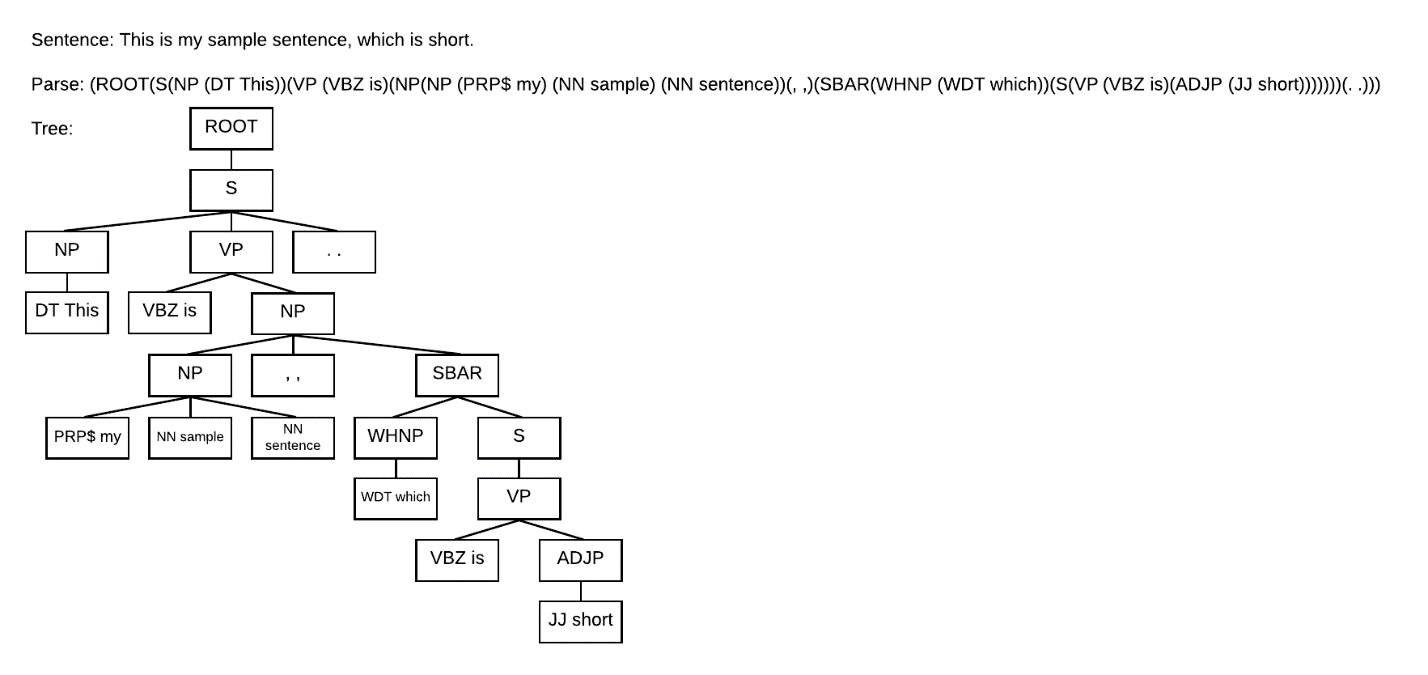
Maintaining Parse Structure for Structural Similarity Analysis in Machine Learning Algorithms

By Garrick James McMickell

# The Existing Structure



# The Problem

Existing textual feature extraction methods for use in machine learning algorithms generally focus of the “Bag of Words” representation, wherein each word is tokenized—given an integer value—and counted. The result of this is a one-dimensional array, where the index is the token value, and the value at that index is the count of that type of token. These individual arrays can then be arranged into a matrix—each row being an individual text sample; each column representing the count of the tokens, as before. This method, however, holds no structural or positional data, allowing only the analysis of term-frequency or tf-idf—inverse-document-frequency—which becomes a problem when trying to analyze and compare structural information within a parse string.

Using the Bag of Words approach, it is possible to use each part of speech in the parse string and tokenize it, but in doing so, we would eliminate the overall parse structure. Outside of the Bag of Words, there appears to be no meaningful way of assessing structural similarity within machine learning algorithms—such algorithms rely on quantitative data, and a structure, the parse tree, is qualitative.

This is the primary issue in attempting to identify structural similarity between two pieces of text: we need to generalize the structure in such a way that it can be evaluated quantitatively.