Consistency scores in text data [*]

Rohan Alexander   *University of Toronto*
Ke-Li Chiu   *University of Toronto*

. . .

Introduction

When we think about a data science project, we may like to think that our job is to 'let the data speak'. But this is rarely the case in practise. Datasets can have errors, be biased, incomplete, or messy. In any case, it is the underlying statistical process of which the dataset is an artifact that is typically of interest. In order to use statistical models to understand that process we often need to prepare the dataset in some way. This is particularly the case when working with text. However, this process requires many decisions. Should we correct obvious errors? What about slightly-less-obvious errors? To what extent have we introduced new errors? Have we made decisions that have affected, or even driven, our results?

In this paper we introduce the concept of consistency for a text corpus. Consistency refers to the proportion of words that are able to be predicted by a trained model based on the previous words and surrounding context. Further, we define internal consistency as when the model is trained on the corpus itself, and external consistency as when the model is trained on a more general corpus. Together, these concepts provide a guide to the cleanliness and consistency of a text dataset. This can be important when deciding whether a dataset is fit for purpose; when carrying out data cleaning and preparation tasks; and as a comparison between datasets.

To provide an example, consider the sentence, 'the cat in the. . . '. A child who has read this book could tell you that the next word should be 'hat'. Hence if the sentence was actually 'the cat in the bat', then that child would know something was likely wrong. A consistency score would likely be lower than if the sentence were 'the cat in the hat'. After the researcher corrects this error, a consistency score would likely increase. By following how the consistency scores change during the data preparation and cleaning stages the researcher can better understand the effect of the changes. Including consistency scores when corpuses are shared allows researchers to be more transparent about their corpus. And finally, the use of consistency scores allows for automation in the cleaning process.

We apply our approach to X (one option is a Hansard, but it's really big, so yeah, I don't really want to do that). Additionally, we construct a Shiny app that computes internal and external consistency scores for smaller corpuses and allows the researcher to make changes and see how it updates.

The remainder of our paper is structured as follows. . .

Background

A typical data science workflow involves
    Internal and external consistency

Stylized example

Application

Discussion

Internal validity vs external- one is their own words the other is a general set of words.
    In the same way that precision and recall provide important measures. . .