Извлечение описаний событий предопределённых типов из потока сообщений пользователей микроблогов

Ефим Пышнограев

Филиал МГУ в городе Севастополе

22 апреля 2014

Введение

Цель работы:

- исследовать существующие подходы по извлечению описаний событий из сообщений пользователей, выделить возникающие проблемы и рассмотреть возможные методы их решения,
- исследовать возможность применения тематических моделей для решения задачи выявления событий,
- разработать метод для извлечения описаний событий из сообщений пользователей сети Твиттер на основе иерархического процесса Дирихле,
- протестировать работу алгоритма на реальных данных.

Постановка задачи

Данные:

- множество документов (сообщений) $\Omega = \left\{ D_i \mid i \in \overline{1,n} \right\}$,
- документ — множество слов и временная метка t_i $D_i = \left\{ w_j \mid j \in \overline{1,l_i} \right\}$.

События:

- резкое увеличение частоты ключевых слов, затем спад до нормального уровня,
- должно быть вызвано реальным событием,
- не носит периодический характер.

Особенности сети Твиттер:

- короткие сообщения (до 140 символов),
- наличие шума и ошибок,
- большая плотность сообщений,
- “взрывной” характер событий.

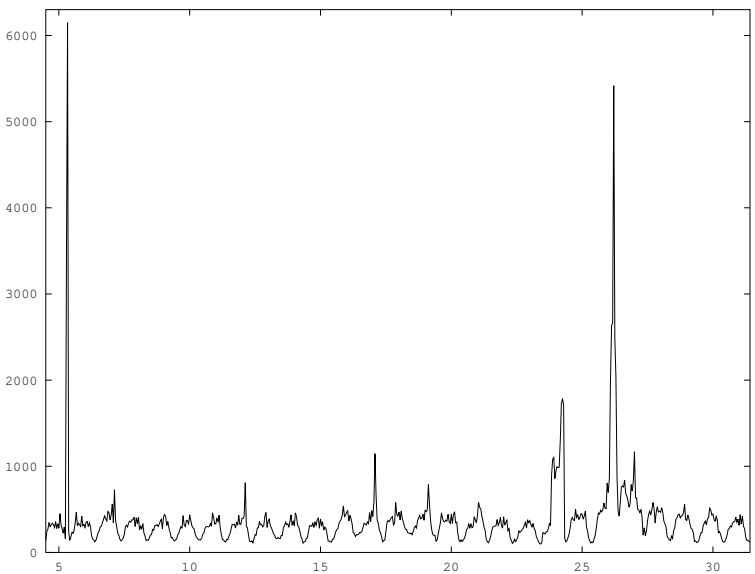
Схема алгоритма

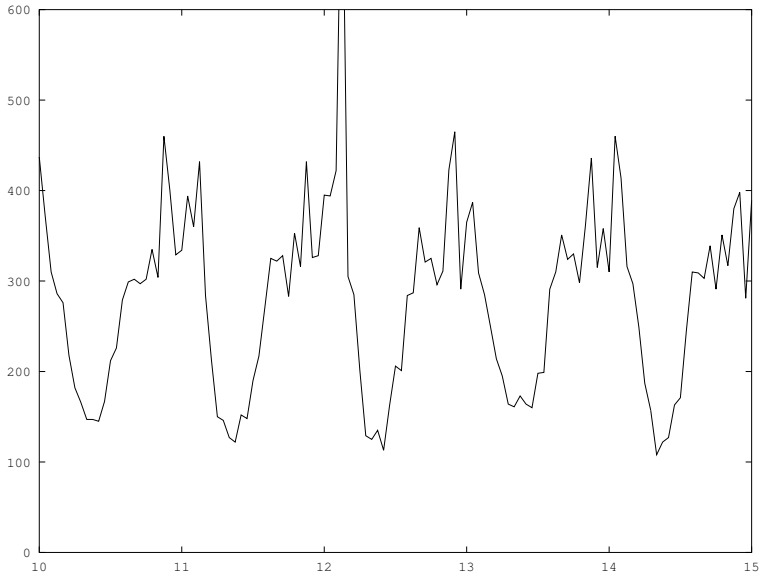
Составные части:

1. нахождение максимума в частотной функции, который будет соответствовать событию,
2. извлечение ключевых слов, характерных этому максимуму,
3. применение модели HDP с частичным обучением для того чтобы выделить все сообщения этой темы,
4. проверка насколько полученный результат соответствует новому событию.

