ReSEED - Social Event dEtection Dataset
ReadMe

Introduction

ReSEED is a dataset for research in the area of social event detection which contains user-contributed images and metadata under a Creative Commons license. It is suitable for clustering and classification tasks in that area.Both the scale and the complexity of the dataset make it more challenging and more representative of real-world problems.

Citation

If you use this dataset in your work, please cite the following paper:

Reuter, T., S. Papadopoulos, V. Mezaris, P. Cimiano. ReSEED: Social Event dEtection Dataset. In *Proceedings of the 5th Multimedia Systems Conference (MMSys '14)*, March 19-21 2014, Singapore. ACM, 2014.

Download

You can find the dataset to download from the following locations:

`http://greententacle.techfak.uni-bielefeld.de/reseed/`

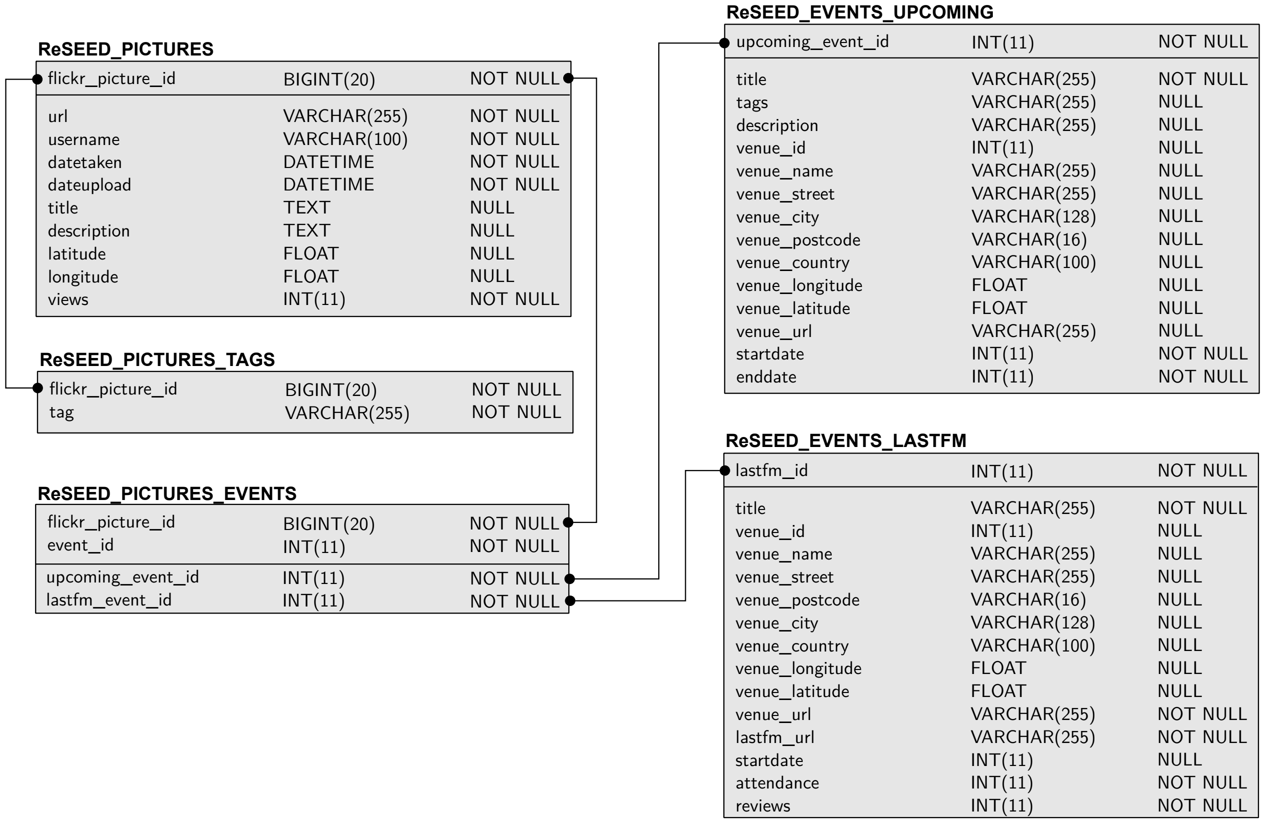
`http://dx.doi.org/10.4119/unibi/citec.2014.10`

Description

The dataset consists of pictures from Flickr together with their associated metadata. The pictures were downloaded using the Flickr API. We considered pictures with an upload time between January 2006 and December 2012, yielding a dataset of 437,370 pictures in total. The assignment of the pictures to events was derived from social media sites. The events in the dataset are heterogeneous, including sport events, protest marches, BBQs, debates, expositions, festivals or concerts. All of the pictures are licensed under a Creative Commons license allowing free distribution. As it is a real world dataset, there are some features like time-stamps and uploader information which are available for every picture, but there are also features (like geographic information) which are available only for a subset of the images.

