Extracting Jamaican geographic locations using the NLTK and Pattern Toolkit for Python from news articles

Extracting Jamaican geographic locations using the NLTK and Pattern Toolkit for Python from news articles

# Jean-Mark Wright

# Abstract

Natural Language Processing (NLP) has long been used to extract information from large bodies of text. Natural Language Processing is often desirable to intelligently parse large volumes of data where the manual alternative may be infeasible. Using Named Entity Recognition, NLP algorithms can be created to extract the mentions of geographic locations of different types from current and archived news articles. This information can be used to auto plot these locations on the map and provide datasets for deriving how hot a particular location is.

# Title of Paper

Extracting Jamaican geographic locations using the NLTK Toolkit for Python from news articles

# Background

Natural Language Processing (NLP) has long been used to extract information from large bodies of text. Natural Language Processing is often desirable to eliminate the need for manual perusing of such sources. The ability to parse human written information produces many economic and practical gains. Even more compelling is the ability to understand the written text. When algorithms are able to understand written text, they can be used to process large volumes of data and make derivations based on the data. This approach has many merits when compared to a manual approach. With the use of training sets, computers can be taught and can arguably be more consistent, focused and unbiased than their human counterparts. Sentiment extraction and analysis; the process of extracting meaning from natural language, is a fruitful area of research within NLP that has also been heavily researched. Many studies and approaches have been presented to extract polarities from written text. NLP has been used to perform document level sentiment summarization and also producing more granular derivations based on sentences or phrases written in natural language. Such algorithms have been used to power commercial grade systems with very high levels of accuracy and precision ([Jeonghee Yi† [1](#_ENREF_1)]). Information that is derived can be used to support intelligent decision making, influence expert systems, and even trigger early warning systems.

Natural Language Processing has been and continues to show great potential for bridging the gap where there is no established format of communication. Traditionally, parsers and their grammars have dictated context and are constructed to avoid ambiguity. However, using NLP, the computer can parse and understand written text across several languages despite language idiosyncrasies and ambiguity. In a sense, ambiguity can be embraced as a normal occurrence that is expected and does not hinder the capturing of accurate and precise conclusions. Also, using NLP, unformatted natural language text can be even be standardised. [CAROL FRIEDMAN [2](#_ENREF_2)] discusses approaches used to automatically encode clinical documents, mapping them to coding used by the Unified medical Language System (UMLS); a well-known and widely used set of codes used in the medical domain.

NLP despite its obvious usefulness represents a domain that is riddled with many challenges. One major challenge presented is the threat to accuracy and precision caused by the absence of context. Ideally, the context of a body of text needs to be derived in order for any meaningful information derivation. This arguably is also necessary for less ambitious approaches that aim just to search natural language for occurrences of a word or phrase. Under different contexts, a single word may have several meanings, requiring that algorithms address ambiguity in documents. To understand exactly what is being described or referenced, a context is often necessary. Obtaining a proper context however still remains an open and challenging problem. Many researchers have avoided this route altogether and instead relied on purely statistical methods instead of attempting to derive a proper context from the text. To understand even a simple sentence often requires knowledge that may or may not be present in the sentence. It follows that to understand a larger article; much of the context required for the article may not be in the article. Also, even when sufficient information exists in an article, it may not be contiguous, connected or structured in a way that makes for simple construction of context using algorithmic approaches. Different methodologies have been employed in research to either directly tackle this problem or avoid dependence on context by using approaches that rely on training sets or other statistical methods. A direct approach to this problem was proposed in [Jeonghee Yi† [1](#_ENREF_1)]. They used a training set and dictionary of sentiments to aid in the understanding of the context. One indirect approach is that of IBM’s Nominator ([Ravin [3](#_ENREF_3)]). Nominator used aggregation of candidate terms across several documents to replace the need for developing a context. Both approaches saw high levels of precision and accuracy.

Named Entity Recognition, is a particular branch of NLP that is also tasked with the challenge of disambiguating information to extract reliable information. Named Entity Recognition is defined as the process of identifying specific names mentioned in a body of text. In general, the three categories recognized are names of persons, organisations and geographic locations. While George Bush and Kofi Annan may immediately be identified as familiar persons in history, it was also noted that Bush is a city of Louisiana and Annan is the name of a small city in Scotland. This example gives a clear picture of the kind of ambiguity that may potentially thwart accuracy and precision. Some semblance of context or alternative approach is needed to understand exactly what is being referenced. Clashes can also occur between the names of places, persons or organisations and other common words in spoken text. One such prime example; Welcome, is the name of a town in St. James and She is the name of a city in India.

Named Entity Recognition however is an area also ripe with potential. There is a direct application of NER for geocoding; news stories, articles and other user produced text can be geo-tagged. With NER algorithms, very large archives and current issues of news item can be parsed relatively quickly. The information garnered from the process can be used to track mentions of locations as time progresses. Despite the many challenges Natural Language Processing has seen many successes; much of the work reviewed has achieved remarkable rates of precision and accuracy. The measure of accuracy and precision are often described using the F-measure. This measure was used by researchers when measuring their results. The results of the research thus far in NLP present a very strong case for its tangible relevance and usefulness as a technology.