Результаты

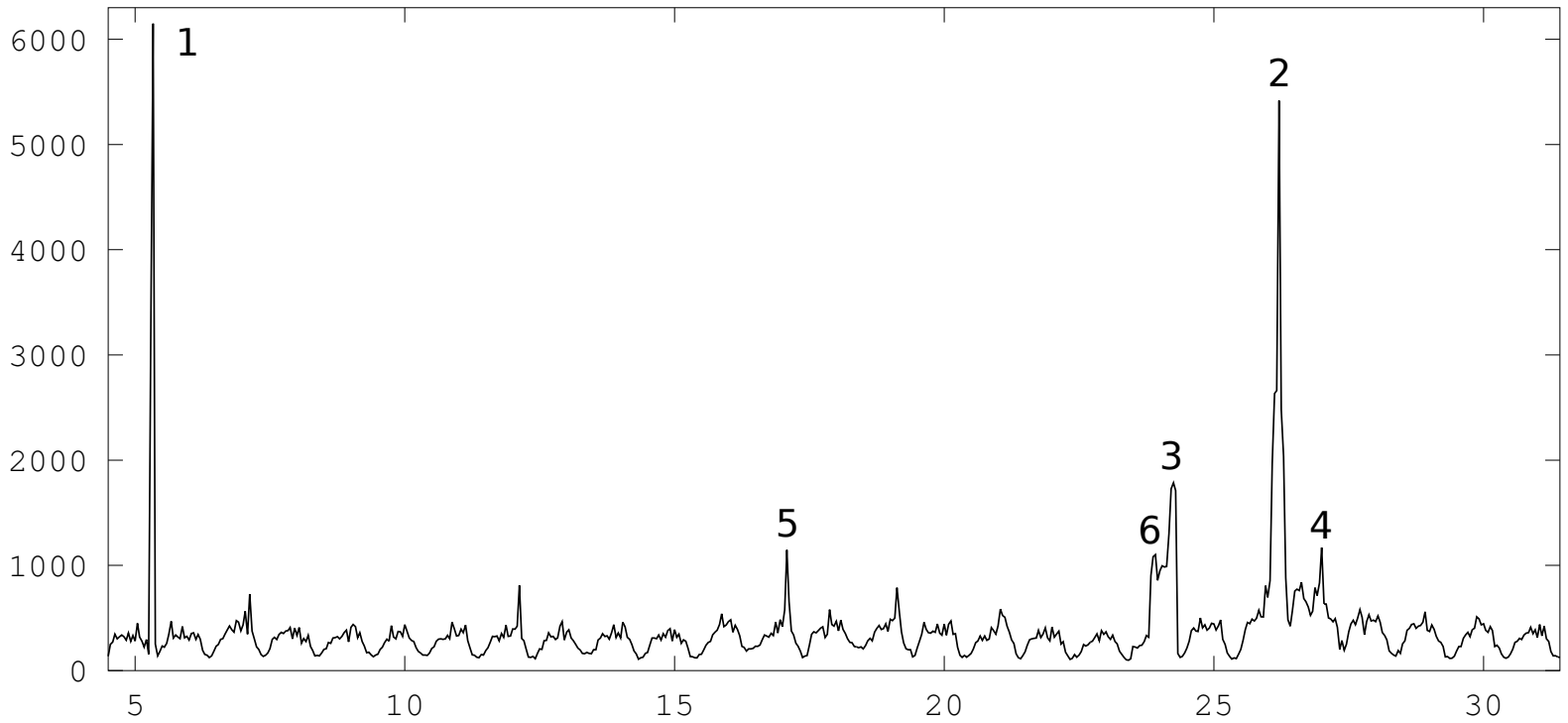
Данные: 240 тысяч сообщений с 4 июня по 1 июля 2013 года, содержащие хэштег #texas.





Результаты

Шаг 1: выделение экстремумов.



