

Interactive visualisation of twitter data

1. **Problem statement:** Analysis of tweets relating to the recent stormy weather (*Storm Emma* and “*the Beast from the East*”)

2. Motivation:

Ireland, UK and most of the northern Europe witnessed one of the most deadly snowstorms in the recent decades. People, mostly confined to their homes, took to twitter to express their feelings, share the hardships faced or just to keep themselves updated with the developments. Civic authorities and the meteorological department made heavy use of twitter to spread awareness among the citizens and keep them abreast of latest weather conditions.

With this background, there are loads of tweets in period from last week of February to first week of March, 2018 that relate to the storm. These tweets are a treasure trove of information that reveal various *issues faced* by the people in the time of crisis. An analysis of these tweets is therefore warranted as we can better *prepare ourselves* should the nature decide to unleash its fury in the future.

3. Methodology:

3.1. **Identifying data sources:** About 13000 tweets were sampled out based on the following criteria:

- Mention keywords/ hashtags like “the beast from the east”, “Storm Emma”, or “Sneachta” or
- were simply created in between 26th Feb, 2018 to 04th Mar, 2018 in the Ireland-UK region.
- Tweets by official Ireland and UK meteorological offices
- **GeoJSON file:** A file that specifies physical boundaries of various regions in UK was procured from UK government open data portal
- **Map tile-layer:** A basic tile layer of world map obtained from <http://openstreetmap.se/>

3.2. **Procuring data:** Twitter API and twitter toolbox were the tools used to query the Twitter data with respect to keywords, time duration and/ or location. All the tweets were dumped to a JSON file (*around 67 MB*).

3.3. Data pre-processing:

Cleaning and enrichment of JSON data was undertaken in Python language as follows:

- **Geocoding:** Most of the tweets lacked any geospatial information, and some of them mentioned ‘place’ from where they were created. For all such tweets where location name was available, latitude/ longitude values were fetched using geocoding. Typical response available was like:

```
In [213]: 1 geolocator = Nominatim()
          2 import time
          3 time.sleep(2)
          4 location_geopy = geolocator.geocode("Limerick")
          5 location_geopy

executed in 2.23s, finished 23:44:56 2018-03-14

Out[213]: Location(Limerick, County Limerick, Munster, V94 A9NX, Ireland, (52.661252, -8.6301239, 0.0))
```

Interactive visualisation of twitter data

- **Preparing GeoJSON file:** As geoJSON format is readily integrated with Leaflet.js maps, a file with required structured was created using Python script. The formatted geoJSON looked like:

```
In [211]: 1 listTweets

Out[211]: [{'geometry': {'coordinates': [-4.2435817, 55.856656], 'type': 'Point'},
  'properties': {'created_at': 'Fri Mar 02 12:24:24 +0000 2018',
    'favorite_count': 1255,
    'location': 'Glasgow',
    'text': '#snowday3 Ooh & some milk https://t.co/bp43THBRaj'},
  'type': 'Feature'},
  {'geometry': {'coordinates': [-6.40734708789517, 53.27367045],
    'type': 'Point'},
  'properties': {'created_at': 'Sat Mar 03 07:36:58 +0000 2018',
    'favorite_count': 135,
    'location': 'South Dublin',
    'text': 'If you're going to share or like or comment about one tweet on #Tallaght'}
```

- **Data for tracking weather warnings:**
UK Meteorological office regularly issues yellow, amber or red weather warnings in case of bad weather conditions. Interestingly, it posts such warnings through its *automated* Twitter handle [@metofficeUK](#) . Daily warnings for different regions were extracted from its tweets which essentially follow a similar structure (as they are automated tweets) Eg.



These warnings then were mapped to respective regions in the geoJSON file for English regions.

- **Popular hashtags:** Number of hashtags in individual tweets were counted in a dictionary (Python), ordered in a descending fashion of occurrence and dumped to a JSON file to be used in a high-chart word cloud representation

Interactive visualisation of twitter data

```
In [205]: 1 listWordCloud = []
          2 for items in topHashtags:
          3     cloud = {}
          4     cloud['name'] = items[0]
          5     cloud['weight'] = items[1]
          6     listWordCloud.append(cloud)
          7 with open('top50Hashtags.json', 'w') as outfile:
          8     json.dump(listWordCloud, outfile)

executed in 31ms, finished 21:25:16 2018-03-14

In [214]: 1 listWordCloud

executed in 16ms, finished 00:03:51 2018-03-15

Out[214]: [{'name': 'snowday3', 'weight': 6612},
            {'name': 'StormEmma', 'weight': 3677},
            {'name': 'BeastFromTheEast', 'weight': 924},
```

- **Timeseries data for food panic:** It was hypothesized that people who panicked about unavailability of popular food products would resort to twitter to talk about it. So, a daily count of tweets mentioning popular food items on each day from 26th Feb to 3rd Mar, 2018 was obtained and added to a dictionary in order to create a JSON data file in an appropriate format. Eg. Checking for bread count in tweets.

```
In [220]: 1 len([(t['text'], t['favorite_count']) for t in locations if 'bread' in t['text'].lower() and 'Feb 28' in t['created_at']])

executed in 31ms, finished 00:15:22 2018-03-15

Out[220]: 153
```

Note: Number of word combinations were tried before the count was obtained. Eg. Bread was tested with related keywords like '#Bread...', '#Brennan', etc. Finally all these datasets were exported in Highcharts and Leaflet friendly JSON formats.

4. Analysis and visualization:

4.1. Plotting locations of popular (most favoured) tweets: Ten tweets from Ireland and UK each, that have been highly liked by other users, have been plotted on the map according their location.

Observations: Most favourite tweets have appeared from urban regions like Glasgow, London in UK and Dublin, Cork in Ireland.

4.2. Visualising daily updates in weather warnings : As mentioned earlier, an automated twitter account was used to find the weather warnings issued during the course of a week from 27th Feb to 4th March.

4.3. Popular hashtags: Since, hashtags related to storm Emma or other snowy weather phenomenon were disproportionately high in data, a filter for taking a subset of top 50 ranked hashtags can be employed using the provided sliders. Hovering upon words in the cloud can reveal their number of occurrences in the tweets corpus.

4.4. Shortage of food items: A line plot for some of the popular food products has been plotted. Observations: Milk and Tea seems to be immensely popular among the people. A shortage in these items and key bakery products, seems to have panicked the people. The surge in demand seems to return back to normal once the storm appears to be subsided.