# LSH based New-Event-Detection on twitter

# (python version)

# Introduction

This is an implementation of New-Event-Detection on twitter microblogging system in python. The algorithm we implement we presented in:

Petrović, S., Osborne, M., & Lavrenko, V. (2010). Streaming first story detection with application to twitter. NAACL HLT 2010 - Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference, (June), 181–189.

# Main Modules

In this section I present the main modules in this implementation.

Main

Clustering

LSH-Cosine Module

LSH Hashtable

Parameters:  
# hashtables

Parameters:  
# documents

Properties

mongoDB

Parameters:  
# hyper planes  
# max bucket size

Parameters:  
\* Clustering threshold

json

Text Reader

DB Reader

Streamer

OR

## Main

## Streamer Framework

### TweetListener

### DB Streamer

### Text File Streamer

## NED Main Module

## LSH Module

### LSH-Cosine Similarity Module

### LSH Hashtable

## Performance Test

I have turned off all logs and then ran the process with different sets of tweets measuring the time every time. The graph plotted shows that the performance is linear to the number of tweets every time.

However, when running the process for 1 million tweets we started to face memory utilization issues.

`