6.806- Advanced Natural Language Processing
Massachusetts Institute of Technology December 14, 2015

Sifter, a New Machine Learning Application for Clustering Medical Research Findings

Name: Winter Guerra

Collaborators: None

Code Repository:
`https://github.com/Winter-Guerra/6.806_nlp_cancer_research`

Dataset Location:
`http://nlp-dataset-6806-2015.s3.amazonaws.com/index.html`

1 Abstract

2 Introduction

The quantity of medical and scientific literature available to the average scientist is increasing at a rapid pace. However, there is currently no good method for easily extracting information from this multitude of data without extensive human interaction. As a result of this inability to easily sift through data, many important findings from cutting edge medical research go unnoticed by the rest of the scientific community. What is needed is a new tool to simplify the act of organizing medical research data based on clusters of findings and topics. This is what my project, Sifter, aims to do.

Utilizing Amazon Web Services's powerful backend, Sifter cross-references NIH's PubMed Open Access dataset of 1,156,698 full-text XML medical research papers and 82,448 meta-research articles to automatically create training clusters of article topics without human interaction. However, after much trail and error testing multiple feed-forward neural network designs on the data that Sifter created, we were unable to perform better than a random data baseline after 75 epochs. Nonetheless, we believe that this result is due to issues with our implementation of our neural network models and could be remedied in the future.

To assist in advancing this goal, we have published the dataset that Sifter created [here] for the convenience of all researchers whom wish to explore the applications of the Sifter dataset to clustering scientific articles.

2.1 Problem

One of the biggest issues with grouping scientific articles is that multiple higher level factors are involved in defining whether one scientific article is similar to the other. For example, two articles

will not be similar if they have done their research on different patient subsets (i.e. male and female test subjects). Because of the complexity that factors into scientific article similarity, labeling of a ground-truth article similarity dataset is commonly done by hand. However, this is very expensive and time-consuming.

2.2 Approach

2.2.1 Creating the training dataset

To create our ground-truth dataset of article similarity metrics, we performed the following steps.

- Downloaded the NIH Open Access Subset to AWS EBS storage.
- Extracted document-type metadata from all articles in the NIH dataset to get an in-memory index of all relevant and irrelevant articles.
- Extracted 41 million citations from all meta-research articles in addition to their in-text citation locations.
- Pruned out all citations that linked to articles outside of our dataset.
- Created a similarity distance matrix from all remaining citations.
- Saved the similarity matrix in the form of a sparse distributed hashtable in the Redis in-memory cache for fast read access.

2.2.2 Calculating Document Similarity Using a Deep Multilayer Perceptron

Using the citation distance data harvested from PubMed, we trained a Deep Multilayer Perceptron neural network that takes in a concatenated bag-of-words vector embedding of the summary sentence of 2 articles, then outputs a binary classification of whether the two articles are similar or not (see 1). This feed

3 Experiment

4 Next Steps

As we can see from our results in 1, there is a lot of room for improvement regarding Sifter's neural network implementation. However, we already have multiple hypothesis for how we can improve Sifter's performance.
First, instead of using a Feed Forward model that uses a vectorized bag-of-words document summary for creating similarity metrics, we would use a sequence-conscious Siamese Convolutional