We have split the dataset in two parts: development (train) and evaluation (test). The development dataset consists of 306,159 pictures, while the evaluation dataset consists of 131,211 pictures. The underlying ground truth enables researchers to use a supervised learning approach, i. e., tune a model to the events.

Data Model



ReSEED_PICTURES

flickr_picture_id	BIGINT(20)	NOT NULL
url	VARCHAR(255)	NOT NULL
username	VARCHAR(100)	NOT NULL
datetaken	DATETIME	NOT NULL
dateupload	DATETIME	NOT NULL
title	TEXT	NULL
description	TEXT	NULL
latitude	FLOAT	NULL
longitude	FLOAT	NULL
views	INT(11)	NOT NULL

ReSEED_PICTURES_TAGS

flickr_picture_id	BIGINT(20)	NOT NULL
tag	VARCHAR(255)	NOT NULL

ReSEED_PICTURES_EVENTS

flickr_picture_id	BIGINT(20)	NOT NULL
event_id	INT(11)	NOT NULL
upcoming_event_id	INT(11)	NOT NULL
lastfm_event_id	INT(11)	NOT NULL

ReSEED_EVENTS_UPCOMING

upcoming_event_id	INT(11)	NOT NULL
title	VARCHAR(255)	NOT NULL
tags	VARCHAR(255)	NULL
description	VARCHAR(255)	NULL
venue_id	INT(11)	NULL
venue_name	VARCHAR(255)	NULL
venue_street	VARCHAR(255)	NULL
venue_city	VARCHAR(128)	NULL
venue_postcode	VARCHAR(16)	NULL
venue_country	VARCHAR(100)	NULL
venue_longitude	FLOAT	NULL
venue_latitude	FLOAT	NULL
venue_url	VARCHAR(255)	NULL
startdate	INT(11)	NOT NULL
enddate	INT(11)	NOT NULL

ReSEED_EVENTS_LASTFM

lastfm_id	INT(11)	NOT NULL
title	VARCHAR(255)	NOT NULL
venue_id	INT(11)	NULL
venue_name	VARCHAR(255)	NULL
venue_street	VARCHAR(255)	NULL
venue_postcode	VARCHAR(16)	NULL
venue_city	VARCHAR(128)	NULL
venue_country	VARCHAR(100)	NULL
venue_longitude	FLOAT	NULL
venue_latitude	FLOAT	NULL
venue_url	VARCHAR(255)	NOT NULL
lastfm_url	VARCHAR(255)	NOT NULL
startdate	INT(11)	NULL
attendance	INT(11)	NOT NULL
reviews	INT(11)	NOT NULL

Figure 1: Database schema for dataset

The database schema für the dataset is illustrated in Figure 1. In particular we use the following features (together with their explanation):

`flickr_picture_id`	Unique ID of the image
`url`	URL of image (in Flickr)
`username`	Username of uploader in Flickr
`datetaken`	Date and time when image has been captured
`dateupload`	Date and time of upload to Flickr
`title`	Title of the image (given by uploader)
`description`	Detailled description of the image (given by uploader)
`latitude`	Latitude of image location
`longitude`	Longitude of image location
`tag`	One or more keywords (tags) that are assigned to the image
`cluster`	This denotes the ground truth events where the image belongs to

Please note the following for the CSV files:

- Files are stored as Unicode UTF-8.
- The first line of the CSV files denotes the name of the columns.
- All columns are tabular separated.
- New lines are encoded as `\n`
- The characters `&`, `'`, `"` are encoded as `&amp;`, `&apos;`, and `&quot;` respectivly.

Contact

Timo Reuter `treuter@cit-ec.uni-bielefeld.de`