Извлечение описаний событий предопределённых типов из потока сообщений пользователей микроблогов

Ефим Пышнограев

Филиал МГУ в городе Севастополе

22 мая 2014

Задача извлечения событий

Входные данные:

- набор коротких сообщений пользователей,
- каждое сообщение состоит из текста, автора и временной метки.

Найти:

- набор событий,
- событие — ключевые слова и время возникновения.

Особенности сети Твиттер:

- небольшая длина сообщений,
- большое количество шума,
- высокая плотность сообщений,
- “взрывной” характер событий.

Цели дипломной работы

Дипломная работа ставит перед собой следующие цели:

- исследовать существующие подходы к задаче извлечения событий,
- исследовать возможность применения тематических моделей для решения задачи выявления событий,
- разработать метод для извлечения событий из сообщений пользователей сети Твиттер на основе иерархического процесса Дирихле,
- оценить качество работы алгоритма.

Схема алгоритма

Составные части:

1. нахождение максимумов частотной функции, в которых будет производиться поиск событий,

   для каждого максимума:

2. извлечение ключевых слов, характерных найденному максимуму,
3. применение модели HDP с частичным обучением для того чтобы выделить все сообщения, которые соответствуют найденным ключевым словам,
4. проверка, являются ли найденные ключевые слова и момент времени событием.