# Research Aims

In general, the aim of this research is to combine techniques used in previous works in order to extract the names of Jamaican geographic features from news articles. We will assume that the article may refer to multiple places simultaneously. These references will be extracted and geo-tagged to the title of the article. In order to accomplish this, the following steps need to be accomplished:

* Build a corpus that contains confirmed references to geographic locations in Jamaica. This corpus will be built with articles from the Jamaica Gleaner, Jamaica Observer, Jamaica Information Services and Discover Jamaica. Requisite tools will be built to extract information from these sources in text. This information will be used to build a training set that can be used to train algorithms to recognize references to Jamaican geographic locations.
* Build a gazetteer for places in Jamaica. No gazetteers are currently available that contains a comprehensive set of places in Jamaica. Thus a gazetteer will be built to complement international gazetteers used. Information from the Global Discovery gazetteer will be used to provide base information. All items in the gazetteer will be given a class and a weighting. Countries in the gazetteer will be given the lowest class and highest weighting. An actual feature will have the highest class but the lowest weighting.
* Develop an algorithm that can perform unsupervised recognition of geographic locations in Jamaica, achieving F-measure of 0.80 or higher. This solution will be developed using the Python programming language and it’s Natural Language Toolkit (NLTK).

# Literature Review

## The Perseus Digital Library

[David A. Smith [4](#_ENREF_4)] set out to disambiguate place names in Historical Digital Library. In their research they described the Perseus digital library and evaluated its performance. In their research they highlighted many of the practical reasons for wanting to utilize natural language techniques. In particular they highlighted the infeasibility of the manual tagging of vast libraries of geographical data. Automated geocoding is a much preferred approach. The process of geocoding mainly involves two broad steps: identification and categorization and also disambiguation. Primarily, they noted that the wide range of data directly affects the process of disambiguation. A gazetteer is used to disambiguate the results found. It was noted that systems whose task is to match entities such as personal names or company names can easily operate without a list of known names. However, for the process of identifying geographic locations, it is necessary to obtain an index such as a gazetteer that supplies names and location based information for each place.

In their review, they spoke of several systems built. BBN's Nymble has a 0.9 score for F-measure. It however requires a large amount of training data. IBM's nominator came close the Nymble with a score of 0.88. Another system reviewed was the Geo-Referenced Information Processing Systems (GIPS). GIPS identifies geographic locations without performing disambiguation. It relies rather on the process of aggregation of the information over several articles to filter out ambiguity. “Reducing ambiguity” was also highlighted as causing a problem in entity recognition. Initially a document may say "President Bill Clinton" but there after refer to the President as "Clinton" or "Mr. Clinton". This presents a problem for name recognition. They also highlighted the fact the names of persons may often clash with the names of geographic locations. Approaches that ignore timelines also present a challenge on the relevance and accuracy of the information. It was noted that across the different continents, there are several places that have multiple names, and also the percentage of names that clash with places names were high in certain regions. It was found that the percentage of name clashes were as high as 57.1% in North and Central America.

To perform the procedure of disambiguation, internal and external sources are considered. The external sources used were gazetteers, bibliographic information and general linguistic knowledge. Internally, honorifics, the linguistic environment and geographic labels are used to aid in the disambiguation process. A listing of over one million names was constructed for the purpose of disambiguation and name identification. Perseus uses the rule of thumb that candidate proper names are generally capitalized. Perseus uses the existence of honorifics as strong indicators of the presence of a name i.e. the presence of topographic labels e.g. mountain or river were used as a strong indicators of the presence of a geographic feature.

The first step in their procedure for identifying and disambiguating place names was to search for all named entities and tag them name, place or date. Their system next looked up all entries tagged as geographic names in the gazetteer. Uncertain names are also matched. A context for the document was determined by the major influence of the dominating or repeated geographic name(s). A score is also calculated for each result based on the proximity to other locations and relative importance (parishes have a higher importance than a city or town). The authors calculated the "centroid" of the article based on the different locations mentioned. They then calculated the standard deviation of the distances from the centroid. Points that were found to be two or more standard deviations away from the centroid were discarded then a new centroid calculated. After the centroid is created, a context consisting of the four geographic references before and after a reference is created. Each geographic reference is now given a score based on its proximity to the geographic references in its context, distance from the new centroid and also the relative importance. The candidate geographic reference with the highest score at the end is deemed the winner.

