NLP Structured Data Investigation on Non-Text

Casey Stella

Hortonworks

Spring, 2015

Table of Contents

Introduction

- I'm a Principal Architect at Hortonworks
- I work primarily doing Data Science in the Hadoop Ecosystem
- Prior to this, I've spent my time and had a lot of fun
  - Doing data mining on medical data at Explorys using the Hadoop ecosystem
  - Doing signal processing on seismic data at Ion Geophysical using MapReduce
  - Being a graduate student in the Math department at Texas A&M in algorithmic complexity theory

Domain Challenges in Data Science

A data scientist has to merge analytical skills with domain expertise.

- Often we're thrown into places where we have insufficient domain experience.
- Gaining this expertise can be challenging and time-consuming.
- Unsupervised machine learning techniques can be very useful to understand complex data relationships.

Domain Challenges in Data Science

A data scientist has to merge analytical skills with domain expertise.

- Often we're thrown into places where we have insufficient domain experience.
- Gaining this expertise can be challenging and time-consuming.
- Unsupervised machine learning techniques can be very useful to understand complex data relationships.

We'll use an unsupervised structure learning algorithm borrowed from NLP to look at medical data.

Word2Vec

Word2Vec is a vectorization model created by Google [1] that attempts to learn relationships between words automatically given a large corpus of sentences.

- Gives us a way to find similar words by finding near neighbors in the vector space with cosine similarity.

[1] http://radimrehurek.com/2014/12/making-sense-of-word2vec/

Word2Vec

Word2Vec is a vectorization model created by Google [1] that attempts to learn relationships between words automatically given a large corpus of sentences.

- Gives us a way to find similar words by finding near neighbors in the vector space with cosine similarity.
- Uses a neural network to learn vector representations.

[1] http://radimrehurek.com/2014/12/making-sense-of-word2vec/

Word2Vec

Word2Vec is a vectorization model created by Google [1] that attempts to learn relationships between words automatically given a large corpus of sentences.

- Gives us a way to find similar words by finding near neighbors in the vector space with cosine similarity.
- Uses a neural network to learn vector representations.
- Recent work by Pennington, Socher, and Manning [2] shows that the word2vec model is equivalent to weighting a word co-occurance matrix weighting based on window distance and lowering the dimension by matrix factorization.

[1] http://radimrehurek.com/2014/12/making-sense-of-word2vec/

Word2Vec

Word2Vec is a vectorization model created by Google [1] that attempts to learn relationships between words automatically given a large corpus of sentences.

- Gives us a way to find similar words by finding near neighbors in the vector space with cosine similarity.
- Uses a neural network to learn vector representations.
- Recent work by Pennington, Socher, and Manning [2] shows that the word2vec model is equivalent to weighting a word co-occurance matrix weighting based on window distance and lowering the dimension by matrix factorization.

Takeaway: The technique boils down, intuitively, to a riff on word co-occurence. See here[1] for more.

[1] http://radimrehurek.com/2014/12/making-sense-of-word2vec/

Clinical Data as Sentences

Clinical encounters form a sort of sentence over time. For a given encounter:

- Vitals are measured (e.g. height, weight, BMI).
- Labs are performed and results are recorded (e.g. blood tests).
- Procedures are performed.
- Diagnoses are made (e.g. Diabetes).
- Drugs are prescribed.

Each of these can be considered clinical "words" and the encounter forms a clinical "sentence".

Clinical Data as Sentences

Clinical encounters form a sort of sentence over time. For a given encounter:

- Vitals are measured (e.g. height, weight, BMI).
- Labs are performed and results are recorded (e.g. blood tests).
- Procedures are performed.
- Diagnoses are made (e.g. Diabetes).
- Drugs are prescribed.

Each of these can be considered clinical "words" and the encounter forms a clinical "sentence".
Idea: We can use word2vec to investigate connections between these clinical concepts.

Demo

As part of a Kaggle competition[2], Practice Fusion, a digital electronic medical records provider released depersonalized clinical records of 10,000 patients. I ingested and preprocessed these records into 197,340 clinical "sentences" using Pig and Hive.

[2] https://www.kaggle.com/c/pf2012-diabetes

Demo

As part of a Kaggle competition[2], Practice Fusion, a digital electronic medical records provider released depersonalized clinical records of 10,000 patients. I ingested and preprocessed these records into 197,340 clinical "sentences" using Pig and Hive.

MLLib from Spark now contains an implementation of word2vec, so let's use pyspark and IPython Notebook to explore this dataset on Hadoop.

[2] https://www.kaggle.com/c/pf2012-diabetes

Questions

Thanks for your attention! Questions?

- Code & scripts for this talk available on my github presentation page.[3]
- Find me at http://caseystella.com
- Twitter handle: @casey_stella
- Email address: cstella@hortonworks.com

[3] http://github.com/cestella/presentations/

Bibliography

[1] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[2] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics, 2014.