DiscoverText: Power Tools for Human & Machine Learning

Dr. Stuart W. Shulman[1]

[1] Founder & CEO, @Texifter stu@texifter.com

Abstract: Participate in this workshop to learn how to build custom machine classifiers for sifting Twitter data. The topics covered include how to: construct precise Twitter data fetch queries, use Boolean search on resulting archives, filter on metadata or other project attributes, tabulate, explore, and set aside duplicates, cluster near-duplicates, crowd source human coding, measure inter-rater reliability, adjudicate coder disagreements, and build high quality word sense and topic disambiguation engines. DiscoverText is designed specifically for collecting and cleaning up messy Twitter and other text data streams. Use basic research measurement tools to improve human and machine performance classifying data over time. The demo covers how to reach and substantiate inferences using a theoretical and applied model informed by a decade of interdisciplinary, National Science Foundation-funded research into the text classification problem. The major idea of the workshop is that when training machines for text analysis, greater reliance should be placed on the input of those humans most likely to create a valid observation. Texifter proposed a unique way to recursively validate, measure, and rank humans on trust and knowledge vectors, and called it CoderRank.

Keywords: coding, adjudication, machine learning, Twitter, metadata

Necessary Resources: A laptop with broadband access to the Internet.

Biographical Note: Entrepreneur, U.S. Soccer C-licensed coach, ODP and NEFC-West, proud parent & son, CEO @Texifter, inventor of @DiscoverText, and Taoist garlic grower.