
  - Remove stop words, most frequent words
  - Set length of each document to be 1000 words

Different ways to represent text data...

- One-hot encoding → dimension = size of vocabulary

Human-Readable

Pet
Cat
Dog
Turtle
Fish
Cat

Machine-Readable

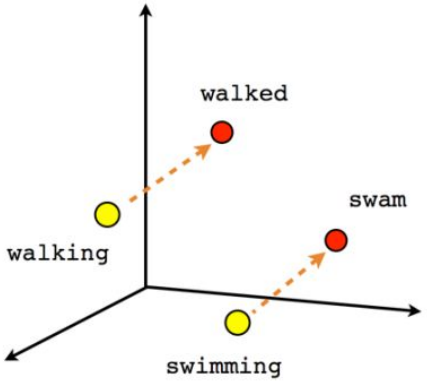
Cat	Dog	Turtle	Fish
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
1	0	0	0

Different ways to represent text data...

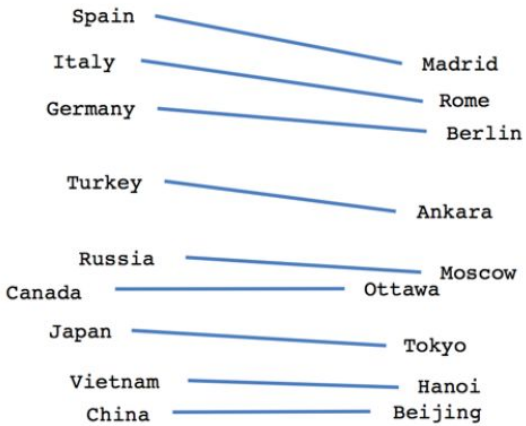
- Bag-of-words representation → dimension = size of vocabulary

Document	the	cat	sat	in	hat	with
the cat sat	1	1	1	0	0	0
the cat sat in the hat	2	1	1	1	1	0
the cat with the hat	2	1	0	0	1	1

