

Predicting focus countries from news articles' text

Prajwal Seth

<https://github.com/prajwalseth/Country-extractor-from-news-articles>

The task

Given the text of a newspaper article:

"Godolphin's Thunder Snow wins \$10m Dubai World Cup. Dubai may have been hot and humid over the past week but thunder and snow rained on the city on Saturday night. And the royal blue silks, the colours of Godolphin, hung over the spectacular Meydan Racecourse as Thunder Snow won the 23 rd renewal of the \$10 million Dubai World Cup..."

Predict the article's focus country as: **AE** (the 2-letter ISO country code for UAE)


Why is predicting the focus country important?


- To be able to filter news articles from particular countries (main reason)
- To know what locations are present within a news article
- To measure how many articles talk about domestic issues vs international issues
- To be able to quickly annotate articles with their countries at large-scale for further analysis


Why not use neural networks for this task?

- Since there are 196 countries, there would be 196 different output classes
- To train the model, a dataset with at least 1000 articles for each country would be needed. This means $196 \times 1000 = 196,000$ articles in total
- Not only did we not have the time to create such a dataset, but also it is not guaranteed that neural networks would give us a better result
- Existing python libraries like [geotext](#) and [spacy](#) already had location extraction features and were much more lightweight

The Approach: Flashgeotext

 flashgeotext

 Search

 iwpond/flashgeotext
17 Stars · 2 Forks

flashgeotext
flashgeotext
examples
tutorials
reference
discussion

>
>
>

flashgeotext

Extract and count countries and cities (+their synonyms) from text, like [GeoText](#) on steroids using [FlashText](#), a Aho-Corasick implementation. Flashgeotext is a fast, batteries-included (and BYOD) and native python library that extracts one or more sets of given city and country names (+ synonyms) from an input text.

documentation: <https://flashgeotext.iwpond.pw/>
introductory blogpost: <https://iwpond.pw/articles/2020-02/flashgeotext-library>

Table of contents
Usage
Getting Started
Installing
Running the tests
Authors
License
Acknowledgments

Usage

```
from flashgeotext.geotext import GeoText

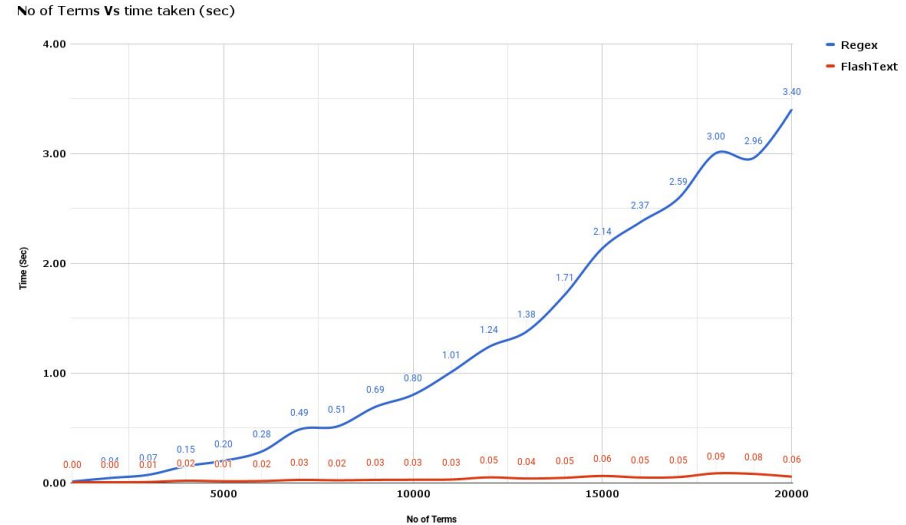
geotext = GeoText(use_demo_data=True)

input_text = '''Shanghai. The Chinese Ministry of Finance in Shanghai said that Ch:
               to cut tariffs on $75 billion worth of goods that the country
               imports from the US. Washington welcomes the decision.'''

geotext.extract(input_text=input_text, span_info=True)
>> {
  'cities': {
    'Shanghai': {
      'count': 2,
      'span_info': [(0, 8), (45, 53)]
    },
    'Washington, D.C.': {
      'count': 1,
      'span_info': [(175, 185)]
    }
  },
  'countries': {
    'China': {
      'count': 1,
      'span_info': [(64, 69)]
    },
    'United States': {
      'count': 1,
      'span_info': [(171, 173)]
    }
  }
}
```

Flashgeotext vs geotext: advantages and drawbacks

- Geotext does not allow the user to add or modify the list of cities, it only relies on a hardcoded list of 22,000 cities
- Flashgeotext includes the inbuilt library of geotext but also allows the user to add their own countries and cities
- However, in practice, I found that the vanilla version of flashgeotext had problems differentiating between normal English words like 'March', 'Of', 'Man', 'Police', 'Central' and city names. It was treating these words as cities



The Four steps to the epiphany

1. Remove punctuation mistakes from the original news article text
2. Delete normal English words that are not supposed to represent city names from the original text
3. Add missing countries and cities which are not in the vanilla version of flashgeotext
4. Assign points to the cities, states, and countries outputted from Flashgeotext based on their occurrence location

The Four steps to the epiphany

1. Remove punctuation mistakes like special characters, words in all-caps, words like 'New York' which were also being recognized as 'York', 'New England' as 'England', etc.

(Original word, replaced word)

("Saint George's", 'Stgeorges')
("Port of Spain", 'Portofspain')
("Porto-Novo", 'Portonovo')
("Porto Novo", 'Portonovo')
("Port-au-Prince", 'Portauprince')
("East Jerusalem", 'Eastjerusalem')
("Equatorial Guinea", 'Equatorialguinea')
("Marshall Islands", 'Marshallislands')
("Papua New Guinea", 'Papuanewguinea')
("Saint Kitts and Nevis", 'Saintkittsandnevis')
("Saint Kitts & Nevis", 'Saintkittsandnevis')
("Saint Kitts", 'Saintkittsandnevis')

(Original word, replaced word)

('New YORK', 'Newyork')
('St. Peter's Square', 'petersq1')
('St. Peters Square', 'petersq1')
('St Peter's Square', 'petersq1')
('Latin America', 'Latin A')
('South America', 'South A')
('Copa America', 'Copa A')
('North America', 'North A')
("Xi'an", 'Xian')
("Democratic Republic of Congo", 'DRC')
("Guinea-Bissau", 'Guineabissau')
("British Columbia", 'Britishcolumbia')

