Geolocation Extraction and Clustering

Belinda Welch, Akshay Rawat, Ruchira Nagle, Annesha Chowdhury, Lingyan Wu

*Abstract*—Automatic identification of geolocation mentioned in online news articles provide vital information for understanding associated events.[1]In this paper,we collect articles from Google News API[2] and extract the geolocation from those articles using Geoparser.io[3], lable the location with latitude and longitude coordinates.Then we cluster the articles based on the coordinates.For each cluster,retrieve the critical words and extract the main topic of the cluster.From the above work,we can find out the dominant topic in a specific district and thus have a better understanding of associated events.

Keywords—Google News API; geolocation; coordinate; Geoparser; Geoparsing; location extraction; Geocoding; k-means; gmplot

# Introduction

With the rapid growth in availability of public, free-format text generated through the accelerating use and adoption of internet-connected devices and the subsequent increase in social media usage, there is a great opportunity for researchers and developers to utilize geographical information. Geographical data adds an additional dimension to the richness of data enabling us to know not just the what and when,but also the where of an event.[4]

The paper is organized as follows. First, a short introduction of the tools and terminations we use during the project. Then we present our approach to this project. Finally our experimental results are described in section IV,comparison with previous work are discussed in section V,and conclusion along with future work are presented in section VI.

# Tools and terminations

## Google News API

Google News API is a simple HTTP REST API for searching and retrieving live article from all over the web.We need an API key to use API,which is a unique key that identifies the requests. It’s free for development, open-source, and non-commercial use. [5]

## Geoparsing and Geoparser

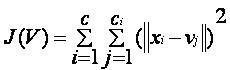
Geoparsing is the process of converting free-text descriptions of places(such as ” twenty miles northeast of Jalalabad ”) into unamibiguous geographic identifiers,such as geographic coordinates expressed latitude-longitude.One can also geoparse location references from other forms of media,for examples audio content in which a speaker mentions a place.With geographic coordinates the features can be mapped and entered into Geographic Information Systems.A geoparser is a piece of software or a (web) service that helps in this process.[6]

## gmplot

It’s an easy way to plot data on Google Maps.A matplotli-like interface to generate the HTML and javascript to render all the data you’d like on top of Google Maps.[7]The gmplot module supports location-based map initialization and zoom-level initialization.

## k-means Clustering

k-means is one of the simplest unsupervised learning algorithm that solve the well known clustering problem.The procedure follows a simple way and easy way to classify a given data set through a certain number of clusters(assume k clusters) fixed apriori.The main idea is to define k centers,one for each cluster.These centers should be placed in a cunning way because of different location causes different result.So,the better choice is to place them as much as possible far away from each other.The next step is to take each point belonging to a given data set and associate it to the nearest center.When no point is pending,the first step is completed and an early group age is done.At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step.After we have these k new centroids,a new binding result of this loop we may notice that the k centers change their location step by step unitl no more changes are done or in other words centers do not move any more.Finally,this algorithm aims at minimizing an objective function know as squared error function given by[8]:

[[](https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm/kmeans.JPG?attredirects=0)](https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm/kmeans.JPG?attredirects=0)

where,

*‘||xi- vj||’* is the Euclidean distance between *xi* and *vj.*

*‘c’* is the number of data points in *ith* cluster.

*‘c’* is the number of cluster centers.

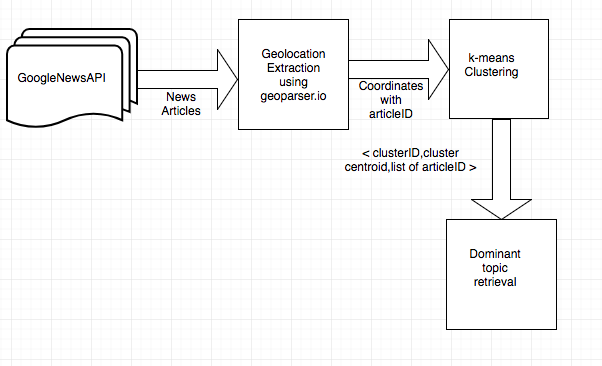
## NLTK Stop Words

A stop word is a commonly used word such as“the”,”a”,”an”,”of”,”we”,“is”,”are”,that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query. We would not want these words taking up space in our database,or taking up valuable processing in time.For this ,we can remove them easily,by storing a list of words that you consider to be stop words.NLTK(Natural Language Toolkit) in python has a list of stopwords stored in 16 different languages.[9]

