Text Selection

Text data is inherently high-dimensional, which makes machine learning regularization techniques natural tools for its analysis. Text is often selected by journalists, speechwriters, and others who cater to an audience with limited attention. We develop an economically-motivated high dimensional selection model that can improve machine learning from text in particular and from sparse counts data more generally. Our highly scalable approach to modeling coverage selection is especially useful in cases where the cover/no-cover choice is separate or more interesting than the coverage quantity choice. We apply this framework to option-implied volatility (VIX) prediction using newspaper coverage, and find that it substantially improves out-of-sample fit relative to alternative state-of-the-art approaches. This advantage increases with the sparsity of the text.