Word vectors

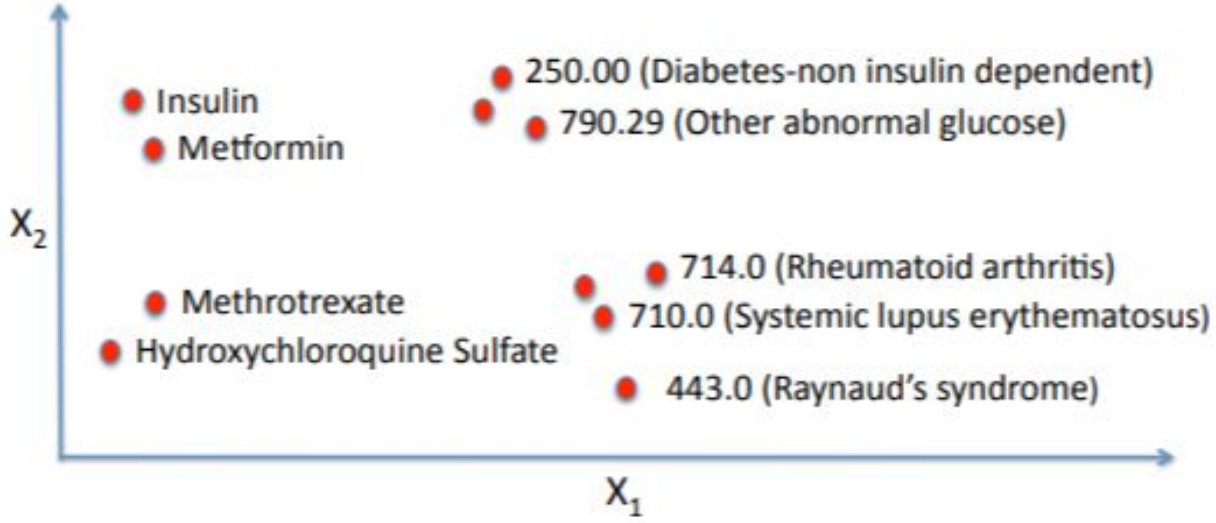


Verb tense



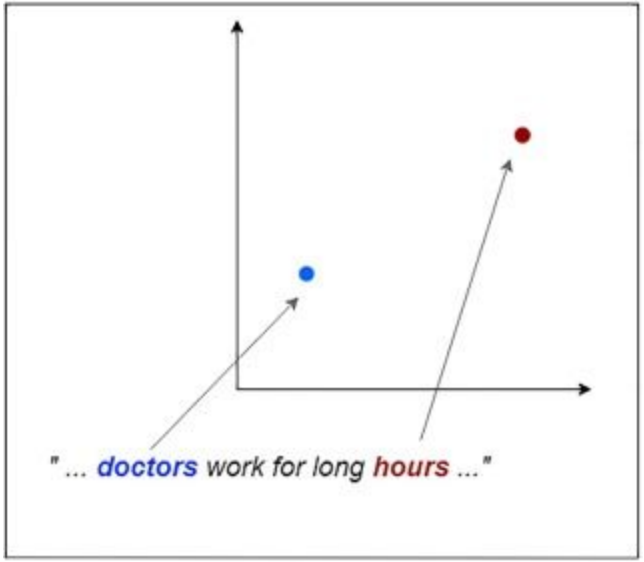
Country-Capital

Word vectors… with healthcare text data

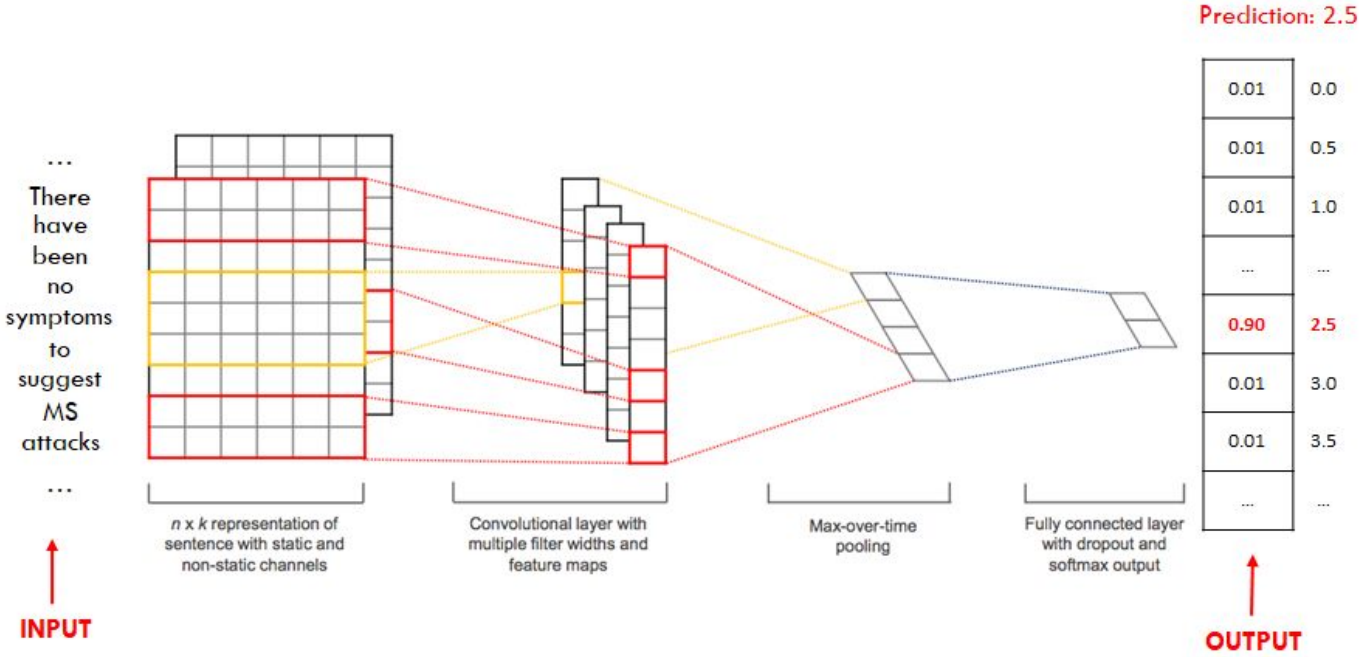


Choi, Y. et al., "Learning Low-Dimensional Representations of Medical Concepts", 2016.