# Approach

The first challenge we came across is how to implement the location extraction.There are serveral location extraction tools,such as Cliff-Clavin,Geoparser,Stanford CoreNLP,Mordecai,and Edinburgh.After lots of research and comparison,we decided to use Geoparser.io,which is a RESTFUL web API that identifies place names mentioned in text,disambiguates those names and returns GeoJSON for the places found in the text.The locations extracted from articles using Geoparser are labeled with latitude and longitude coordinates.Note that each article may include more than one locations.The locations extracted from the same article are associated with a unique articleID. Due to tight deadline,we decide not to implement focus location extraction.You can find detailed information about focus location extraction in the paper “Focus Location Extraction from Political News Reports with Biased Correction”.

The next challenge is how to cluster those locations.An article might have multiple locations,we handle each location independently.Note that we allow an article assigned to different clusters if this article contains different locations.After clustering,we need to extract the critical words from each cluster and thus retrieve the main topic.We choose to use spark Word Count to obtain the important words.We need to preprocess the data and remove verbs,prepositons,conjunctions from the articles first.How to filter out useless data?We use NLTK stopwords to remove the unwanted words.



**Fig 1.** Flow Chart

# Experimental Results

## Plot the centroid of each cluster using gmplot,when a mouse hover over a point(centroid),it displays the dominant topic of this district.../../../../Screen%20Shot%202017-12-14%20at%2011.27.35%20PM.pn

**Fig 2.** Visualization

# Comparison with other work

In the work of “What’s missing in geographical parsing”,they evaluated and analysed the performance of a number of leading geoparsers on a number of corpora and highlight the challenges in detail.They also published an automatically geotagged Wikepedia corpus to alleviate the dearth of (open source) corpora in this domain.[4]

In the work of “Focus Location Extraction from Political News Reports with Bias Correction”,they have presented and developed a focus location extraction method executable on unstructured text-based new reports.They extract all of the possible locations with a named entity recognition tool,and propose a semantic approach to find the focus location.[1]

In our project,we combined the geoparser technology and “Focus Location Extraction” idea from the prior work, furthermore, we implemented k-mean clustering algorithm on the extracted geolocations.Then we retrieved the main topic from each cluster.We didn’t make an intensive study of the methods or algorithms like the previous work.However, the result of our work-getting to know the dominant topic of a specific district-provides a different perspective to human geography research.

# Conclusion and future work

We have implemented geolocation extraction from unstructured text\_based online news articles.Further more,we cluster the geolocations using k-means algorithm.Lastly,we retrieve the main topic from each cluster using spark Word Count, after preprocessing the articles to filter out useless words using NLTK stopwords and then extracting the top 10 frequent words.

Our future work includes: (i) To extract focus geolocation in new articles in which more than one event is reported; (ii) To use different location extraction tools.

##### Acknowledgment

We greatly appreciate the help from Maryam Bahojb Imani. Some of the terminations and ideas came from her paper “Focus Location Extraction from Political News Reports with Biased Correction”.

##### References

[1] Focus Location Extraction from Political News Reports with Biased Correction.

[2] https://newsapi.org/s/google-news-api

[3] https://geoparser.io/

[4] What’s Missing in Geographical Parsing?

[5] https://newsapi.org/docs

[6] https://en.wikipedia.org/wiki/Geoparsing

[7] <https://github.com/vgm64/gmplot>

[8]https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm

[9] http://www.geeksforgeeks.org/removing-stop-words-nltk-python/