The Four steps to the epiphany

2. Delete normal English words that are not supposed to represent city names from the output of vanilla flashgeotext. If they are country or state names ('Palestine', 'Goa', etc.) then be sure to add them later manually.

```
delete_these_words = ['March', 'Of', 'Union', 'Police', 'Normal', 'Central', 'Most', 'Man', 'Mary', 'Roy', 'Bedi', 'Date', 'Along', 'Much',  
'University', 'Split', 'Tata', 'Nagar', 'Golden', 'George', 'Sale', 'Stuart', 'Bar', 'Antony', 'Chidambaram', 'Wedding',  
'Antony', 'District', 'Correspondent', 'March', 'Photo', 'Mode', 'IST', 'City', 'George', 'Corona', 'Ath', 'Frome', 'Federal', 'count', 'Onex', 'David',  
'Indija', 'Plantation', 'Delta', 'Mission', 'Punch', 'Wedi', 'Metro', 'West', 'Independence', 'Bay', 'Mon', 'Mons', 'Babu',  
'Begun', 'Sabha', 'Since', 'Zlin', 'Kars', 'Rugby', 'Marks', 'Gary', 'Country Club', 'Van', 'Lopez', 'Tours', 'Parys', 'Stade',  
'Mora', 'Morsi', 'Temple', 'Nelson', 'Sparks', 'Siena', 'Anna', 'Rama', 'Since', 'Caen', 'Apex', 'Sabha', 'Sincé', 'Mataram', 'Clinton',  
'Obama', 'Kandahār', 'Tamils', 'Alliance', 'Nokia', 'Tak', 'Maun', 'Samara', 'Rodriguez', 'Kati', 'Soma', 'Tomar', 'Dara', 'Marshall',  
'Paka', 'Best', 'Wādi', 'Thatta', 'The Valley', 'Swords', 'Pandi', 'Pearl', 'Santee', 'Goa', 'Nago', 'Karaman', 'Nandu', 'Hammond', 'Buta',  
'Bid', 'Hassan', 'Mon', 'Pen', 'Amb', 'Martin', 'Hercules', 'Pen', 'Asia', 'Male', 'Robert', 'Le Robert', 'Bani', 'Bern', 'Cordeirópolis', 'Pinto', 'Whitney',  
'Middleton', 'Edison', 'Alvin', 'Barry', 'Melo', 'Carnegie', 'Clive', 'Mercedes', 'Ron', 'Liberal', 'Nigel', 'Borna',  
'Matthews', 'Warren', 'Mitchell', 'Stanton', 'Fairfax Media', 'Pinewood', 'Bath', 'Bell', 'Brak', 'Lindi', 'Metz', 'San', 'Tyler', 'Wright', 'Elizabeth', 'Nancy',  
'Roth', 'Toba', 'Posse', 'Laurel', 'Holiday', 'Hastings', 'Graham', 'Greeley', 'Deal', 'Bello', 'Alma', 'Young', 'Norman',  
'Columbia', 'Pop', 'Sake', 'Papa', 'Vic', 'Spring', 'Franklin', 'Griffin', 'Como', 'Baar', 'Hamilton', 'Marina', 'Sunset', 'Bais', 'Lakewood', 'Green',  
'Pérez', 'Roseburg', 'Sebastian', 'Adam', 'Westlake', 'Cary', 'Harper', 'Latin America', 'Stanley', 'Gap', 'Bryan', 'Cerro', 'Jos',  
'Machado', 'Torres', 'Garcia', 'Bentley', 'Nehe', 'Helena', 'Damme', 'Mobile', 'Begün', 'Nevers', 'Tara', 'Adler', 'Vladimir', 'Harper',  
'Bela', 'Dome', 'Gap', 'Reading', 'Frederick', 'Biga', 'Arona', 'Salt', 'Phrae', 'Vitória', 'Santos', 'Izmayil', 'Dixon', 'Pérez', 'Molina',  
'Gálvez', 'Khanna', 'Kara', 'Ghat', 'Indija', 'Split', 'Castro', 'Minas', 'Fleet', 'Ipoh', 'Summit', 'Manga', 'English', 'Manage', 'Worms', 'Banning', 'Prince  
George', 'Vogan', 'Le Port', 'Flint', 'Dig', 'Durg', 'Osh', 'Trento', 'Wa', 'Longview', 'Buy', 'Erba', 'Bear', 'Born', 'La Trinidad', 'Nuku'alofa',  
'Port-au-Prince', 'Port-of-Spain', 'Saint George's', 'St. John's', 'Gay', 'Sulphur', 'Co.', 'Brent', 'Una', 'Goes', 'Hi', 'Grasse', 'Münster',  
'Opportunity', 'Wyckoff', 'Gallup', 'Honda', 'Plunge', 'Luce', 'Surprise', 'Thän', 'Than', 'Scarborough', 'La Possession', 'Superior', 'Voi', 'Prince Albert',  
'Drama', 'Bam', 'Mi', 'Drama', 'Brits', 'Dax', 'Kula', 'Sim', 'Shingū', 'Boo', 'Klin', 'Lucé', 'Lota', 'Aba', 'Peer', 'Ieper', 'Cañete',  
'Canning', 'Colón', 'Babat', 'Dour', 'Fatwa', 'Iba', 'Laon', 'Sama', 'Uman', 'Saki', 'Bagé', 'Fargo', 'Logan', 'Pátra', 'Eagle', 'Palín', 'Chebba', 'Humble', 'Rio  
Tinto', 'Gao', 'Palo', 'Centenario', 'Hong Kong', 'Rio Tinto', 'Crystal', 'Magna', 'Itu', 'Butterworth', 'Pio', 'Salwá', 'Jena', 'Praia',  
'Cocoa', 'Asha', 'Benoy', 'Wedding', 'Leer', 'Liberty', 'Melle', 'Kaka', 'Nidda', 'Ripley', 'Achim', 'Safed', 'Auerbach', 'Albany', 'Cambridge', 'Nederland',  
'Loni', 'Derby', 'Kashi', 'Dalai', 'Lamu', 'Shangri-La', 'Terrace', 'Sami', 'Barking', 'Fleetwood', 'Weston', 'Porto', 'George Town', 'Granada', 'Palestine',  
'Guinea-Bissau', 'Jawhar', 'Mexico', 'Mexico City', 'Ho', 'Roses', 'Rifu', 'Ode', 'Or.', 'Polish', 'Dabou', 'Lala', 'Nysa', 'Vidin', 'Kleve', 'Perm']
```

This list was created manually. However, there is also an automated way to come up with a list like this (next slide)