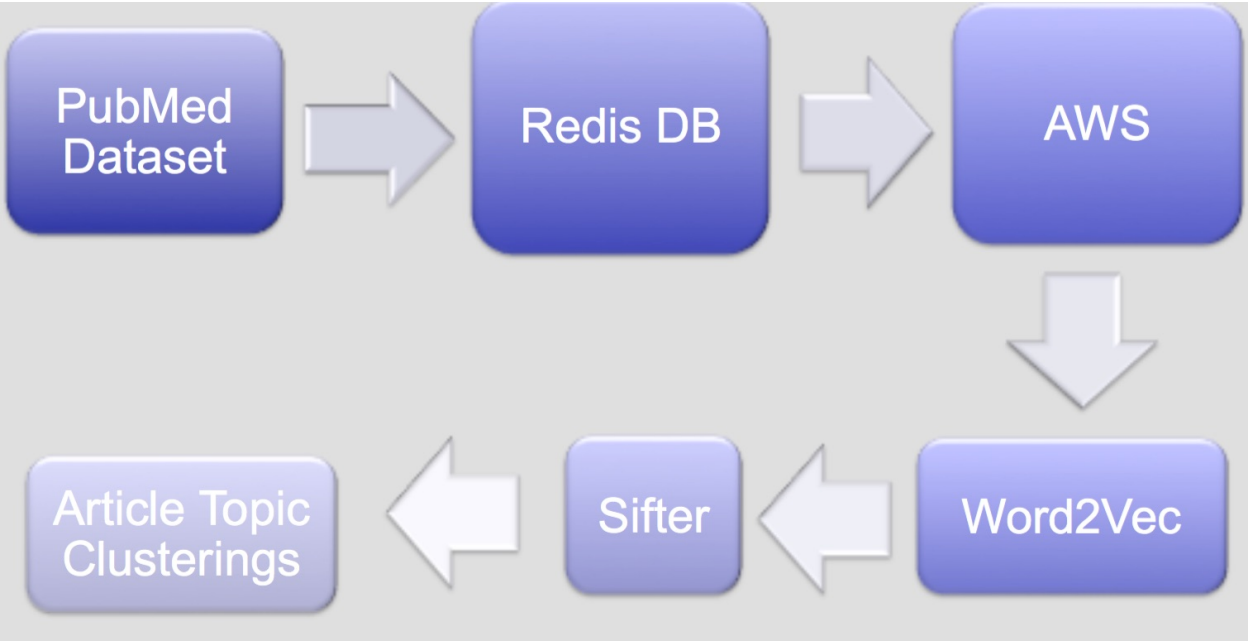

Figure 1: Flowchart of data flow in Sifter.

Neural Network (see figure 3). This type of model has already been proven to work well for calculating similarity metrics between image inputs (in the form of the MNIST Digit Dataset), but has also seen some use for comparing similarity metrics of documents on a sentence level.

5 Conclusion

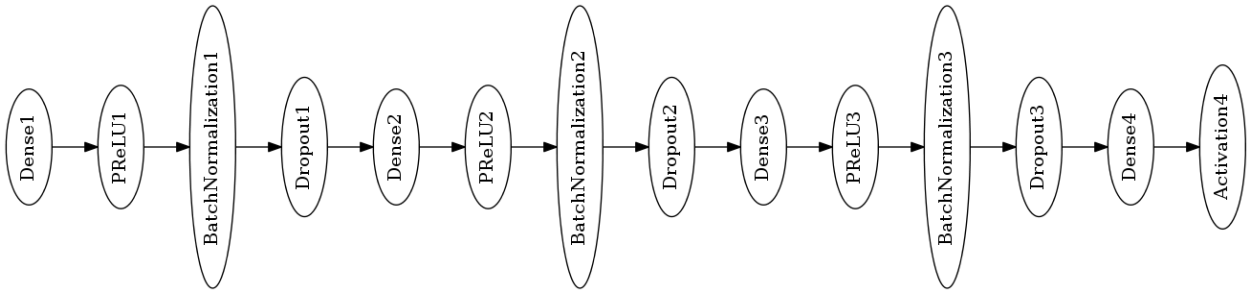

Figure 2: The Deep Multilayer Perceptron feed forward network architecture tested in this writeup for Sifter.

Table 1: Sifter Deep Multilayer Perceptron Neural Network Accuracy vs Random Choice Baselines

Method	Accuracy
Sifter DMLP Network with 75 training epochs	0.500
Random +/- choice on training set	0.500
Random +/- choice on test set	0.500
Total ratio of +/- labels in test & training set	0.489

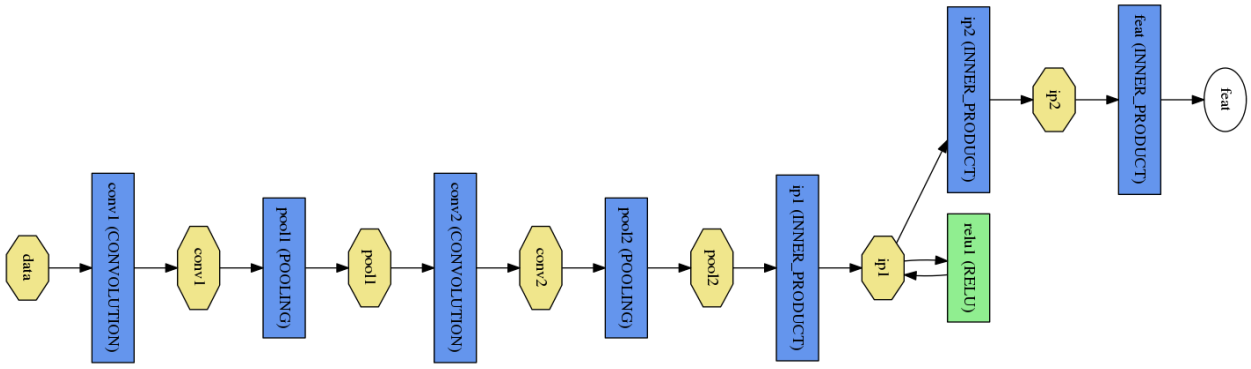

Figure 3: The convolutional Siamese neural network model that we plan to test in the future against Sifter's training data (which performs well on MNIST character similarity tests).