Flickr Dataset

Ryan McDyer - 13431038
Jordan Healy - 13379226

Group 37

Dataset & Objective

- Flickr is Yahoo’s image-hosting website
- Data for 109,593 images scraped from Flickr using the Flickr API
    - Images taken from the center of Dublin within 3km radius
- Flickr has a publicly accessible API, so we wrote a bash script using `wget` to scrape it
- The goal: predict the views of an image based on its location (latitude and longitude), date & time taken
- Training dataset of 54,593 tuples

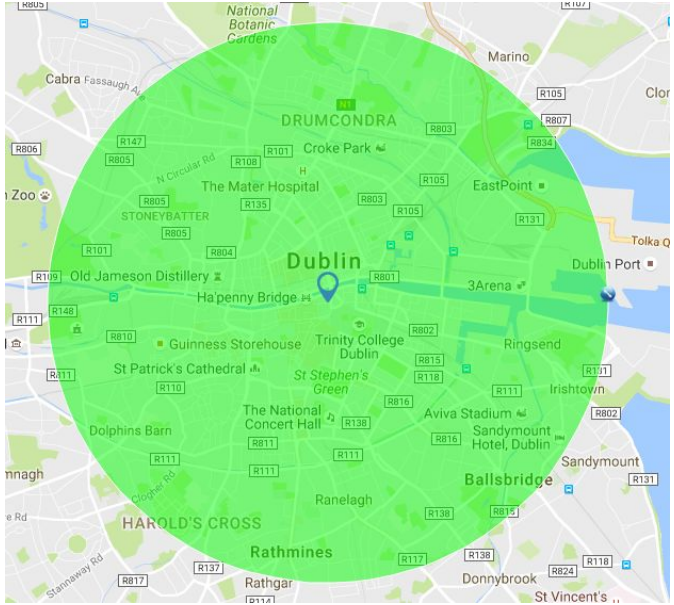


Figure 1: Radius of dataset

id	owner	title	datetaken	datetakengranularity	datetakenunknown	views
9789136583	12350145@N04	Dublinia	2013-09-08 14:09:46	0	0	67062
15620512370	33577523@N08	T. Cranfield CDV	2016-01-01 00:00:00	8	0	1069
17060618259	95093559@N06	215 at Heuston, 22/4/15	2015-04-22 18:28:05	0	0	159
30005537656	59126743@N04	A Walk on the Southside of the Liffey. Dublin Fair City.	2016-09-28 12:26:32	0	0	9

latitude	longitude	tags	accuracy	place_id	woeid	context
53.343467	-6.27124	ireland dublin dublinia	16	fjmHzRZUVLpKRgI	560743	0
53.343841	-6.259179	officer lieutenant hussar edmondhegankennard	15	fjmHzRZUVLpKRgI	560743	0
53.346349	-6.295058	dublin irish train gm rail railway trains railways	15	v29QVFxUVLszepE	561338	0
53.334011	-6.263509		16	fjmHzRZUVLpKRgI	560743	0

Creating Classifications

- No classification in our dataset
  - We must create it
- Views attribute are numbers ≥ 0
- Create 4 classes based on Five Number Summary
  - Low = 0 -> 17
  - Medium = 18 -> 63
  - High = 64 -> 229
  - Very High = 230 -> 98,001
- Looked at standard number ranges
- Add this attribute, views_classification, to dataset

Min	0
Q1	17
Median	63
Q3	229
Max	98001

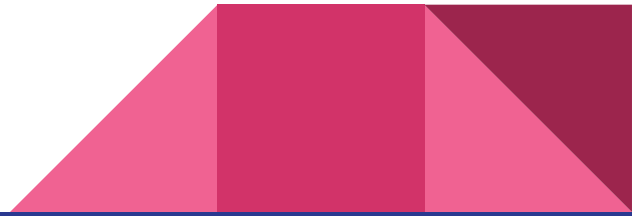
Figure 2: Five Number Summary

Missing Values

- Some tuples in our dataset have missing attribute values
  - Tags
  - Titles
- We will leave these blank as this isn’t an error in our data, the user has intentionally left no title or given no tags
  - Use this for predicting views (use number of unique tags or length of title)
  - From this we will create a new attribute in each tuple called “numberoftags”and “titlelength”

Remove Unnecessary Attributes

- There are a few attributes given to us by Flickr that we deem useless in our prediction
  - Is_family
  - Is_friend
  - is_public
- Although Flickr does give us some metadata. For instance;
  - Context - whether the photo was taken indoor, outdoor, or unknown
  - Accuracy - how accurate the latt. & long. values are (0 low -> 16 high)
  - Granularity - accuracy of date (whether format is missing time or day, for example)



Discrepancy Detection

- Found “Field Overloading” in datetaken attribute
  - When two or more concepts are being used in a single data field
    - Create two separate attributes
- Metadata in tags attribute.
  - Make new attributes from uploaded:by=instagram, uploaded:by=flickrmobile, and foursquare:venue=$hexstring
  - Then remove these from the tags field so they aren’t accounted for twice.

id	owner	title	title_length	datetaken	timetaken	views	views _classification	context
8607568914	12504159@N06	SKBFC Academy Cup Final Celtic v West Bromwich Albion	53	2013-03-31	13:13:11	1236	Very High	0
25381889510	84221313@N04	St. Patrick's Cathedral, Dublin, Ireland	40	2015-11-07	13:03:25	986	Very High	0
9445244328	47545877@N06	AFL Europe - Aussie Rules Football	34	2013-08-03	18:09:36	103	High	0

place_id	woeid	tags	count_tags	latitude	longitude	flickr _app	instagram _app	foursquare _venue
fjmHzRZUVL pKRgI	560743	ireland dublin dublinia ...	6	53.343467	-6.27124	TRUE	FALSE	TRUE
fjmHzRZUVL pKRgI	560743	officer lieutenant hussar ...	15	53.343841	-6.259179	FALSE	TRUE	FLASE
v29QVFxUVL szepE	561338	dublin irish train gm rail railway trains railways ...	18	53.346349	-6.295058	FALSE	FLASE	FLASE

Conclusion

- We learned that the data you receive from a source isn't always suitable to be analysed immediately.
  - We needed to remove useless tuples (where datetakengranularity is not 0 or accuracy is 0) and remove useless attributes (13 in total).
  - But we did need to add some attributes too (we added 5).
- Clean enough, not to our liking, must add some attributes to make it better for analysis
- Can transform data to make it easier to make prediction
  - Not changing values, just the way it looks