The Four steps to the epiphany

2A. Use the Stanford NER (Named Entity Recognition) tool to remove names from the list of cities

```
delete_these_also=['Widnes','Sabhā','Columbus','Indi','Wakefield','Blaine','Christiana','Molina','Rheine','Samara','Esperanza',  
'Clifton','Burton','Dole','Babu','Safi','Melville','Fuwah','Bria','Fontaine','Horst','Perris','Dādri','Numan','Kari','Keene',  
'Napoli','Harwich','Keller','Altena','Haydock','Fondi','Chekhov','Randolph','Tokio','Rezé','Saky','Gaspar','Massa','Takahashi',  
'Brody','Norton','Easley','Alameda','Eden','Sherwood','Irving','Voerde','Sutton','Caivano','Antony','Terrell','Estelle','Gresham',  
'Mercedes','Ara','Thornton','Selby','Markham','Hull','Vigo','Medina','Sari','Kasūr','Matiāri','Ōhara','Pau','Warburton',  
'Tupi','Roosevelt','Bartlett','Mariano','Carrefour','Saint-André','Arona','Rafah','Ghazni','Hatfield','Ajalpan','Ramsey','Carolina',  
'Buffalo','Soria','Milas','Nizhniy','Novgorod','McDonough','Blyth','Assen','Burke','Salvador','Albemarle','Dyer','Nyborg',  
'Saint-Louis','Kanda','Kalyān','Pinto','Baar','Rosenberg','Siuri','Bedford','Oviedo','Tiel','Mendoza','Morton','Yola','Chiba',  
'Saint','Mahe','Radcliffe','Hille','Slonim','Oran','Sidi','Slimane','Aki','Werne','Huntley','Griffith','Gori','Ruma','Teresa',  
'Girona','Hammond','Morsi','Westlake','Amos','Córdoba','Ānand','Kanekomachi','Melton','Miri','Canela','Murakami','Belah',  
'Cranston','Monzón','Mariupol','Mariana','Marrero','Barrie','Vernon','Alicia','Cornelius','Hof','Barbosa','Zama','Sogcho','Mansfield',  
'Troyan','Aso','Zemun','Hod','HaSharon','Enna','Nāwa','Sheridan','Kazan','Milton','Keynes','Kōnan','Roxas','Valls','Hatton',  
'Speyer','Hagen','Lafayette','Matsumoto','Cairns','Lawton','Lancaster','Cueto','Patti','Alegre','Mut','Gulu','Alès','Goiás',  
'Homer','Glen','Bilgi','Juba','Harper','Stirling','Garland','Wāling','Durant','Santee','Ajaccio','Tipton','Halle','Lacey',  
'Biga','Monroe','Castro','Mandi','Sousa','Vista','Mendes','Pimentel','Borås','Kenner','Franca','Roth','Plato','La','Madeleine',  
'Nanterre','Sevilla','Griffin','Middleton','Rosso','Sinan','Glan','Al','Mukallā','Stafford','Empoli','Palu','Rahden','Ramos',  
'Leduc','Conde','Passos','Izumi','Bo','Westminster','Iglesias','Balfour','Rhondda','McKinney','Bissau','Preston','Ariana',  
'Nehe','Metz','Marino','Parys','Hurst','Talladega','Saint-Germain-en-Laye','Lens','Jonesboro','Maldonado','Amancio','Brak',  
'Malyn','Newton','Antioch','Norwood','Mons','Bais','Nowa','Sól','Vaughan','Jāwad','Prospect','David','Elizabeth','Chester',  
'Lopez','Howard','Rosario','Morales','Rodriguez','Moore','Kyle','Corona','Ādam','Walker','Le','Robert','Los','Reyes','Whitley',  
'Bay','Allen','Graham','Ron','Martin','Bello','León','Nancy','Arnold','Tucker','Lamont','Franklin','Tyler','Anderson','Roy',  
'Nirmal','Douglas','Masuda','Langley','Harvey','Carney','Mitcham','Yara'...]
```

The Four steps to the epiphany

3. Add missing countries and cities to the vanilla version of flashgeotext

- A. Indian cities
- B. US states and their abbreviations
- C. Canadian states and their abbreviations
- D. Plural versions of nationalities
- E. 31 countries for which vanilla flashgeotext was not giving any output at all
- F. Numerous other cities which were mislabeled (for example, the capital of India 'Delhi' was being labeled as 'US')
- G. City-states like Singapore, Vatican City, Monaco and transnational entities like UN, EU

The Four steps to the epiphany

3A. Added Indian cities

- I was provided with a list of 2339 Indian cities by my advisor. This list had been manually created by another team previously. Adding them significantly improved the quality of predictions for news articles talking about India
- Additionally, I created a list with all Indian states and added it to the list of Indian cities as well so that they would be picked up by Flashgeotext

```
indian_state_names = ["Andhra Pradesh", "Arunachal Pradesh",  
    ", "Assam", "Bihar", "Chhattisgarh", "Goa", "Gujarat", "Haryana", "Himachal Pradesh",  
    ", "Jammu and Kashmir", "Jharkhand", "Karnataka", "Kerala", "Madhya Pradesh",  
    ", "Maharashtra", "Manipur", "Meghalaya", "Mizoram", "Nagaland", "Odisha",  
    ", "Punjab", "Rajasthan", "Sikkim", "Tamil Nadu", "Telangana", "Tripura", "Uttar Pradesh",  
    ", "Uttarakhand", "West Bengal", "Andaman and Nicobar Islands", "Chandigarh",  
    ", "Dadra and Nagar Haveli", "Daman and Diu", "Lakshadweep", "National Capital Territory of Delhi", "Puducherry"]
```

	State	City	District	COORD_X	COORD_Y
1	Uttar Pradesh	Agra	Agra	78.5325729	26.88660294
2	Uttar Pradesh	Agra	Agra	77.9902797	27.15731954
3	Rajasthan	Pali	Bali	73.15426383	25.14652137
4	Maharashtra	Pune	Bhor	73.82704729	18.18031484
5	Madhya Pradesh	Dhar	Dhar	75.38378363	22.55063457
6	Chhattisgarh	Durg	Durg	81.32568244	21.14344278
7	Uttar Pradesh	Etah	Etah	78.66418024	27.55456811
8	Bihar	Gaya	Gaya	84.94307465	24.72639441
9	Maharashtra	Pune	Ghod	73.85364002	19.04152249
10	Madhya Pradesh	Guna	Guna	77.19997779	24.80387644
11	Haryana	Jind	Jind	76.35810997	29.34641657
12	Maharashtra	Kaij	Kaij	76.07783271	18.75886362
13	Rajasthan	Kota	Kota	75.88464036	25.0454217
14	Puducherry	Mahe	Mahe	75.28279161	12.00638841
15	Punjab	Moga	Moga	75.15987779	30.75492753
16	Rajasthan	Pali	Pali	73.19444088	25.77594508
17	Maharashtra	Pune	Paud	73.54902371	18.51629811
18	Nagaland	Phek	Phek	94.56595408	25.63264549
19	Odisha	Puri	Puri	86.07152387	19.91992727
20	Madhya Pradesh	Rewa	Rewa	81.3742883	24.55198482
21	Rajasthan	Tonk	Tonk	75.76826196	26.28768003
22	Rajasthan	Jalor	Ahor	72.84914157	25.50151092
23	Maharashtra	Akola	Akot	77.07880323	21.03654431
24	Himachal Pradesh	Solan	Arki	76.9176186	31.21578827
25	Rajasthan	Baran	Atru	76.65398792	24.75239841
26	Maharashtra	Latur	Ausa	76.54271136	18.19802315
27	Bihar	Patna	Barh	85.72936064	25.40263112
28	Gujarat	Patan	Sami	71.70550604	23.62428897
29	Arunachal Pradesh	Anjaw	Tezu	96.95527941	28.0527876
30	Arunachal Pradesh	Lohit	Tezu	96.18558787	27.86239019
31	Madhya Pradesh	Ratlam	Alot	75.46822199	23.72825235
32	Karnataka	Hassan	Alur	75.95380874	12.91910205
33	Rajasthan	Jaipur	Amer	75.75994432	27.0263653
34	Maharashtra	Wardha	Arvi	78.35269813	20.96256276
35	Rajasthan	Jaipur	Dudu	75.13583697	26.67481896