Используемые данные

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

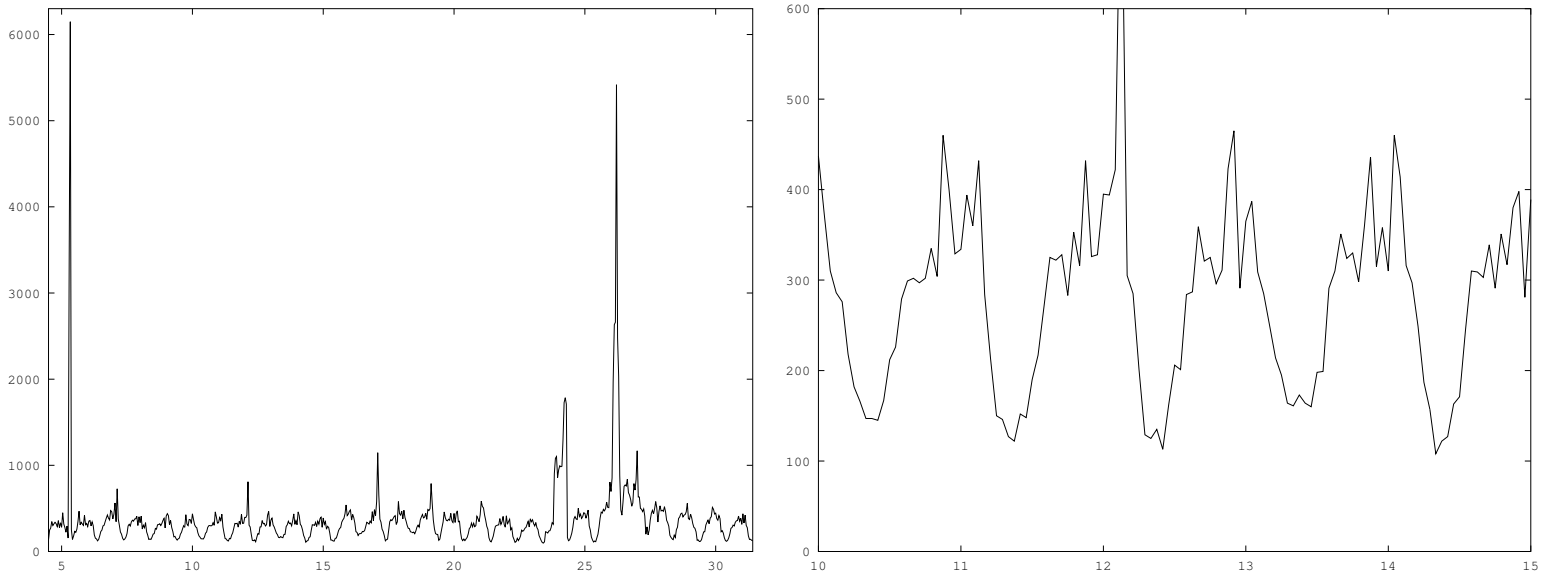
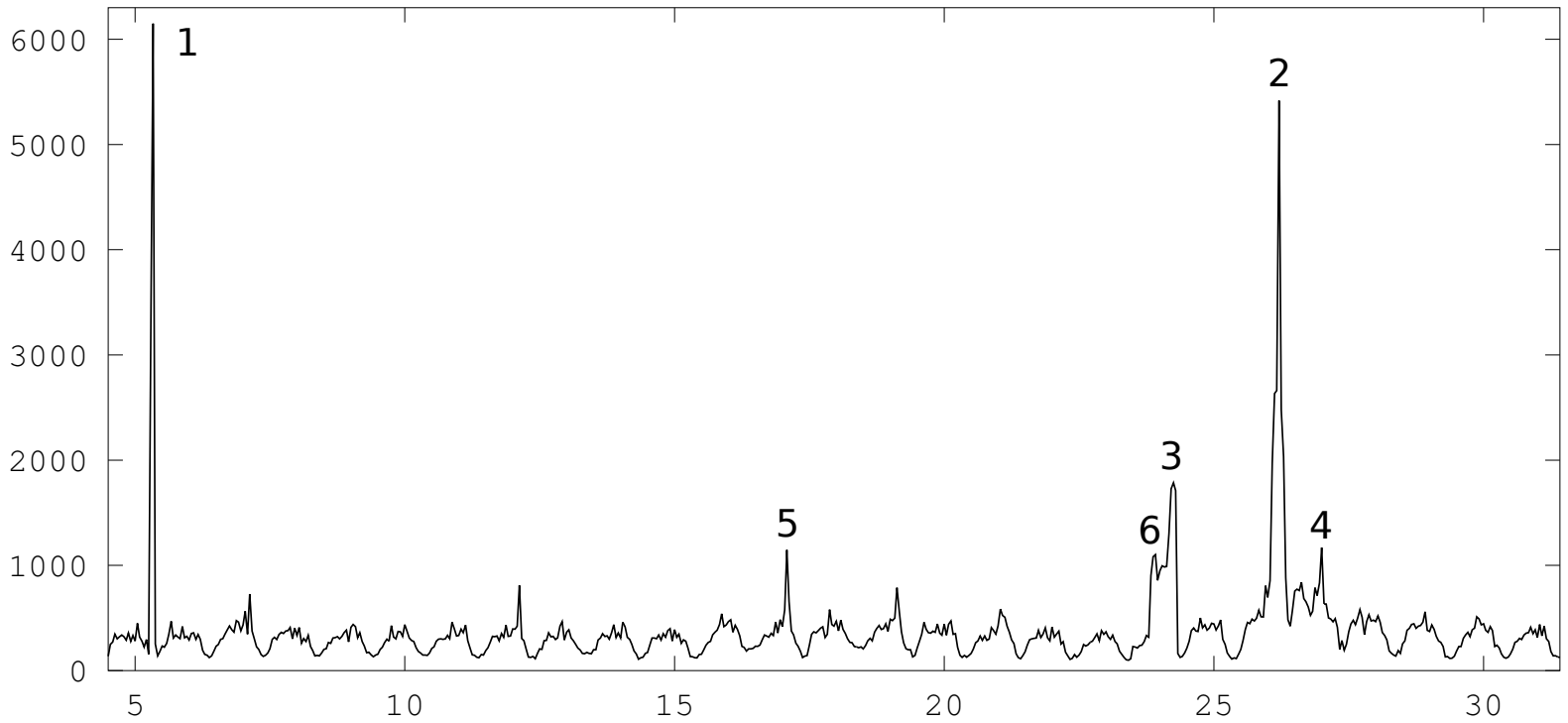


Рисунок. Количество сообщений в час.

