6.806- Advanced Natural Language Processing
Massachusetts Institute of Technology December 13, 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

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. Using this training set, Sifter is able to generate a neural network model that can also cluster new incoming articles in an online manner. The specifics of Sifter's neural network accuracy will be revealed during the 6.806 final poster presentation for the class.

2 Introduction

2.1 Problem

2.2 Approach

3 Background: Related Work

4 Approach

5 Experiment

6 Conclusion