The Four steps to the epiphany

3B. Added US states and their abbreviations

- I created a list of traditional abbreviations used by US states and added it to the list of locations for the US
- I also added all US state names and their 2 letter abbreviations mixed with special characters, such as 'N.Y.', 'N.Y.', 'NY.', 'NY.', 'Ca.'



abb1	abb2	abb3
Ala.		
Alaska		
Ariz.		
Ark.		
Calif.	Cal.	
Colo.	Col.	
Conn.		
Del.		
Fla.	Flor.	
Ga.		
	H.I.	
Idaho	Ida.	Id.
Ill.		
Ind.		
Iowa	Ia.	
Kans.	Kan.	
Ky.	Ken.	Kent.
La.		
	Me.	
Md.		
Mass.		
Mich.		
Minn.	M.N.	MN
Miss.		
Mo.		
Mont.		
Nebr.	Neb.	
Nev.		
N.H.		

```
us_state_names = ["Alaska", "Alabama", "Arkansas", "American Samoa",  
"Arizona", "California", "Colorado", "Connecticut", "Delaware", "Florida",  
"Guam", "Hawaii", "Iowa", "Idaho", "Illinois", "Indiana", "Kansas",  
"Kentucky", "Louisiana", "Massachusetts", "Maryland", "Maine", "Michigan",  
"Minnesota", "Missouri", "Mississippi", "Montana", "North Carolina", "North  
Dakota", "Nebraska", "New Hampshire", "New Jersey", "New Mexico", "Nevada",  
"Ohio", "Oklahoma", "Oregon", "Pennsylvania", "Puerto Rico", "Rhode Island",  
"South Carolina", "South Dakota", "Tennessee", "Texas", "Utah", "Virginia",  
"Virgin Islands", "Vermont", "Wisconsin", "West Virginia", "Wyoming"]
```

```
us_abbreviations = ["AL", "AK", "AZ", "AR", "CA", "CO", "CT", "DC", "DE",  
"FL", "GA", "HI", "ID", "IL", "IN", "IA", "KS", "KY", "LA", "ME", "MD",  
"MA", "MI", "MN", "MS", "MO", "MT", "NE", "NV", "NH", "NJ", "NM", "NY",  
"NC", "ND", "OH", "OK", "OR", "PA", "RI", "SC", "SD", "TN", "TX", "UT",  
"VT", "VA", "WA", "WV", "WI", "WY", "PR", "AS", "MP", "CZ", "VI"]
```

The Four steps to the epiphany

3C. Added Canadian states and their abbreviations along with special characters. Removed 'Oh.', 'On,' as they were occurring as English words in news articles

```
can_province_names = { 'AB': 'Alberta', 'BC': 'British Columbia',  
                        'MB': 'Manitoba', 'NB': 'New Brunswick', 'NL': 'Newfoundland and  
                        Labrador', 'NS': 'Nova Scotia', 'NT': 'Northwest Territories',  
                        'NU': 'Nunavut', 'ON': 'Ontario', 'PE': 'Prince Edward Island',  
                        'QC': 'Quebec', 'SK': 'Saskatchewan', 'YT': 'Yukon' }  
  
can_abbreviations2=  
[ 'AB', 'Ab.', 'AB.', 'AB', 'Ab', 'A.B.', 'A.b.', 'BC', 'Bc.', 'BC.', 'BC',  
  'Bc', 'B.C.', 'B.c.', 'MB', 'Mb.', 'MB.', 'MB', 'Mb', 'M.B.', 'M.b.', 'N  
B', 'Nb.', 'NB.', 'NB', 'Nb', 'N.B.', 'N.b.', 'NL', 'NL.', 'NL.', 'NL', 'N  
L', 'N.L.', 'N.l.', 'NS', 'Ns.', 'NS.', 'NS', 'Ns', 'N.S.', 'N.s.', 'NT',  
  'Nt.', 'NT.', 'NT', 'Nt', 'N.T.', 'N.t.', 'NU', 'Nu.', 'NU.', 'NU', 'Nu',  
  'N.U.', 'N.u.', 'ON', 'ON.', 'ON', 'O.N.', 'O.n.', 'PE', 'Pe.', 'PE.', 'PE  
', 'Pe', 'P.E.', 'P.e.', 'QC', 'Qc.', 'QC.', 'QC', 'Qc', 'Q.C.', 'Q.c.',  
  'SK', 'Sk.', 'SK.', 'SK', 'Sk', 'S.K.', 'S.k.', 'YT', 'Yt.', 'YT.', 'YT',  
  'Yt', 'Y.T.', 'Y.t.' ]
```

The Four steps to the epiphany

3D. Added 276 plural nationalities with their country code

Cuba	Cuban	CU
Curacao	Curacaoan	CW
Cyprus	Cypriot	CY
Djibouti	Djiboutian	DJ
Dominica	Dominican	DM
East Timor	Timorese	TL
Equatorial Guinea	Equatorial Guinean	GQ
Eritrea	Eritrean	ER
Falkland Islands	Falkland Islander	FK
Faroe Islands	Faroese	FO
Fiji	Fijian	FJ
French Polynesia	French Polynesian	PF
Gabon	Gabonese	GA
Gambia	Gambian	GM
Georgia	Georgian	GE
Gibraltar	Gibraltar	GI
Greenland	Greenlandic	GL
Grenada	Grenadian	GD
Guam	Guamanian	GU
Guernsey	Channel Islander	GG
Guinea	Guinean	GN
Guinea-Bissau	Bissau-Guinean	GW
Guyana	Guyanese	GY
Hong Kong	Cantonese	HK
Isle of Man	Manx	IM
Jersey	Channel Islander	JE
Kiribati	Kiribati	KI
Kosovo	Kosovar	XK
Kuwait	Kuwaiti	KW
Kyrgyzstan	Kyrgyzstani	KG