Работа алгоритма

Шаг 1: выделение экстремумов.



Работа алгоритма

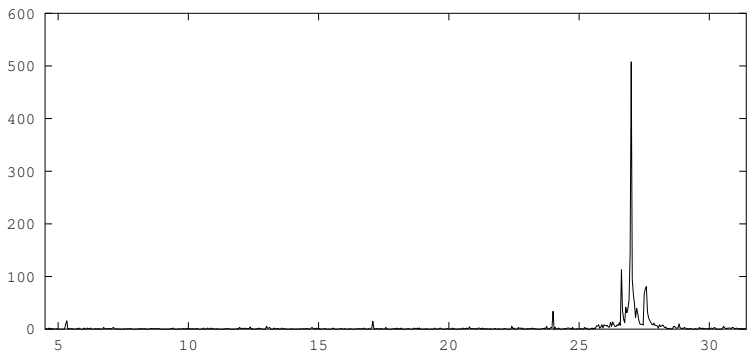
Шаг 2: выделение ключевых слов.

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

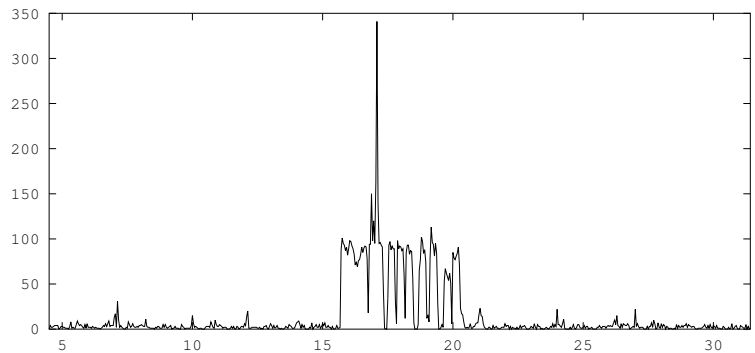
Работа алгоритма

Шаг 3, 4: применение модели HDP, фильтрация:

- на основе частоты сообщений, принадлежащих событию:



а) событие



б) не событие

- на основе отношения количества уникальных авторов к числу сообщений.

Работа алгоритма

Ключевые слова	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.
execut, 500th, mccarthi, kimberli, 1976	В штате Техас приведен в исполнение 500-й по счету смертный приговор. Смертная казнь имеет место с 1976 года.

Оценка качества

Для оценки качества было сделано предположение:

- если данные содержат событие, оно находится в максимуме функции частоты.

Сравнение с базовым алгоритмом New event detection:

Алгоритм	Точность	Полнота	F-мера
предложенный алгоритм	0.8	1.0	0.89
New event detection	0.01	0.8	0.02

Таблица. Оценка качества исходя из предположения

Почему у NED такая низкая точность на данных Твиттера:

- большая плотность новых слов из-за шума в данных,
- предположение, сделанное для оценки качества.

Детали оценки качества

Пусть:

- n событий было выдано алгоритмом,
- m событий находится во входных данных,
- k событий были верно найдены алгоритмом.

Критерии качества:

- точность $p = k/n$,
- полнота $r = k/m$,
- F-мера $F = 2pr/(p + r)$.

Способ оценки:

- точность: просмотреть n сообщений и выделить верные события,
- полнота:
  - с предположением об экстремумах: просмотреть небольшое количество максимумов функции частоты,
  - без предположения: просмотреть все сообщения для выделения событий.

Заключение

В ходе выполения дипломной работы было достигнуто:

- исследованы существующие подходы к задаче извлечения событий,
- исследованно и продемонстрировано применение тематических моделей для решения задачи,
- разработан метод для извлечения событий из сообщений пользователей сети Твиттер на основе иерархического процесса Дирихле,
- изучено качество работы предложенного алгоритма на имеющихся данных.