International Journal of Population Data Science

Journal Website: www.ijpds.org

Learning Unsupervised Representations from Biomedical Text

Meaney, C[1], Tu, K[1], Jaakkimainen, L[2], Escobar, M[1], Rudzicz, F[3], and Widdifield, J[2]

[1]University of Toronto

[2]Institute for Clinical Evaluative Sciences

[3]University Health Network

Introduction

Healthcare settings are becoming increasingly technological. Interactions/events involving healthcare providers and the patients they service are captured as digital text. Healthcare organizations are amassing increasingly large/complex collections of biomedical text data. Researchers and policy makers are beginning to explore these text data holdings for structure, patterns, and meaning.

Objectives and Approach

EMRALD is a primary care electronic medical record (EMR) database, comprised of over 40 family medicine clinics, nearly 400 primary care physicians and over 500,000 patients. EMRALD includes full-chart extractions, including all clinical narrative information/data in a variety of fields.

The input data (raw text strings) are discrete, sparse and high dimensional. We assessed scalable statistical models for high dimensional discrete data, including fitting, assessing and exploring models from three broad statistical areas: i) matrix factorization/decomposition models ii) probabilistic topic models and iii) word-vector embedding models.

Results

EMRALD is comprised of 12 text data streams. EMRALD text data is structured into 84 million clinical notes (3.5 billion word/language tokens) and is approximately 18Gb in storage size. We employ a "text as data" pipeline, i) mapping raw strings to sequences of word/language tokens, ii) mapping token sequences to numeric arrays, and finally iii) using numeric arrays as inputs to statistical models.

Fitted topic models yield useful thematic summaries of the EMRALD corpora. Topics discovered reflect core responsibilities of primary care physicians (e.g. women's health, pain management, nutrition/diet, etc.).

Fitted vector embedding models capture structure of discourse/syntax. Related words are mapped to similar locations of vector spaces. Analogical reasoning is possible in the embedding space.

Conclusion/Implications

"Text as data" requires an understanding of statistical models for discrete, sparse, high dimensional data. We fit a variety of unsupervised statistical models to biomedical text data. Preliminary results suggest that the learned low dimensional representations of the biomedical text data are effective at uncovering meaningful patterns/structure.

http://dx.doi.org/10.23889/ijpds.v3i4.760