Using word vectors



Model: Convolutional Neural Networks



Kim, Y., 2014. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

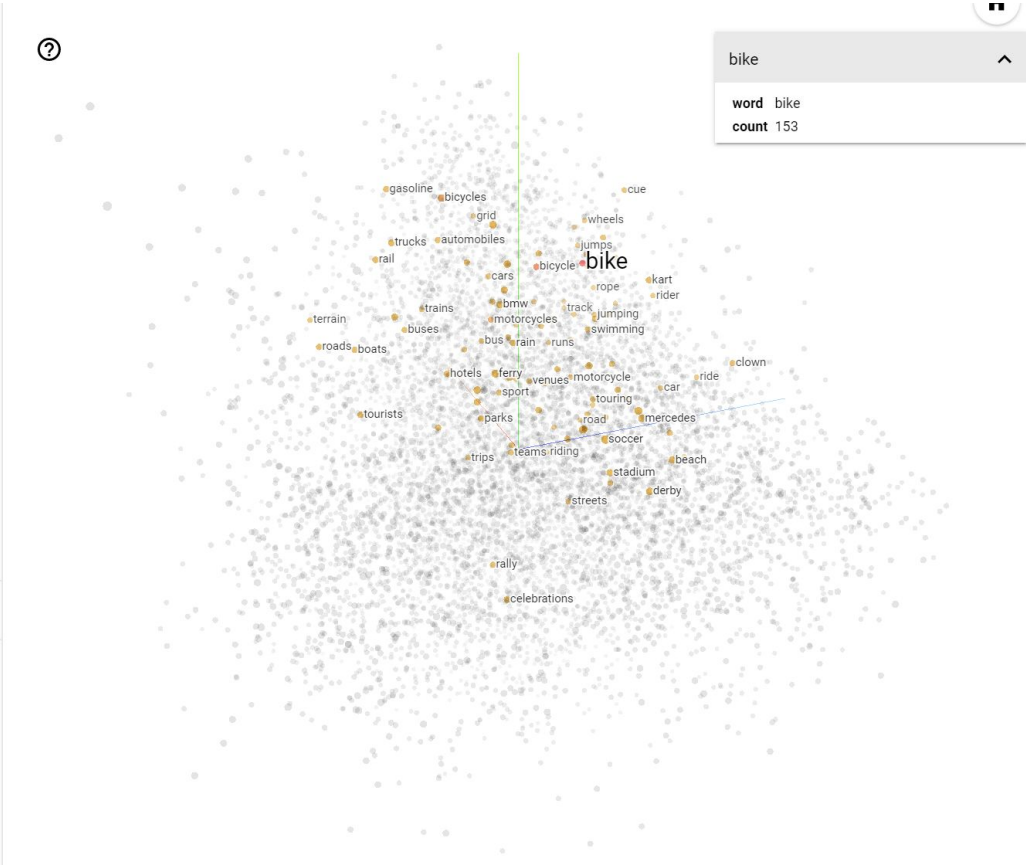
Rule-based model + ML model

Different word embeddings for healthcare data

- PubMed+PMC word2vec: https://bio.nlplab.org/
- MS-BERT: https://nlp4h.com/blog/ms_bert_intro/
- And more!!!

Tools used in this project

- Spacy: https://spacy.io/ → text pre-processing
- Gensim: https://radimrehurek.com/gensim/models/word2vec.html → train you own word2vec models
- Tensorflow embedding projector: https://projector.tensorflow.org/ → visualize word vectors



Outro

Bad labels = Bad model

- Labels/Annotations:
  - Use case #1 (TB risk factors): each note needs to be labeled
  - Use case #2 (GIM): outcome is present in the structure data
  - Use case #3 (MS): experts labeled the data… some variables are hard to label though…
- Be mindful of the number of people who are labeling your data.
- Is your model learning behavior X? Or is it learning what labeler A thinks of the data?
- Inter-annotator agreement.

