Analysis of the Spatial Distribution of Semantic Meaning using Tweets from Manhattan

Angela Zhou[1]

1. Princeton University
*Contact author: angelaz@princeton.edu

Keywords: topic modeling, social media, spatial clustering, natural language processing, remote sensing

How do people talk about their immediate surroundings: where they are, what they're doing, and how they feel about the places they inhabit, work, and live in? We consider this broad question in the framework of natural language processing and data science, and model the spatial distribution of natural language and its semantic meaning. The abundance of information from social media, including metadata such as GPS location, and the ease of programmatic access, motivates our use of Twitter data as a source of geocoded natural language utterances. We develop a framework for developing customized natural-language data sets from social media, using singular value decomposition (SVD) for topic modeling from the **gensim** package to transform raw unstructured text into a semantic vector space. In particular, we use historical archives of tweets from the Internet Archive and Twitter's Streaming API to collect a dataset of geocoded tweets from Manhattan, which we then aggregate by their location to finally arrive at documents originating from a grid of locations in Manhattan, each document representing tweets from a 500m radius. We analyze the spatial distribution of these semantic vectors in a new transformed vector space which incorporates aspects of semantic similarity between vectors and standard distances between locations in Manhattan. The noise in the dataset prevents the use of the gap statistic, a measure of intra-cluster variance, to choose an optimal number of clusters for K-Medoids clustering, but we devise a custom distance metric and find that this improves cluster performance. Finally, we refine the dataset using SVM classification on manually annotated tweets to arrive at a cleaner dataset of tweets that are explicitly about place, and repeat the clustering analysis. We find that there appears to be more internal structure, and the intra-cluster variance decreases on our clusterings. In the analysis, we illustrate the relevance of this methodology for "remote sensing" of the spatial distribution of human text and semantic information, and also note interesting features that would be of interest to city planners and urban developers. Analysis and data collection, which was primarily conducted in *Python*, could easily be converted to a streaming application.