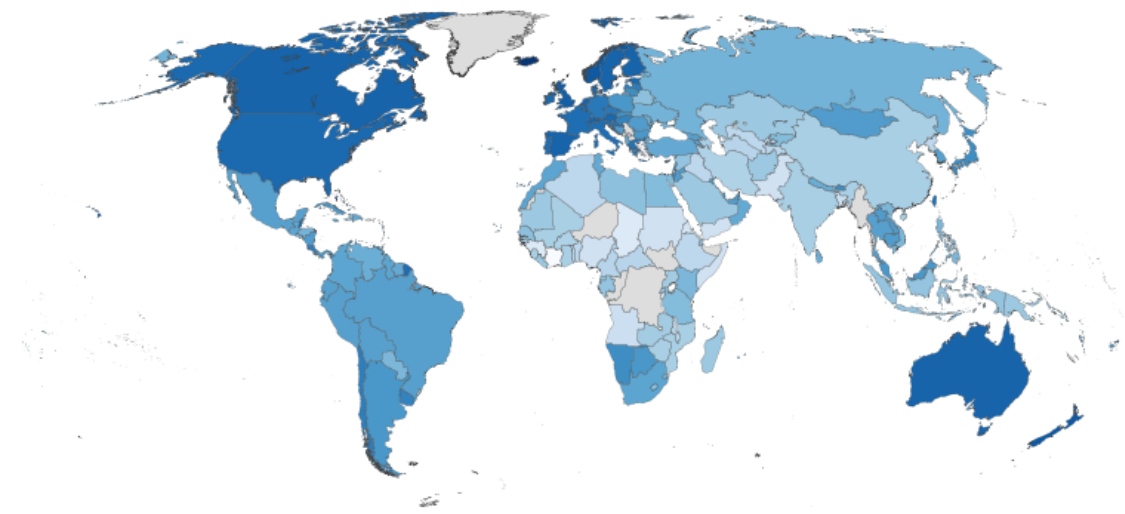# Geography-Based Summary

Overview Statistics

(M5) Geographic representation of dataset scaled by country population, colored on a logarithmic scale.
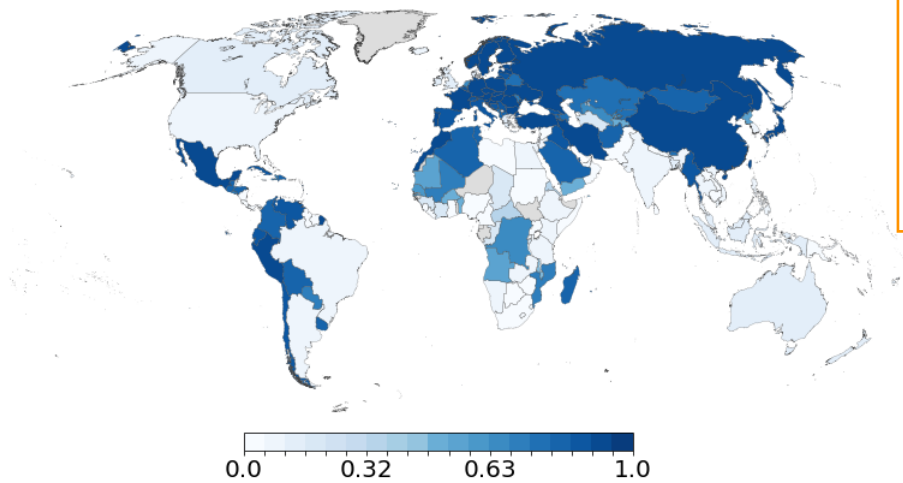

(M10) Most popular languages:
English: 85130   85.1862%
Italian: 2341   2.3425%
French: 2254   2.2555%

(M10) Map representing the fraction of tags in a country that are not labeled in a local language.

Percentage of tags in non-local language

Comparing this map to the one above, we can see that even when underrepresented countries are represented, it might be by visitors rather than locals.

Sample Interesting Findings

(M6) Overrepresentations of tags by country (tag in country vs tag in rest of the countries):
0.0: East+Timor-reading (7/75 vs 508/1220353)

0.0: Namibia-desert (39/631 vs 1943/1219797)
0.0: Chile-skate (57/5647 vs 411/1214781)
0.0: Egypt-mosque (43/2733 vs 523/1217695)

Underrepresentations of tags by country (tag in country vs tag in rest of the countries):
0.0: United+States-travel (3143/376080 vs 14772/844348)
0.0: United+States-turkey (77/376080 vs 2921/844348)
0.0: United+States-wales (5/376080 vs 2077/844348)
0.0: United+Kingdom-travel (910/144959 vs 17005/1075469)

(M6) To discern if there is an appearance difference in how certain subregions represent a tag, we extract scene-level features from each image, and fit a linear SVM to distinguish between the tag in the subregion and out of the subregion.
An example of the most linearly separable tag: southwest has an accuracy of 0.762 when classifying in Northern America vs outside Northern America.

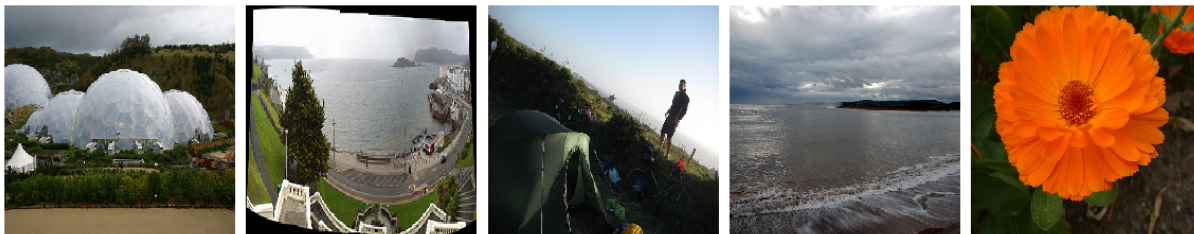In: Correct

In: Incorrect

The tool automatically surfaces the most linearly separable tag, but it is up to the user to determine if the difference is actually meaningful. In this case, it is not actually clear that "southwest" is being portrayed in a significantly different way.

Out: Incorrect

No images in this category

Out: Correct

Some of the other metrics in the notebook

- (M5) Image breakdown by country and subregion
- (M5) Dataset representation map

- (M6) Over/under representations of tags by subregion
- (M10) Visual representation of what languages are represented
- (M10) What languages each country's tags are in