Some NLP libraries

Python:

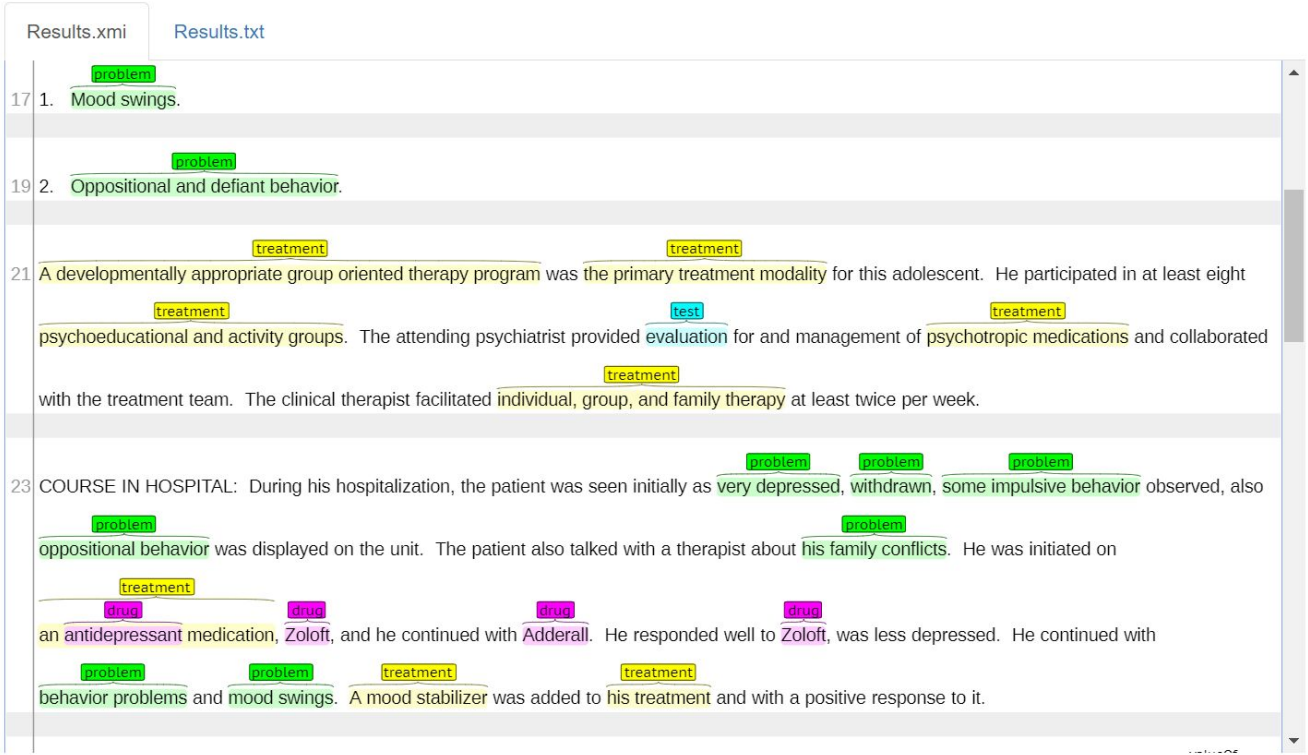
- spaCy: https://spacy.io/
- nltk: https://www.nltk.org/
- Gensim: https://radimrehurek.com/gensim/

R:

- Tidytext: https://cran.r-project.org/web/packages/tidytext/vignettes/tidytext.html
- Hunspell: https://cran.r-project.org/web/packages/hunspell/index.html

More tools!!!

- Information extraction tools: cTakes, MetaMap, CLAMP
- Wang, Y. et al., “Clinical information extraction applications: A literature review”, 2017.



https://clamp.uth.edu/clampdemo.php