country	plural_country	match	country_code
Alghanistan	Alghan	240	AF
Algeria	Algerian	239	DZ
Angola	Angolan	238	AO
Argentina	Argentine	237	AR
Austria	Austrian	236	AT
Australia	Aussie	235	AU
Australia	Australian	235	AU
Australia	Austn		AU
Bangladesh	Bangladeshi	234	BD
Belarus	Belarusian	233	BY
Belgium	Belgian	232	BE
Bolivia	Bolivian	231	BO
Bosnia and Herzegovina	Bosnian	230	BA
Bosnia and Herzegovina	Herzegovinian	230	BA
Brazil	Brazilian	229	BR
Britain	Brit	#N/A	GBX1
Britain	British	#N/A	GBX1
Bulgaria	Bulgarian	228	BG
Cambodia	Cambodian	227	KH
Cameroon	Cameroonian	226	CM
Canada	Canadian	225	CA
Central African Republic	Central African	224	CF
Chad	Chadian	223	TD
China	Chinese	222	CN
Colombia	Colombian	221	CO
Costa Rica	Costa Rican	220	CR
Croatia	Croatian	219	HR
Czech Republic	Czech	218	CZ
Democratic Republic of the Congo	Congolese	217	CD
Denmark	Danish	216	DK
Ecuador	Ecuadorian	215	EC
Egypt	Egyptian	214	EG
El Salvador	Salvadoran	213	SV
England		#N/A	GB
Estonia	Estonian	212	EE
Ethiopia	Ethiopian	211	ET
Finland	Finnish	210	FI
France	French	209	FR
Germany	German	208	DE
Ghana	Ghanaian	207	GH
Greece	Greek	206	GR
Guatemala	Guatemalan	205	GT
Holland	Dutch	#N/A	NL

The Four steps to the epiphany

3E. Added 31 countries and their capitals for which vanilla Flashgeotext was not giving any output

```
missing_countries =  
['IT','JM','GD','CL','MV','AG','CR','TT','LR','TO','CH','PE','BO','BG','TD','HT','PW','PA','FM','IE','CA','P  
R','NL','NR','AZ','SC','LK','MD','BN','GY','MM']  
missing_countries_names = {'IT': 'Italy', 'JM': 'Jamaica', 'GD': 'Grenada', 'CL': 'Chile', 'MV': 'Maldives',  
'AG': 'Antigua and Barbuda', 'CR': 'Costa Rica', 'TT': 'Trinidad and Tobago', 'LR': 'Liberia', 'TO':  
'Tonga', 'CH': 'Switzerland', 'PE': 'Peru', 'BO': 'Bolivia', 'BG': 'Bulgaria', 'TD': 'Chad', 'HT': 'Haiti',  
'PW': 'Palau', 'PA': 'Panama', 'FM': 'Federated States of Micronesia', 'IE': 'Republic of Ireland', 'CA':  
'Canada', 'PR': 'Puerto Rico', 'NL': 'Kingdom of the Netherlands', 'NR': 'Nauru', 'AZ': 'Azerbaijan', 'SC':  
'Seychelles', 'LK': 'Sri Lanka', 'MD': 'Moldova', 'BN': 'Brunei Darussalam', 'GY': 'Guyana', 'MM':  
'Myanmar'}  
missing_countries_capitals = {'IT': 'Rome', 'JM': 'Kingston', 'GD': "St. George's", 'CL': 'Santiago', 'MV':  
'MalÃ©', 'AG': "St. John's", 'CR': 'San Jose', 'TT': 'Port of Spain', 'LR': 'Monrovia', 'TO': "Nuku'alofa",  
'CH': 'Bern', 'PE': 'Lima', 'BO': 'Sucre', 'BG': 'Sofia', 'TD': "N'Djamena", 'HT': 'Port-au-Prince', 'PW':  
'Ngerulmud', 'PA': 'Panama City', 'FM': 'Palikir', 'IE': 'Dublin', 'CA': 'Ottawa', 'PR': 'San Juan', 'NL':  
'Amsterdam', 'NR': 'Yaren', 'AZ': 'Baku', 'SC': 'Victoria, Seychelles', 'LK': 'Sri Jayewardenepura Kotte',  
'MD': 'Chisnau', 'BN': 'Bandar Seri Begawan', 'GY': 'Georgetown', 'MM': 'Naypyidaw'}
```


The Four steps to the epiphany

3F. Numerous other cities and entities which were mislabeled or not present in vanilla Flashgeotext

- (["Stjohns", 'AG17')
- (["St. George's", 'GD17')
- (["Portofspain", 'TT17')
- (["Malé", 'MV17')
- (["Nuku'alofa", 'Nukualofa', 'T017')
- (["BSE", 'Bse', 'B.S.E.', 'B.s.e.', 'Rbi', 'RBI', 'Cbi', 'Cid', 'CBI', 'CID', 'IN17')
- (['Kempegowda Nagar', 'Kempegowdanagar', 'Perambur', 'Vallanadu', 'Thoothukudi', 'Bombay Stock Exchange', 'Kashmir', 'Kashmiri', 'Erragadda', 'Nagapattinam', 'Nagappattinam', 'Taj Mahal', 'Tiruchi', 'Bollywood', 'Keerapakkam', 'Tambaram', 'Kannada', 'Indrakeeladri', 'Dhinkia', 'Havelock Island', 'Budaun', 'IN18')
- (["dharna", 'Dharna', 'IN19')
- (['Hobart', 'Sydney', 'Perth', 'Melbourne', 'NSW', 'New South Wales', 'Queensland', 'Tuggeranong', 'Bondi Beach', 'Flemington', 'Warracknabeal', 'Tasmania', 'Belconnen', 'Lesmurdie', 'Donvale', 'Weetangera', 'Lancefield', 'Cobaw', 'Benloch', 'Pastoria', 'Baynton', 'Saint George', 'St George', 'Kogarah', 'Victoria', 'AUX')
- (['Dhaka', 'BDX')
- (['Athens', 'Moria', 'GRX')
- (['Jalalabad', 'Kandahar', 'Herat', 'AFX')
- (['Paris', 'Lyon', 'Dordogne', 'Stade de France', 'FRX')
- (['Niger Delta', 'Borno', 'Nigeria', 'Lagos', 'Ogun', 'NGX')
- (['EU', 'European Union', 'E.U.', 'Eu', 'European Commission', 'European Investment Bank', 'Euro', 'euro', 'EUX')
- (['UN', 'United Nations', 'U.N.', 'Un', 'UNX')
- (['Bali', 'IDX')
- (['Damascus', 'SYX')
- (['Santa Maria', 'Rio Olympic', 'Rio Olympics', 'Rio 2016', 'BRX')