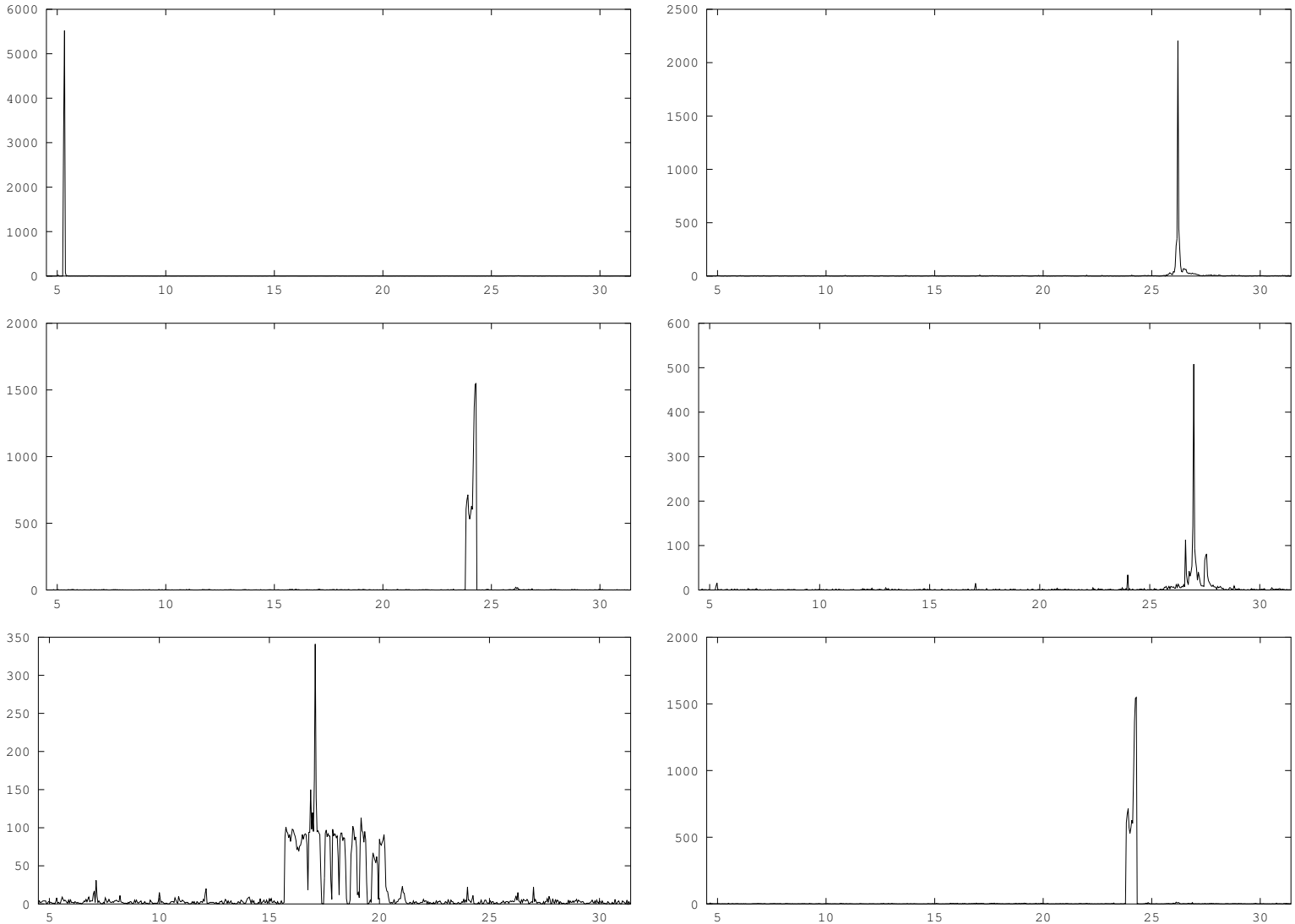
Результаты

Шаг 2: выделение ключевых слов.

#1	year, read, prison, dragon, amazon,
#2	sb5, standwithwendi, talk, made, histori
#3	houston, california, england, artist, rt
#4	execut, 500th, mccarthi, kimberli, 1976
#5	florida, z1, america, usa, bahrain
#6	houston, california, rt, watch, england

Результаты

Шаг 3, 4: применение модели HDP, фильтрация.



Результаты

Ключевые слова	Вызвавшее событие
year, read, prison, dragon, amazon, written, bestseller	Написанная в техасской тюрьме книга The Sword and the Dragon является бестселлером на Amazon уже более трех лет.
sb5, standwithwendi, abort, senat	Обсуждение сенатом закона по запрету абортов, который носит название Senate Bill 5. Wendi Davis — политик, которая боролась с принятием этого закона.
houston, california, england, artist, rt, watch, defjam	Хип хоп исполнитель из Хьюстон Devin the Dude объявил, что его восьмой студийный альбом будет называться One for the Road и будет выпущен в продажу в сентябре 2013 года. Альбом будет выпущен лейблом Def Jam Recordings.
execut, 500th, mccarthi, kimberli, 1976	В штате Техас приведен в исполнение 500-й по счету смертный приговор. Смертная казнь имеет место с 1976 года.

Заключение

В ходе выполения дипломной работы было достигнуто:

- изучены существующие методы извлечения описаний событий из коротких пользовательских сообщений,
- изучена возможность применения тематических моделей для это задачи,
- разработан и проанализирован алгоритм извлечения описаний событий из социальной сети Твиттер на основе иерархического процесса Дирихле,
- алгоритм был реализован, параметры подобраны эксперементальным путем,
- тестирование показало, что алгоритм успешно распознает события, имеющиеся во входных данных.