NLP-driven interactive healthcare data visualization

Team 50 – The Data Tavern

Team members: Stefan Lehman, Lei Han, Kelly Lewis, Ruby Truong, Ying Wang, Luke Williams

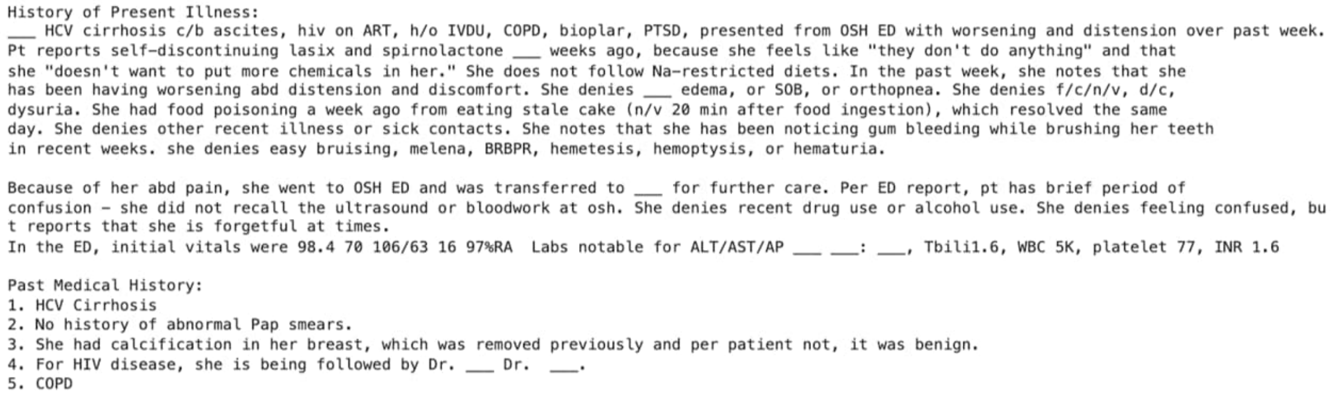
Georgia Tech

Project Objectives

- **Objectives:**
  - Utilize Natural Language Processing (NLP) to transforms raw clinical data into visual reports, and by identifying vital medical terms from clinical notes, expedite healthcare workers' review process by providing interactive visualizations, thereby enhancing their interactions with electronic health records (Abudiyab, 2022 [1]; Chishtie, 2022 [2]).

- **Project beneficiaries:**
  - These visualizations aim to assist healthcare professionals and stakeholders (hospitals, clinics, and billing departments) to access and interpret patient data effectively.

Raw clinical data

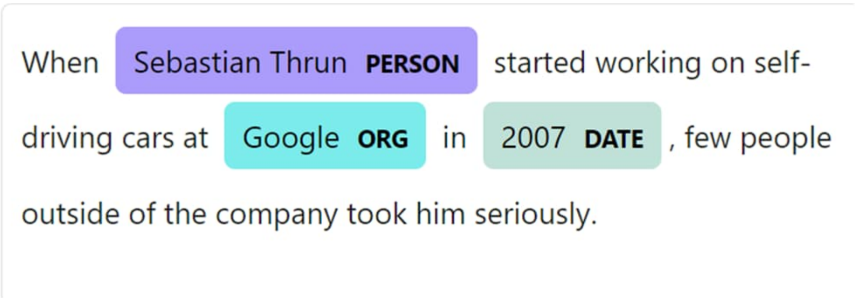


```
History of Present Illness:
___ HCV cirrhosis c/b ascites, hiv on ART, h/o IVDU, COPD, bioplar, PTSD, presented from OSH ED with worsening and distension over past week.
Pt reports self-discontinuing lasix and spirnolactone ___ weeks ago, because she feels like "they don't do anything" and that
she "doesn't want to put more chemicals in her." She does not follow Na-restricted diets. In the past week, she notes that she
has been having worsening abd distension and discomfort. She denies ___ edema, or SOB, or orthopnea. She denies f/c/n/v, d/c,
dysuria. She had food poisoning a week ago from eating stale cake (n/v 20 min after food ingestion), which resolved the same
day. She denies other recent illness or sick contacts. She notes that she has been noticing gum bleeding while brushing her teeth
in recent weeks. she denies easy bruising, melena, BRBPR, hemetesis, hemoptysis, or hematuria.

Because of her abd pain, she went to OSH ED and was transferred to ___ for further care. Per ED report, pt has brief period of
confusion - she did not recall the ultrasound or bloodwork at osh. She denies recent drug use or alcohol use. She denies feeling confused, bu
t reports that she is forgetful at times.
In the ED, initial vitals were 98.4 70 106/63 16 97%RA  Labs notable for ALT/AST/AP ___ ___: ___, Tbili1.6, WBC 5K, platelet 77, INR 1.6

Past Medical History:
1. HCV Cirrhosis
2. No history of abnormal Pap smears.
3. She had calcification in her breast, which was removed previously and per patient not, it was benign.
4. For HIV disease, she is being followed by Dr. ___ Dr.  ___.
5. COPD
```

Sample visual report



When Sebastian Thrun PERSON started working on self-driving cars at Google ORG in 2007 DATE , few people outside of the company took him seriously.