The Four steps to the epiphany (3F)

- (['Moscow', 'Saint Petersburg', 'St. Petersburg'], 'RUX')
- (['Newengland', 'Albany', 'Tippecanoe', 'St. Augustine', 'Vilano Beach', 'St. Tammany', 'District of Columbia', 'Bryant Park', 'Charlotte', 'Boro Park', 'Roseburg', 'Marrero', 'Montauk', 'Long Island', 'Capitol Hill', 'Syracuse', 'Georgetown', 'New Smyrna Beach', 'Palm Beach', 'North Little Rock', 'St. Louis', 'Port Saint Lucie', 'Central Park', 'America', 'Queens', 'Lawrence', 'Williamsburg', 'Washington', 'Washington, D.C.', 'Prairie View', 'Santa Clara', 'San Jose', 'Santa Cruz', 'Hudson Valley', 'Kitsap County', 'Harvard', 'Mount Sinai', 'Putnam County', 'Goddard Park', 'Hawaii', 'Davis', 'Calaveras County', 'Wall Street', 'Carroll County', 'Humboldt County', 'Neshannock', 'Penn', 'Yale', 'Federal Reserve', 'NYC', 'N.Y.C.', 'San Francisco', 'San Mateo', 'Worcester', 'Johnson City', 'Cdc', 'Harlem', 'NYPD', 'St Louis', 'Fox Lake', 'Pentagon', 'Richmond', 'Telluride', 'White House', 'York', 'Newyork'], 'USX1')
- (['Newmexico', 'Usa', 'Dow', 'Buffalo', 'BUFFALO', 'Democrat', 'Democrats', 'Republican', 'Republicans', 'Senator', 'FBI', 'Fbi', 'F.B.I.', 'Us', 'Nasa', 'NASA', 'Sen.', 'Fed', 'GOP', 'Gop', 'Centers for Disease Control and Prevention'], 'USX2')
- (us_state_names, 'US11')
- (['Haiti', 'Portauprince'], 'HTX')
- (['Mali'], 'MLX')
- (['Cabo Delgado', 'Ancuabe', 'Mozambique'], 'MZX')
- (['Pama'], 'BFX')
- (['Papuanewguinea'], 'PGX')
- (['Grimsby', 'Lincolnshire', 'Widnes', 'Middlesex', 'Cambridge', 'Oxford', 'London', 'Ascension Island', 'Newcastle', 'Somerset', 'Bristol', 'Carterton', 'Derbyshire', 'West Hampstead', 'Romsey', 'Tottenham', 'Sky Sports', 'West Hampstead', 'Kent', 'Brighton', 'Cornwall', 'Westminster', 'Boosbeck', 'Yorkshire', 'Chelsea', 'Manchester', 'Norwich', 'Norfolk', 'Birmingham', 'England', 'Wales', 'Northern Ireland', 'Scotland', 'Manchester', 'Richmond and Twickenham', 'Britain', 'Clogher', 'Wandsworth', 'Cronton', 'Inverurie', 'Peak District', 'Snowdonia', 'Dorset', 'Dundee', 'Salisbury', 'Suffolk', 'York'], 'GBX3')
- (['Uk', 'Gbr', 'GBR', 'Boris Johnson', 'NHS', 'Nhs', 'nhs', '£'], 'GBX4')
- (['Vigui'], 'PAX')
- (['Somalia'], 'S011')
- (['Wellington', 'Albert Town', 'Queenstown', 'Napier', 'Waitarere', 'Levin'], 'NZ11')
- (['Frankfurt', 'Nürnberg', 'Nuremberg', 'Cologne', 'Jena', 'Melle', 'Achim', 'Auerbach'], 'DE11')
- (['Venice', 'Milan', 'San Siro', 'San Donato Milanese'], 'IT11')

The Four steps to the epiphany (3F)

- (['Vienna', 'Nickelsdorf', 'Sölden', 'Soelden'], 'AT11')
- (['Syria', 'Idlib'], 'SY11')
- (['Yemen', 'Mokha', 'Mocha', "Sa'dah", 'Saada'], 'YE11')
- (['Vernon', 'Trans Mountain', 'TransMountain', 'Newfoundland', 'Labrador', 'Harrison Hot Springs', 'Victoria'], 'CA11')
- (['Tiananmen Square', 'Guangdong', 'Peking', 'Urumqi', "Xi'an", "Xi'an", 'Xian', 'Jiangsu', 'Hunan', 'Kashgar'], 'CN11')
- (['Duma', 'Hamas', 'Eastjerusalem'], 'PS11')
- (['Roszke'], 'HU11')
- (['El Salvador'], 'SV11')
- (['DRC', 'Kasai', 'Drc'], 'CD11')
- (['Rcongo'], 'CG11')
- (['Naguru', 'Kampala', 'Uganda', 'Entebbe'], 'UG11')
- (['Cotopaxi', 'Ecuador'], 'EC11')
- (['Jerusalem', 'Hadera', 'Yerushalayim', 'Bnei Brak', 'Safed'], 'IL11')
- (['Cyprus'], 'CY11')
- (['Santo Domingo', 'Dominican Republic'], 'D011')
- (['Chiang Mai'], 'TH11')
- (['Holland', 'Nederland'], 'NL12')
- (['Bermuda'], 'BM11')
- (['Kim Jong Un', 'DPRK', 'Dprk', 'D.P.R.K.'], 'KP11')
- (['Tenerife', 'Canary Islands', 'Barcelona', 'Catalonia'], 'ES11')
- (['Tshwane', 'Limpopo', 'Maritzburg', 'Mankweng', 'Manenberg', 'Ficksburg', 'Bellville', 'Ekurhuleni', 'Suikerbosrand', 'Eastern Cape', 'Western Cape', 'De Doorns', 'Daveyton', 'Gauteng', 'De Deur', 'Nongoma', 'Khayelitsha', 'Modderpoort', 'Free State', 'Mafikeng', 'Mosselbay', 'Mossel Bay', 'Pollsmoor', 'Tlokwe', 'Brandfort', 'Boikhutso', 'Welgevonden', 'Goedgevonden', 'Kakamas', 'Sterkspruit', 'Engcobo', 'Ngcobo', 'Gugwini', 'Silindile', 'Moyeni'], 'ZA11')
- (['Mahé', 'Seychelles'], 'SC11')
- (['Gorgan', 'Golestan', 'Tehran', 'Bandar Abbas'], 'IR11')
- (['Qatar'], 'QA11')

The Four steps to the epiphany (3F)

