Feature-rich event detection in social media streams.

Abstract:

Social networking platforms such as Twitter have emerged in recent years, creating a radically new mode of communication between people. Monitoring and analysing rich and continuous flow of user-generated content can yield unprecedentedly valuable information, which would not have been available from traditional media outlets.However, learning from Twitter streams poses new challenges, as compared to traditional media.

Traditional approaches to information extraction from microblog streams involve clustering based on semantic features of tweets. In this project, we implement and study different clustering models based on certain Twitter-specific features which include geo-positional data, timestamps, hashtags and check-ins. We also perform clustering based on bursty behaviour of n-grams in Twitter documents.

The resulting clusters contain valuable information about live events, but they also do contain a lot of noise. We treat this as a binary classification problem and use Machine Learning algorithms to classify and rank the clusters based on their textual, social and temporal features. Finally, we summarise top-ranking clusters and output the results to the end user.