# Twitter Odee

## 摘要

过去的ee总是集中在新闻领域，而这篇文章则着眼于社交媒体。并且是第一篇提出开领域，无监督的论文。采用的是隐变量模型进行分类，输出是以日历的形式。

社交媒体的特点：

海量，冗余。

推文是有时间标注的。（这其实一直是ee中的一个大难题，但在社交媒体中相对容易）

杂乱无章（这也是我们自动提取，聚合，分类，起源动力，无结构到有结构）

由于推文的简洁性，推文的语法结构一般较简单，但通常缺乏足够上下文来将它分类。

由于推文的非正式性，推文一般语法比较自由，但大多数基于规则的工具都是建立在规整的语法上。

推文有大量无价值的事件。信息过载。

社交媒体话题广泛，事件类型难以界定。

输出：（named entity, event phrase, calendar date, event type ）

实体是agent

流程：先打pos标记，再提取命名实体与事件短语，解析时间，事件分类，根据冗余数量来确定是否重大。（根据每个命名实体共发的推文数量来衡量每个命名实体与日期之间的关联强度，以确定事件是否重大。）

NER采取在twitter上训练的NER分类器。。（有监督。）

那个事件提取也是有监督的。。（CRF）

时间提取采用的TempEx

采用隐变量模型。

Given a raw stream of tweets, our system extracts named entities in association with event phrases and unambiguous dates which are involved in significant events. (实体与事件短语与匿名时间相关)

First the tweets are POS tagged, （有监督词性标注）

then named entities and event phrases are extracted, temporal expressions resolved, and the extracted events are categorized into types.（NE,EE,TE,事件分类）

Finally we measure the strength of association between each named entity and date based on the number of tweets they co-occur in, in order to determine whether an event is significant.（事件检查。）

命名实体，pos标注器是他们自己之前训练的，

事件标注器是他们格外训练的。（全有监督。）（）

时间提取用的Tempex

事件分类是无监督的。。。

We adopt an approach based on latent variable models inspired by recent work on modeling

selectional preferences [47, 39, 22, 52, 48], and unsupervised information extraction [4, 55, 7].

每个事件指示短语均建模为混合的类型。

事件类型既和命名实体有关，也和时间有关。同一时期的事件类型更有可能一样。（odee是事件应该和更多的实体相关。）

为了解决 Twitter 的嘈杂风格，我们在嘈杂的文本[46，31，19]中跟踪了 NLP 的最新工作，为带有事件的推文注解，然后将其用作序列标签模型的训练数据，以识别数以百万计的事件提及消息。

由于推文的简洁性，有时是平凡的但高度冗余的性质，我们被激励着重于提取事件的聚合表示形式，这为诸如事件分类之类的任务提供了额外的上下文，并且还通过利用信息的冗余性来过滤掉了世俗的事件。我们建议将重要事件识别为那些提到的事件与对唯一日期的引用紧密相关，而不是在整个日历中均匀分布的日期。

将来工作的可能途径是提取更多更丰富的事件表示，同时保持域独立性。例如：将相关实体分组在一起，根据事件中实体的角色对实体进行分类，从而提取事件的基于帧的表示形式。

采取的G2对数似然来给事件排序。（准确度测试）

同类的还有塌方对数似然，与fisher准确度测试。

论文问题：

Linklda

G2测试

如何通过gibbs推断lda

变分推断lda