- (['Zurich', 'Geneva'], 'CH11')
- (['Sint Maarten', 'St. Maarten', 'St Maarten'], 'SX11')
- (['UAE', 'Uae', 'U.A.E.'], 'AE11')
- (['Bahamas'], 'BS11')
- (['Hanoi', 'Ha Noi', 'Vu Duc Dam', 'Ho Chi Mihn City', 'Hà Nội'], 'VN11')
- (['Gothenburg', 'Sweden'], 'SE11')
- (['Muranga'], 'KE11')
- (['Belize', 'Bdf'], 'BZ11')
- (['Antarctica'], 'AQ11')
- (['Niger'], 'NE11')
- (['Murilo'], 'FM11')
- (['Fukushima', 'Omura', 'Fukuoka'], 'JP11')
- (['Porto'], 'PT11')
- (['Apia', 'Samoa'], 'WS')
- (['Libya'], 'LY11')
- (['Istanbul'], 'TR11')
- (['Fiji'], 'FJ11')
- (['Nuuk', 'Greenland'], 'GL11')
- (['Tajikistan'], 'TJ11')
- (['Uzbekistan'], 'UZ11')
- (['Auki', 'Solomon Islands'], 'SB11')
- (['Chernobyl'], 'UA11')
- (['Hong KONG'], 'HK12')
- (['Malta'], 'MT11')
- (['Namibia'], 'NAX11')
- (['Bosnia-Herzegovina'], 'BA11')
- (['Dominica'], 'DM11')

The Four steps to the epiphany (3F)

- (['Togo'], 'TG11')
- (['St. Peter's Square', 'Saint Peter's Square', 'St. Peter's Basilica', 'Saint Peter's Basilica', 'petersq1', 'peterbas2'], 'VA11')
- (['Ashanti Region', 'Ho Municipal', 'Ho Municipality'], 'GH11')
- (['Praia'], 'CV11')
- (['Mullaitivu'], 'LK14')
- (['MexicoLocality', 'MexicoCountry', 'Iguala'], 'MX14')
- (['Sierra Leone'], 'SL14')
- (['Porto-Novo', 'Porto Novo', 'Portonovo'], 'BJ14')
- (['San José'], 'CR14')
- (["Stgeorges"], 'GD14')
- (["Chisinau"], 'MD14')
- (["Victoria"], 'SC14')
- (["Juba"], 'SS14')
- (["Sri Jayawardenapura Kotte"], 'LK15')
- (['George Town'], 'MY15')
- (['Angola'], 'AO21')
- (['Antigua and Barbuda', 'Antigua & Barbuda', 'Antigua', 'Barbuda'], 'AG21')
- (['Bosnia and Herzegovina', 'Bosnia & Herzegovina', 'Bosnia', 'Herzegovina'], 'BA21')
- (['EquatorialGuinea'], 'GQ21')
- (['Eritrea'], 'ER21')
- (['Gambia'], 'GM21')
- (['Granada'], 'NI21')
- (['Guinea'], 'GN21')
- (['Honduras'], 'HN21')
- (['Liechtenstein'], 'LI21')
- (['Macedonia'], 'MK21')
- (['Marshallislands'], 'MH21')

The Four steps to the epiphany (3F)

- (['Mauritius'], 'MU21')
- (['Micronesia'], 'FM21')
- (['Mongolia'], 'MN21')
- (['Nicaragua'], 'NI21')
- (['Palestine'], 'PS21')
- (['Paraguay'], 'PY21')
- (['Senegal'], 'SN21')
- (['Swaziland'], 'SZ21')
- (['Turkmenistan'], 'TM21')
- (['Uruguay'], 'UY21')
- (['Saintkittsandnevis'], 'KN21')
- (['Saintlucia'], 'LC21')
- (['Saintvincentandthegrenadines'], 'VC21')
- (['Southsudan'], 'SS21')
- (['Trinidadandtobago'], 'TT21')
- (['Unitedarabemirates'], 'AE21')
- (['Guineabissau'], 'Bissau'], 'GW21')
- (['Georgia'], 'GE21')
- (['Britishcolumbia'], 'CA21')
- (['Ciudad Constitución'], 'PE21')
- (['Polish2'], 'PL21')
- (['Dabou'], 'CI21')
- (['Balochistan' , 'Khyber Pakhtunkhwa' , 'Sindh'], 'PK11')
- (['Killarney' , 'Dublin'], 'IE11')
- (['Saint Joseph' , 'St. Joseph'], 'TT11')
- (['Hong Kong' , 'HK' , 'H.K.' , 'Causeway Bay' , 'Mong Kok' , 'Lung Wo Road'], 'HK11')

The Four steps to the epiphany

4. Assign points to the cities, states, and countries outputted from Flashgeotext based on their occurrence in the body of the news article text

- **+17 points** if the city/state/country is in the first 20 characters of the news article text
- **+10 points** if the city/state/country is between the first 20 & 100 characters of the news article text
- **+8 points** if the city/state/country is between the first 100 & 250 characters of the news article text
- **+2 points** if the city/state/country is between the first 250 & 1000 characters of the news article text
- **+1 point** if the city/state/country is after 1000 characters of the news article text
- **+3 points** if the city/state/country has no month names (January, February, etc.) within 20 characters in front of it or behind it, and if it is located in the first 250 characters of the news article text
- **+25 points** if 'in' is in the 3 characters preceding the location's first character and the location is in the first 50 characters of the news article text
- **+20 points** if 'in' is in the 3 characters preceding the location's first character and the location is between the first 50 and 100 characters of the news article text
- **+15 points** if 'in' is in the 3 characters preceding the location's first character and the location is between the first 100 and 250 characters of the news article text

The Four steps to the epiphany

4. Assign points to the cities, states, and countries outputted from Flashgeotext based on their occurrence in the body of the news article text

```
{
  'ae11_cities': {
    'Uae': {
      'count': 2,
      'span_info': [(427, 430), (1965, 1968)]},
    'aeyc106_cities': {
      'Emirati': {
        'count': 2,
        'span_info': [(724, 731), (2858, 2866)]},
      'cities': {
        'Dubai': {
          'count': 9,
          'span_info': [(35, 40), (52, 57), (317, 322), (444, 449), (521, 526), (600, 605), (1268, 1273), (1403, 1408), (1625, 1630)]},
          'Pisa': {
            'count': 1,
            'span_info': [(3928, 3932)]},
          'countries': {
            'United States': {
              'count': 1,
              'span_info': [(910, 913)]},
            'us11_cities': {
              'California': {
                'count': 2,
                'span_info': [(1717, 1727), (4229, 4239)]},
                'Kentucky': {
                  'count': 1,
                  'span_info': [(3622, 3630)]},
                'usx1_cities': {
                  'America': {
                    'count': 1,
                    'span_info': [(1952, 1960)]},
                    'usx2_cities': {},
                    'usyc107_cities': {
                      'American': {
                        'count': 2,
                        'span_info': [(961, 969), (2387, 2395)]},
                        'Americans': {
                          'count': 1,
                          'span_info': [(2139, 2148)]}
                    }
                  }
                }
              }
            }
          }
        }
      }
    }
  }
```

