Mining Flickr

Armin Eftekhari
Alejandro Weinstein

Division of Engineering
Colorado School of Mines

December 7, 2009

Key idea

Is there any relationship between the popularity of a photo and the camera used?

http://www.flickr.com/photos/acastellano/181730235/

- 2573 comments
- 69 tags
- 8539 users call this photo a favorite
- Taken with a Canon EOS 10D (about $1000).

Key idea

http://www.flickr.com/photos/luigistrano/354904253

- 17 comments
- 15 tags
- 7 users call this photo a favorite
- Taken with a Panasonic DMC-FX7 (about $400).

Building the data set

Build using the Flickr API, using the Python “Flickr API kit”. We get the following attributes:

Id: the unique id used by Flickr to identify a photo.

Views: the number of views of the photo.

Location: the location of the photo.

Comments: the number of comments of the photo.

Tags: the number of tags of the photo.

Favorites: the number of Flickr users that call the photo a favorite

Make: the maker of the camera used to take the photo.

Model: the model of the camera used to take the photo.

Building the data set

Sampling strategy:

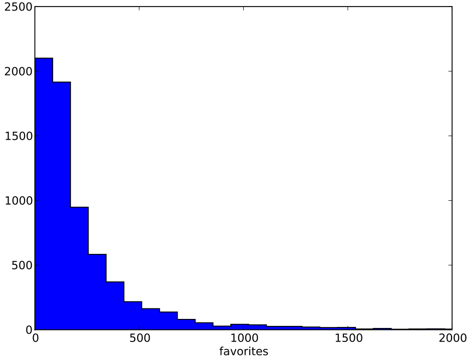
- Get 100 photos for each of the “All time most popular tags”.
- The photos must be at least one year old.
- Four steps process:
  1. Get the 100 ids for each of the poplular tags. We got 14400 records.
  2. Eliminate the duplicate ids. We reduced the 14400 records to 7342.
  3. Get the data for each photo id.
  4. Write the data as a CSV file.

Building the dataset took several hours.

Dataset cleaning

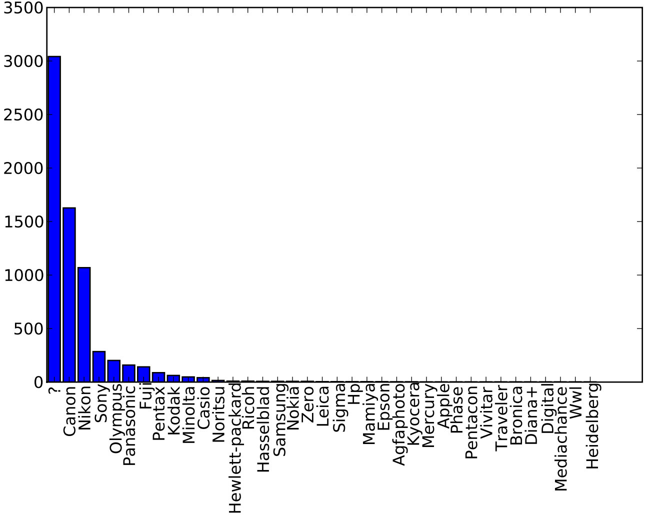
- The *Make* attribute was inconsistent.
- Slightly different names for a given maker.
- For instance, “Olympus” cameras were labeled as “Olympus Imaging Corp.”, “Olympus optical Co. Ltd” or “Olympus corporation”.
- Another Python script solved this inconsistency.
- We reduced the number of unique manufacturers from 62 to 36.

