Computer Science Colloquium
"Bayesian Models for Mining Public Health Information from Twitter"

Dr. Mark Dredze

Assistant Research Professor
Johns Hopkins University

Abstract

Twitter and other social media sites contain a wealth of information about populations and have been used to track sentiment towards products, measure political attitudes, and study social linguistics. In this talk, we investigate the potential for Twitter to impact public health research. Specifically, we consider population surveillance, a major focus of public health that typically depends on clinical encounters with health professionals to collect patient data. Individual users often broadcast salient health information, such as "sick with this flu fever taking over my body ughhhh time for tylenol", which indicates that not only does this person have the flu, but also a fever and is self-medicating with tylenol. Aggregating such content across millions of users could provide information about numerous aspects of illnesses in the population.

In this work we present the Ailment Topic Aspect Model (ATAM), a new Bayesian graphical model for Twitter that associates symptoms, treatments and general words with diseases (ailments.) When applied to 1.6 million health related tweets, ATAM discovers descriptions of diseases in terms of collections of words (symptoms and treatments) and partitions messages based on the referenced disease. The model discovers diseases corresponding to influenza, infections, obesity, insomnia, and several others. Furthermore, we demonstrate the effectiveness of this model at several tasks: tracking illnesses over times (syndromic surveillance), measuring behavioral risk factors, localizing illnesses by geographic region, and analyzing symptoms and medication usage. We show quantitative correlations with public health data and qualitative evaluations of model output. Our results suggest that Twitter has broad applicability for public health research.

Biography

Mark Dredze is an Assistant Research Professor in Computer Science at Johns Hopkins University and a research scientist at the Human Language Technology Center of Excellence. He is also affiliated with the Center for Language and Speech Processing and the Center for Population Health Information Technology. His research in natural language processing and machine learning has focused on graphical models, semi-supervised learning, information extraction, large-scale learning, and speech processing. His recent work includes health information applications, including information extraction from social media, biomedical and clinical texts. He obtained his PhD from the University of Pennsylvania in 2009.

Date: Friday, March 9, 2012
Time: 1:00pm to 2:00pm
Location: ECS South 2.410

Refreshments will be served at 12:45pm