Output from Flashgeotext

AE is the country code with maximum points (only look at the first 2 characters, ignore numbers or extra alphabets)



OrderedDict([('BEYC10', 1), ('IT', 2), ('USX1', 13), ('US11', 13), ('USYC107', 13), ('US', 13), ('AE11', 52), ('AEYC106', 52), ('AE', 52)])

Output and conclusion

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
	text	locations	cleaned_text	country_list_flash	country_list_names											
0	Krewe du Vieux	['Frenchmen', 'New Orleans']	Krewe du Vieux 2018: Take a virtual tour	Ordered[US]		['Mandeville', 'Chartres', 'New Orleans', 'Washington', 'Louisiana', 'Mississippi', 'French', 'German', 'Greek', 'Irish', 'Americans']										
1	Turkey shells	['Afrin', 'Bri']	Turkey shells Syrian city as it pushes into	Ordered[TR]		['Beirut', 'Afrin', 'Turkey', 'United Kingdom', 'Britain', 'Syria', 'Syrian', 'Turkish']										
2	A new development	['Jamison', 'A new development in Gainesville, Va.,']	A new development in Gainesville, Va.,	Ordered[US]		['Gainesville', 'Charleston', 'Granville', 'Somerville', 'Manassas', 'Fairfax', 'Centreville', 'Long Island', 'Va.', 'Virginia', 'Vienna']										
3	Laura Ingraham	['The Wash']	Laura Ingraham takes an Easter break at	Ordered[US]		['Parkland', 'San Marcos', 'Anglet', 'Gillette', 'Washington', 'America', 'Florida', 'California']										
4	Perm Secre	['Transform']	Perm Secretary Launches Book for Africa	Ordered[NG]		['Abuja', 'Nigeria', 'Nigerian']										
5	Putin's spokesman	['Russian H']	Putins spokesman likens Weinstein accusation	Ordered[RU]		['Hollywood', 'Pskov', 'Duma', 'Russia', 'Duma', 'Russian']										
6	News and notes	['corner', 'y']	News and notes from the Ncaa women's	Ordered[US]		['Louisville', 'Victoria', 'Charlotte', 'Connecticut', 'Ohio', 'Mississippi', 'North Carolina', 'Victoria', 'Victoria', 'Irish', 'American']										
7	Pope in Easter	['Nigerian', 'Pope in Easter Vigil to baptize Nigerian r']	Pope in Easter Vigil to baptize Nigerian r	Ordered[NG]		['Rome', 'Nigerian']										
8	Muscat Film	['Oman', 'C']	Muscat Film Festival ends amid fanfare.	Ordered[OM]		['Muscat', 'Salalah', 'Oman', 'Bollywood', 'Omani']										
9	Barr cites conspiracy	['bondage', 'Barr cites conspiracy theory in support c']	Barr cites conspiracy theory in support c	Ordered[US]		['New York City', 'White House', 'York', 'Democrats', 'Illinois', 'York']										
10	New 'Below Deck	['pageants', 'New 'Below Deck Mediterranean' crew f']	New 'Below Deck Mediterranean' crew f	Ordered[US]		['Oceanside', 'Bellmore']										
11	Cardinal Tagle	['tomb', 'to']	Cardinal Tagle on Easter 2018: Roll away	Ordered[PH]		['Manila', 'Bethany', 'Philippines', 'Filipinos']										
12	Air France, easyJet	['EU', 'aviat']	Air France, easyJet, lag, Lufthansa and R	Ordered[FR]		['Brussels', 'France', 'Belgium', 'Greece', 'Spain', 'Eu', 'Cyprus', 'Italy']										
13	Story Course: A chef's	['Iran', 'sch']	Story Course: A chef's life, in six dishes.	Ordered[US]		['New Orleans', 'New York City', 'South Korea', 'Iran', 'America', 'York', 'York', 'Korean']										
14	Charles Barkley	['schools', 'Charles Barkley lashes out at Ncaa, Fbi a']	Charles Barkley lashes out at Ncaa, Fbi a	Ordered[US]		['San Antonio', 'Louisville', 'Russia', 'Fbi', 'Democrats', 'Republicans', 'Arizona']										
15	Iran's non-oil	['regions', 'Irans non-oil trade with Cis on fall. By Tr']	Irans non-oil trade with Cis on fall. By Tr	Ordered[IR]		['Iran']										

Screenshot from the output of the Country Extractor program run on the GeoWebNews dataset

Output and conclusion

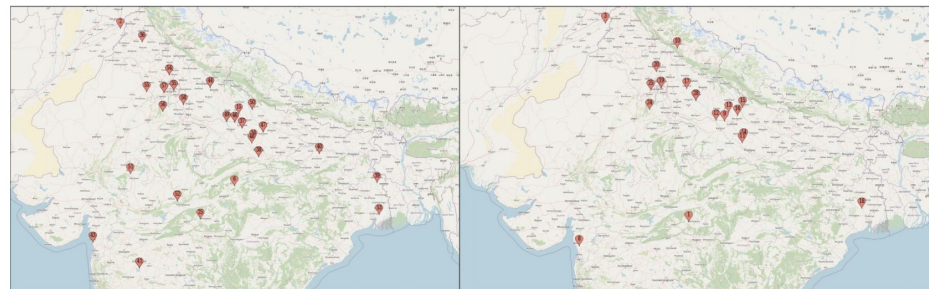
- 99.5% news articles in the 200 GeoWebNews benchmark dataset have 1+ locations in common with the locations found by Country Extractor
- It took approximately 43 seconds to label 200 news articles of length 800+ words with their focus country (running on Google Colab)
- The ability to add and remove countries/cities/states is present in Country Extractor, but not in geotext or spacy. Moreover, Country Extractor improves upon vanilla Flashgeotext in a number of ways, such as by removing names that were identified as cities, adding numerous locations and state abbreviations, etc.

Comparison with locations from [GeoWebNews](#)

Number of common locations	Count	Percentage
0	1	0.5%
≥ 1	199	99.5%
≥ 2	184	92%

Comparison with [spaCy's Part-of-Speech tagging](#)

Map plot showing extracted locations from
Ch. 1 of 'Poetry of Belonging' by Ali Khan Mahmudabad



From Country extractor

From spaCy

Further improvements

- Preprocessing the news article text to remove more typing errors, falsely identified cities/countries
- Adding more cities for each country (this is a pertinent issue)
- Cleaning up the code, writing more helper functions, organizing the notebook, adding comments
- More robustly testing the Country Extractor program by comparing its output to hand-coded focus countries
- Finding ways to collaborate with news agencies, researchers, or anyone else who want to use the Country Extractor program in their operations (currently it is only being used by Koç University)

Thank You