Summary statistics: Favorites

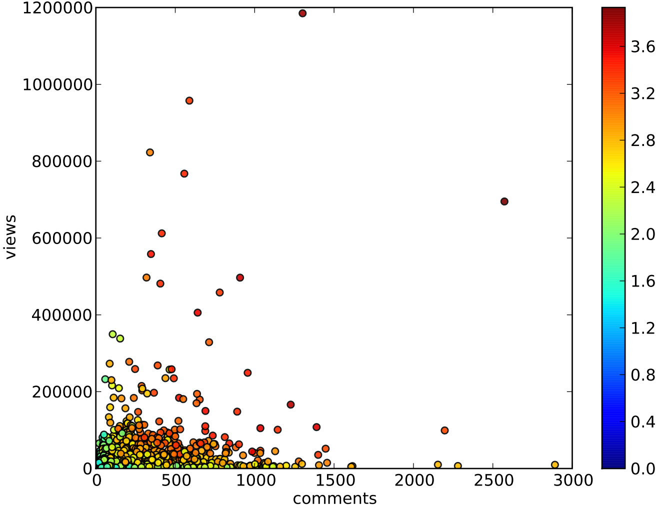


favorites	
Min	0
Max	8539
Mean	243.0
Std dev	348.2

Summary statistics: Makers



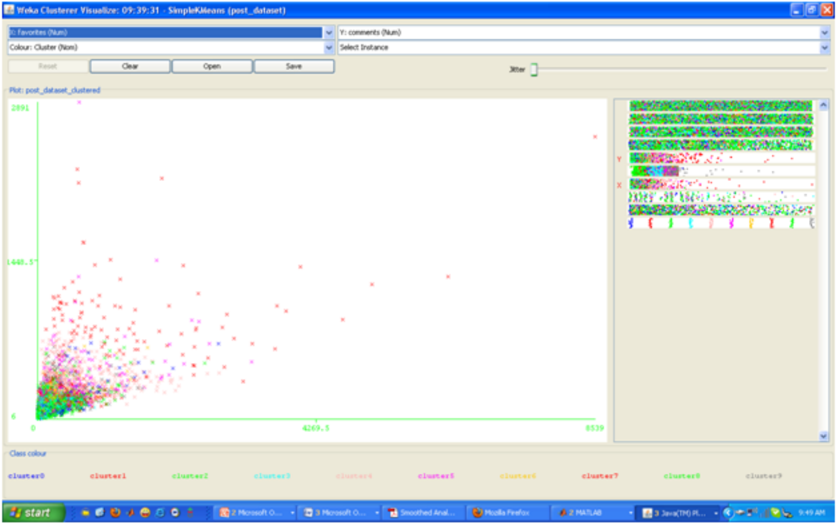
Scatter plot



Clustering

```
\includegraphics[scale=0.5]{scatter.pdf}
```

Clustering



Weka Clusterer Visualize: 09:39:31 - SimpleKMeans (post_dataset)

X: favorites (Num)

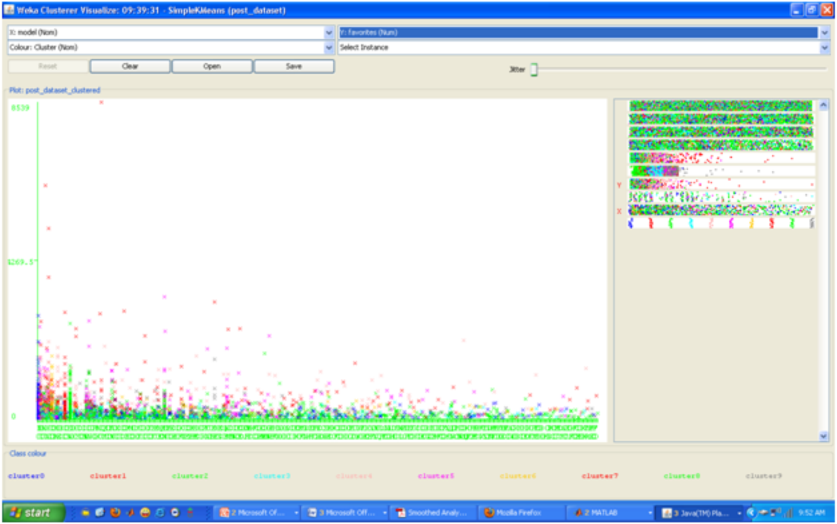
Y: comments (Num)

Colour: Cluster (Nom)

Select Instance

Plot: post_dataset_clustered

Clustering



Weka Clusterer Visualize: 09:39:31 - SimpleKMeans (post_dataset)

X: model (Nom)

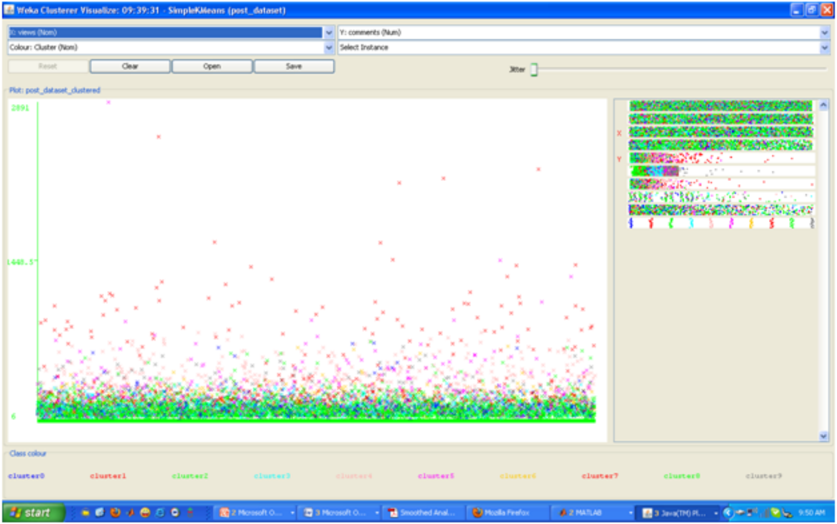
Y: favorites (Num)

Colour: Cluster (Nom)

Select Instance

Plot: post_dataset_clustered

Clustering



Weka Clusterer Visualize: 09:39:31 - SimpleKMeans (post_dataset)

X: views (Nom)

Y: comments (Num)

Colour: Cluster (Nom)

Select Instance

Plot: post_dataset_clustered