Expected innovations

- **Current limitations**
  - A 2017 study found that physicians spent 5.9 hours interacting with a computer per day (Arndt et al. 2017 [3])
  - Three primary tasks are examining chart review records, compiling visit documentation, and order entry (Overhage & McCallie, 2020 [4])
  - Contributes to physician burden, burnout, and dissatisfaction
- **Differences from prior approaches**
  - NLP techniques to highlight important details such as medication names, conditions, and procedures
  - Provide a patient filtering mechanism by ID, enabling personalized insights.
  - Comparison of patients with similar conditions, aiding in treatment planning
  - Will be successful by offering comprehensive interface for healthcare workers

Measuring Success and Risk Factors

- **Successful Impact and Measurement**
  - Healthcare professionals will experience time savings and patient care could be improved.
  - Success can be measured through comparing manual use of EHRs to our NLP-derived visualization tool to determine if time savings are statistically significant.
- **Risk and Payoffs**
  - Data security risk minimized with deidentification may reduce the performance and reliability of the NLP algorithm by eliminating information (Noor et al., 2022 [5]).
  - Lack the subject matter expertise, so we are treating this tool as a proof-of-concept.

Plan of Activities, Timeline and Cost

- ## Plan of Activities and Timeline

Week	Activity	Members	Key Milestones	Description
March 3 - 6	Conduct EDA (Part 1)	All members	Deliverable 1: Initial Insights	Initial exploratory data analysis to understand dataset characteristics and data quality issues.
March 7 - 11	Conduct EDA (Part 2)	All members	Deliverable 2: EDA Report	Continue EDA focusing on identifying patterns, anomalies, and relationships between variables.
March 12 - 17	Data Cleaning	Kelly, Han	Deliverable 3: Cleaned Dataset	Clean the dataset based on insights from EDA, handling missing values, outliers, and errors.
March 19 - 24	EHR Structuring	Han, Ying	Midterm 1: Process EHR Data	Break down the electronic health records into structured sections for each record, preparing for efficient querying and analysis.
March 25 - 31	Handling Big Data with SQLite or Pickle	Ruby, Ying	Midterm 1: Data Split for ML	Implement SQLite or use Pickle files for managing big data. Split the dataset into training, testing, and validation datasets.
April 1 - 6	SpaCy Development for Structured EHR Data	Kelly, Stefan	Deliverable 4: NLP Model Development	Develop and train custom SpaCy models for NLP tasks specific to the structured EHR data.
April 7 - 11	Integration of Data Querying with NLP	Luke, Ruby	Final: NLP Integration with Data	Ensure efficient querying of structured data for use with NLP models, optimizing for performance.
April 12 - 19	App Development with Visualization	Stefan, Luke	Final: Visualization Web App	Develop a web application for visualizing EHR data and NLP analysis results, focusing on UI
April 20 - 21	Conduct Statistical Analysis and Impact	All members	Final: Final Analysis Report	Apply statistical methods and machine learning models for in-depth analysis of EHR data. Interpret results to derive insights.

- ## Cost Estimation:
  - Local development will cost $0. AWS's total costs for 8 weeks of development include $72.00 for SageMaker, $12.24 for EC2, and $0.076 for S3, totaling $84.31. GCP's costs would be $314.64 for Vertex AI, $79.20 for AppEngine, $0.066 for Cloud Storage, summing up to $393.9.

Reference

- [1] Abudiyab NA, Alanazi AT. Visualization Techniques in Healthcare Applications: A Narrative Review. Cureus. 2022 Nov 11;14(11):e31355. doi: 10.7759/cureus.31355. PMID: 36514654; PMCID: PMC9741729. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9741729/
- [2] Chishtie J, Bielska I, Barrera A, Marchand J, Imran M, Tirmizi S, Turcotte L, Munce S, Shepherd J, Senthinathan A, Cepoiu-Martin M, Irvine M, Babineau J, Abudiab S, Bjelica M, Collins C, Craven B, Guilcher S, Jeji T, Naraei P, Jaglal S. Interactive Visualization Applications in Population Health and Health Services Research: Systematic Scoping Review. J Med Internet Res 2022;24(2):e27534. https://www.jmir.org/2022/2/e27534
- [3] Brian G. Arndt, John W. Beasley, Michelle D. Watkinson, Jonathan L. Temte, Wen-Jan Tuan, Christine A. Sinsky & Valerie J. Gilchrist. (2017). Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations. The Annals of Family Medicine September 2017, 15 (5) 419-426. https://www.annfammed.org/content/15/5/419.full
- [4] J. Marc Overhage & David McCallie Jr. (2020, January 14). Physician Time Spent Using the Electronic Health Record During Outpatient Encounters. Annals of Internal Medicine. https://www.acpjournals.org/doi/10.7326/M18-3684
- [5] Noor Abu-el-rub, Jay Urbain, George Kowalski, Kristen Osinski, Robert Spaniol, Mei Liu, Bradley Taylor, & Lemuel R. Waitman. (2022, May 23). Natural Language Processing for Enterprise-scale De-identification of Protected Health Information in Clinical Notes. AMIA Summits on Translational Science Proceedings 2022: 92–101. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9285160/