Department of Mathematics, Statistics and Computer Science

COLLOQUIUM ANNOUNCEMENT

Preserving Privacy in Statistical Models of Text

XandaSchofield

Departmentof Computer Science
Cornell University

1:00 PM, WednesdayJanuary 16, 2019

Cudahy Hall, Room 401

Abstract

Machine learning models, such as topic models and word embeddings, detect statistical patterns in text collections much too large for a human to read. However, these tools may reveal sensitive information particular to individuals, like a unique conversation or medical report. In this talk, I discuss a way to preserve people's privacy when inferring topic models by introducing randomness to word counts before model inference. I show that, by accounting for how this random noise changes the statistical distributions of words in a collection, we can obtain topic models similar to those we would get from the original text while enforcing specific privacy guarantees.

Bio

Xanda Schofield is a PhD candidate in Cornell University's Department of Computer Science, where she is advised by David Mimno. Her work focuses on the practical aspects of using distributional semantic models for analysis of real-world datasets, with problems ranging from understanding the consequences of data preprocessing on model inference to enforcing text privacy for these models. Prior to her Ph.D. work, Xanda received her B.S. from Harvey Mudd College in Computer Science and Mathematics. She is the recipient of the NDSEG Fellowship in 2016. When not doing computer science, she likes to bake computationally inspired cookies and practicing Aikido.

3267&2//248,805()5(6+0(176 6(5,19('
&8'$+<+$//52200 $700:30