Confounder

Inferring errors of omission

OVERVIEW

News articles and studies that attempt to infer a causal relationship (or lack thereof) between two variables often neglect to take important factors into account. They may fail to ***accurately measure*** the two things being compared, they may ***omit important variables and trade-offs***, and they may fail to demonstrate ***which variable affects the other*** (confusing cause and effect). Confounder is a machine learning program that aims to detect such errors using *domain-specific criteria* and corresponding *text samples* that are representative of such errors/omissions.

Example: *What impact does college education have on lifetime earnings?*

This question often has a number of omitted variable and trade-offs, which include:

1. The sunk cost of not working as much during school years
2. Any tuition costs including interest on student loans, residency, etc
3. Super-earner outliers skewing the average for degree recipients
4. Aggregating all majors and all non-degree holders into two groups
5. Those that complete college may be prone to succeed beforehand

Example Workflow

Sample articles and studies that discuss the “college premium” are gathered, read, and labeled by which of the five above errors/omissions are made. Once a representative corpus of text has been gathered, they are all pre-processed into a dataframe containing the name of the text, the raw text itself, and five more columns indicating which of the five factors were mentioned. These are used to iteratively train a natural language processing classifier to accurately detect whether any of the five factors were omitted or not in future articles and studies, rather than having to read them all by hand long after they have been published.

Development:

Future improvements and serious usage of Confounder will largely consist of the following:

1. Applying the workflow above to new subjects involving alleged causes and effects
2. Improving the predictive accuracy of the program as well as updating topic criteria
3. Deploying Confounder for the purpose of tracking news and research in real-time