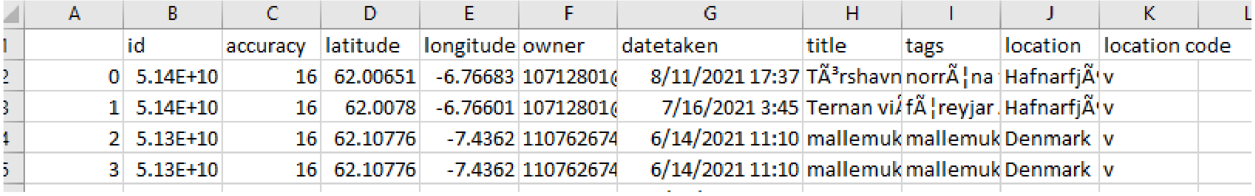
Flickr BOG class

Bag of words is a representation for text analysis that counts each instance of a word. This program goes through all the tags for each photo and counts how many times they are encountered. It has two methods for data processing, get_tag_info and tags_to_BOG.

get_tag_info

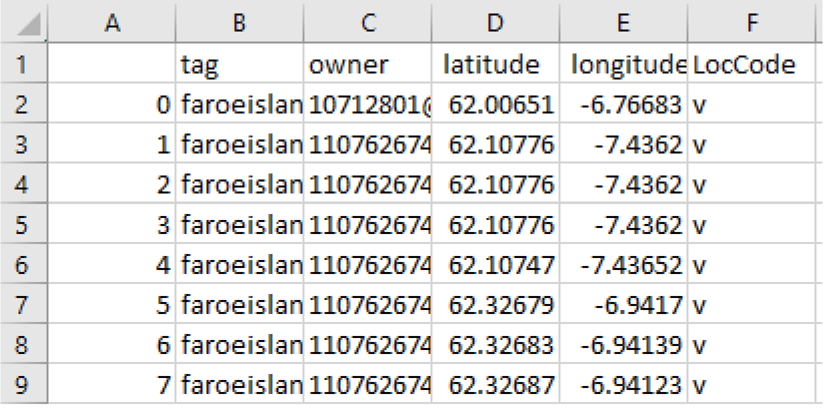
Takes in data frame and a 'tag' and outputs a row for the location of each tag. Only one tag can be pulled at time. All text is converted to lower case to avoid multiple cases that can occur with different capitalizations

Figure 1: Sample input csv.



	A	B	C	D	E	F	G	H	I	J	K	L
1		id	accuracy	latitude	longitude	owner	datetaken	title	tags	location	location code	
2	0	5.14E+10	16	62.00651	-6.76683	10712801(	8/11/2021 17:37	TÃ³rshavn	norrÃ¦na	HafnarfjÃ	v	
3	1	5.14E+10	16	62.0078	-6.76601	10712801(	7/16/2021 3:45	Ternan viÃ	fÃ¦reyjar	HafnarfjÃ	v	
4	2	5.13E+10	16	62.10776	-7.4362	110762674	6/14/2021 11:10	mallemuk	mallemuk	Denmark	v	
5	3	5.13E+10	16	62.10776	-7.4362	110762674	6/14/2021 11:10	mallemuk	mallemuk	Denmark	v	

Figure 2: Sample output csv for the tag "faroeislands". Outputs a row for each instance of a tag. This output is for mapping tags in a geographic information system. This output contains 87 tags, see figure 4.

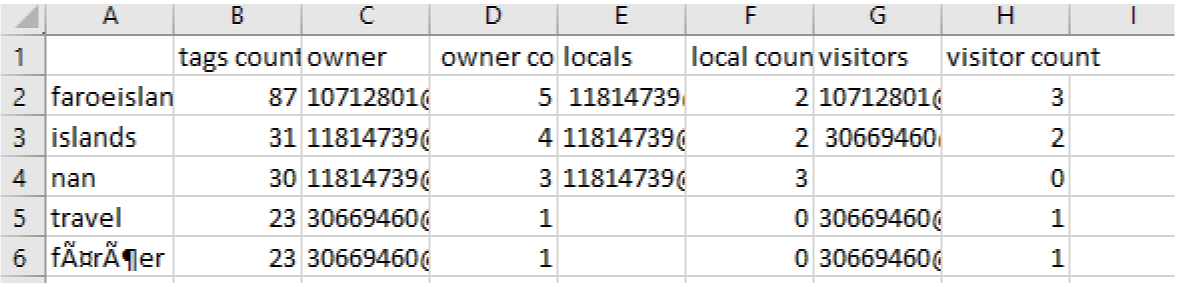


	A	B	C	D	E	F
1		tag	owner	latitude	longitude	LocCode
2	0	faroeislan	10712801(	62.00651	-6.76683	v
3	1	faroeislan	110762674	62.10776	-7.4362	v
4	2	faroeislan	110762674	62.10776	-7.4362	v
5	3	faroeislan	110762674	62.10776	-7.4362	v
6	4	faroeislan	110762674	62.10747	-7.43652	v
7	5	faroeislan	110762674	62.32679	-6.9417	v
8	6	faroeislan	110762674	62.32683	-6.94139	v
9	7	faroeislan	110762674	62.32687	-6.94123	v

tags_to_BOG

Takes in a data frame and Outputs a Bag of words representation for tags and some additional information. Information provided includes tag, tag count, list of owners and a break down of locals and visitors. Nan = not a number and represents photos that have no tags. The sum of the local count and the visitor count should equal the owner count. The output is not sorted in any way, the results pictured below were sorted in excel.

Figure 4: Sample output csv



	A	B	C	D	E	F	G	H	I
1		tags count	owner	owner co	locals	local coun	visitors	visitor count	
2	faroeislan	87	10712801@	5	11814739	2	10712801@	3	
3	islands	31	11814739@	4	11814739@	2	30669460	2	
4	nan	30	11814739@	3	11814739@	3		0	
5	travel	23	30669460@	1		0	30669460@	1	
6	fÃ¤rÃ¶er	23	30669460@	1		0	30669460@	1	