

Product name: Understanding Healthcare Data

Team name: Anthem Sponsored

Release name: V2.0

Release date: 3/13/2020

Revision number: 2.0

Revision date: 4/9/2020

#### High-Level Goals:

- Can you predict recurrence or rehospitalization for a given chronic disease given an arbitrarily sized window of medical history?

#### User Stories:

- Sprint 1 (3/30/2020 - 4/12/2020)
  - Create a model to train on word embeddings created given the English descriptions of medical codes present in a patient's data in a given month.
- Sprint 2 (4/13/2020 - 4/26/2020)
  - Get running on hummingbird
    - Figure out if we need access to AWS accounts
  - Work on analyzing the model to determine which factors it uses to determine whether a patient will be rehospitalized
    - Determine if current model is statistically reliable
  - Determine if the model can be generalized to myocardial infarction
  - Work with larger data set to train model
  - Start specifying how to package up model and data pipeline so Rob can run it
    - Build some trivial container as a start
- Sprint 3 (4/27/2020 - 5/10/2020)
  - Apply model to a chronic disease: myocardial infarction
  - Combine scripts to generate sentences and embedding matrices
    - Generate 6 month window list of lists of strings containing all of the medical codes that show up for a given patient in that given window (per bucket) [matrix][month][code]
    - In parallel, collapse all of the 6 month windows, removing any empty windows, for the purpose of creating a master list of sentences
    - Train a word2vec model on this master list of sentences

- Use the word embeddings generated by the word2vec model to convert each of the 6 month windows into 6 x 100 matrices by averaging the embedding vectors for each of the sentences per window
      - If a given sentence is empty, replace that sentence with the 0 vector
    - Dump the resulting matrix, label pairs into a .npy file for consumption by the training script
  - Turn training notebook into a python script
    - Download and refactor code such that it will read in a .npy file and perform the necessary transformations to feed the data into NNs
  - Prepare a docker container to run all of the scripts, starting from JSON files and ending with trained models
  - ASK ROB:
    - Duplicate codes?
  - “Engineering”
    - Tighten pipeline and ceremonies
    - Package up the above and send to Rob to run on real data
- Sprint 4 (5/11/2020 - 5/24/2020)
  - Have Rob run on real data
  - Have first draft of three figures
  - First VC pitch for our service
- Sprint 5 (5/25/2020 - 6/7/2020)
  -