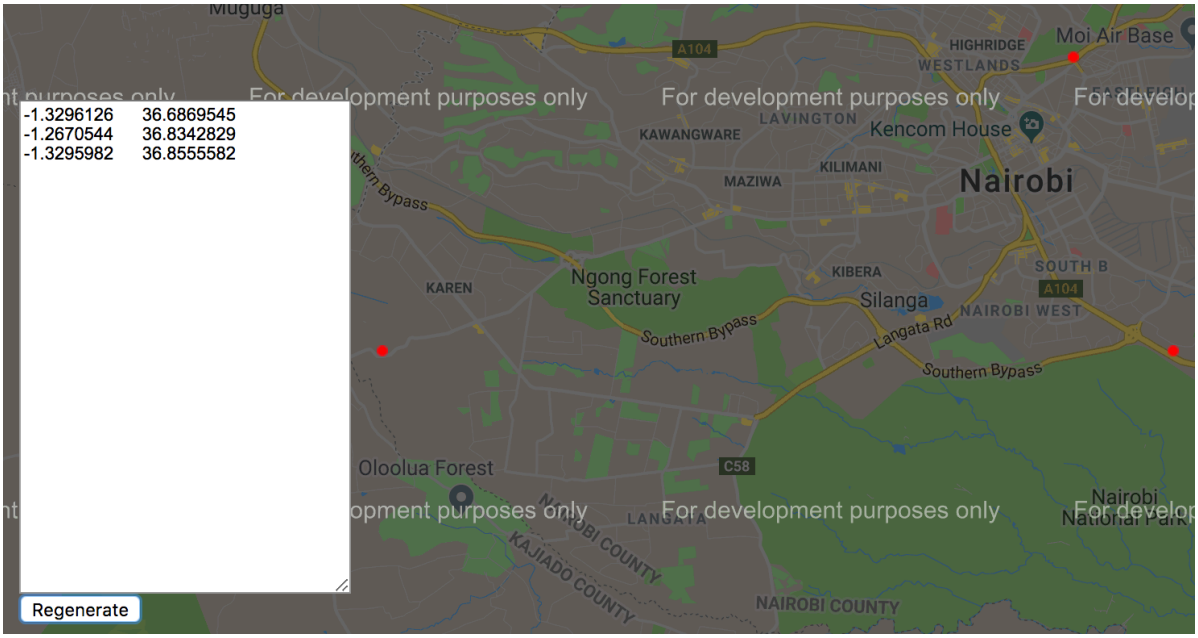
Tweet Crash Clustering Protocol

Goal: Often multiple people tweet about the same crash. In this exercise, we're interested in identifying which tweets refer to which crash. This will help us identify things such as: (1) how many minutes/hours apart do people tweet about the same crash? (2) if people refer to the same crash but refer to different landmarks, how far apart are they?

How to determine crash clusters: Read the tweet, look at the time it occurred, and check the landmark used & location. If two tweets were tweeted at similar times and are close together, code them as the same crash. Look for clues in the tweet: for example, if one tweets says a "car and a motorcycle" got in a crash, and another says a "matatu and truck" got in a crash, then they're referring to different crashes. We realize this will be a bit subjective; just try to use your best judgement.

How to easily determine if two crashes are close together:

- **Suggestion 1:** Plug coordinates individually into Google Maps
  Suggestion 2: This website (http://www.hamstermap.com/quickmap.php) allows you to copy and paste a bunch of coordinates into a map and have them all immediately appear. See the screenshot below: we copy & pasted three pairs of coordinates directly into the window, clicked "regenerate", and can clearly see the points are far apart and therefore do no refer to the same crash.



How to Code Crashes: Each tweet has a unique id. The crash cluster id should be the lowest tweet id used among all crashes in a cluster. For example, if tweets with ids of 3, 6 10, 16 all refer to the same crash, each should be given a crash cluster id of 3.

If only one tweet refers to a crash, the crash cluster id should just be the tweet id. For example, if tweet with id 4 is the only tweet that refers to a crash, the crash cluster id should just be 4.

Task: In the excel file, fill in the crash_cluster_id column. If you have any comments (e.g., are uncertain), feel free to write something in the comments column but this is not required.

Example Coding

id	time	tweet	landmark	latitude	longitude	Crash cluster id
15	2015-12-09 17:39:40 UTC	15:39 An accident on Mbs Rd toward town opp Simba Colt small car overturned. Rwaken bus also on sight. Ppl slowing to check via @ckageha	Simba Colt	-1.3295982	36.8555582	15
16	2015-12-09 18:15:20 UTC	16:14 Accident on Msa rd inbound opp Panari hotel #kot https://t.co/B2z5pAKUPJ via @mistamodesti	Panari	-1.3295982	36.8555582	15
17	2015-12-09 18:24:30 UTC	16:23 #NAIROBI Accident on Mombasa rd inbound opp Panari hotel , saloon gymnasium via @MombasaCGW	Panari	-1.3295982	36.8555582	15
18	2015-12-09 18:33:20 UTC	16:32 panari crash @GulfKavirondo https://t.co/ltH7sJgNSi via @tycoontraffic	Panari	-1.3295982	36.8555582	15
19	2015-12-09 18:37:40 UTC	16:37 accident in mombasa road next to panari hotel https://t.co/t234ufzjz7 via @OyugiEryc	Panari	-1.3295982	36.8555582	15
20	2015-12-09 19:09:30 UTC	17:08 Accident on mombasa road near panari is nearly cleared https://t.co/y47t6EdcFu via @kenyanspider	Panari	-1.3295982	36.8555582	15
21	2015-12-09 19:48:30 UTC	17:47 Accident opposite Panari Msa Rd via @Uhurunomics	Panari	-1.3295982	36.8555582	15
22	2015-12-09 22:14:00 UTC	20:13 heavy traffic in Karen towards Ngong. Accident at KCB/ECC area via @_DavidOdhiambo	KCB Leadership centre	-1.3296126	36.6869545	22
23	2015-12-09 22:41:00 UTC	20:40 accident near KCB training centre ngong road. Mat n govt saloon involved both seem to have done a few flips via @edugooner	KCB Leadership centre	-1.3296126	36.6869545	22
24	2015-12-10 07:31:41 UTC	05:31 fatal accident at pangani interjection.no policemen so far. via @domimoas	Pangani roundabout	-1.2670544	36.8342829	24
25	2015-12-10 07:50:01 UTC	05:49 @ntsa_kenya Accident at pangani rounderbout https://t.co/aHUpOkW3k3 via @ABISOMOSH	Pangani roundabout	-1.2670544	36.8342829	24
26	2015-12-10 08:50:01 UTC	Bad accident near garden city	Garden city	-1.231286	36.877484	26
27	2015-12-10 08:52:01 UTC	Really bad crash near yaya center	Yaya center	-1.292358	36.788111	27