When tested using standard precision and recall means, the system scored very well, getting as high as 0.93 in precision for Greek places and a high of 1.00 in recall for Roman names. The F-measure was highest for Greek names, with a value of 0.96. Scores obtained for London, California and the Upper Midwest were lower. This was attributed to the duplicity of names and places. The author noted that proper names are not generally inflected in English, and hence it was deemed that the use of stemming would do more harm than good for the research. More information could have published about the details of the testing procedures. The authors did explain that their tests were biased toward system recall and that corpuses were used for different languages for the evaluation of the system and that humans were also used to manually work through 20% of the corpus of each language. They also provided analysis for the results obtained.

## Nominator

[Ravin [3](#_ENREF_3)] in their work on Extracting names from Natural Language Text, aimed to supplement the technologies used at IBM. They built a system called Nominator for this purpose. Nominator identifies proper names in text, classifying them as place, name or organization. It was built to deal mostly with data generated from news items.

When names are identified in a body of text, each group is assigned a canonical name. The canonical name is chosen to be the least ambiguous and most permanent name. The canonical name serves for a grouping of names. Hence, many words or phrases may refer to a single name. Nominator uses 1600 tags and other name indicators (e.g. Mr., Mrs.), 700 most common places names and approximately 20,000 first names. Nominator also relies of capitalization, punctuation and on contextualization to adequately classify names. This assumption they based on the conventions normally observed in the texts that are properly edited. Nominator also assumes that most names are referred to first in its most explicit sense then are referred to using shorter forms. An important point made in their research was that names often are phrases that cannot be separated; they must be preserved to keep their meaning.

In their review, very little was mentioned of previous research. They however did briefly mention the StenStep application. StenStep performs tokenization of its input. This process is however minimal.

In order to extract names from a document, Nominator performed the following steps:

* Parse the document for a list of candidate names. The document presented was first searched to generate a list of potential candidate names.
* Split the names into smaller names. After parsing, a second perusal the names are compared to each other. In this phase, some names are discarded and others are edited or adjusted. This second parsing is used to narrow the list.
* Group the names into equivalence classes. The list is then divided into groups of equivalent names and a canonical name is selected for each group. Nominator produces a database of names which are each related to a collection of documents.
* Aggregate the classes across multiple documents. After all the documents have been analyzed and canonical names have been selected for each group, groups with the same canonical name are merged. When the aggregation is done, the stronger canonical name always overrides the weaker. This aggregation is used as an indirect approach to combat ambiguity in the document.

IBM's approach in building Nominator focused less on intensive Natural Language Processing techniques and more on simpler operations based on characters and strings. Nominator also does not perform intensive analysis of contexts. In selecting this approach, the authors were able to gain much more speed in processing but limited Nominator's ability to understand what is being discussed or directly tackle ambiguity. The method used instead to tackle ambiguity was to aggregate the equivalence classes across many documents. By taking this approach, the aggregation helps to remove the noise presented by ambiguities.

Nominator was tested using 88 articles in the Wall Street Journal. Proper names were first manually identified from the articles, and then Nominator was given the text to identify proper names. In testing the Nominator performed very well, achieving a high recall of 97.8% and a precision of 91.8%.

The researchers from IBM presented a solid case that was well tested and very well documented. Even though, the approach negated the use of a direct context, the use of aggregation and equivalence classes proved to be a useful and remarkable approach. This was evidenced in the high scores of recall and precision received in the tests performed.

## Geocoding across multiple languages

[Bruno Pouliquen and Blackler [5](#_ENREF_5)] aimed to pick up references to geographic place names in multiple languages. They aimed not only to state that a name is a place, but also to disambiguate homographic place names. One of the primary approaches taken by the authors here was to only tag a place as such after is has been confirmed that the place is not instead the name of a person. Place name identification is done with the help of a gazetteer. The following sources were used to form the gazetteer:

* Global Discovery 2006 – this document contains over 500,000 place names world-wide. This also contains a class attribute that assigns a weighting to all the places.
* KNAB database
* European commission internal document – this document stores the names of capitals, inhabitant names, currency and country adjectives used.

After places are identified in the gazetteer, the process of disambiguation is run.

The process of disambiguation takes into consideration that a place name may clash with the name of a person. The disambiguation process also considers the importance of the place (nation as opposed to city), the main country or countries that the text is about, the presence or absence of geo-stop words, the words in the surrounding lexical contextand the minimum distance to any other places. After person names have been recognized in the text, those names are not considered as candidates for geographic locations. A weighting was assigned to geographic names in the Global Discovery. Countries, country capitals and main cities were all given a high rating of 80. Province level or small cities were given weightings of 30 and 20 respectively.

Shallow geoparsing is first used to identify the highest classes of geographic names in the document. Then deep geoparsing is used to recognize lower class geographic names only if they exist within one of the names previously identified while performing shallow geoparsing. Therefore cities are only retained in the deep geoparsing process if its country or some geometric place which contains it had previously been identified during shallow processing.

A geo stop word tool was also built to help with the ambiguity caused by words being homonymic with each other. The stop word list uses two approaches. The first analyses a list of first names and tags those that are potential names of geographic locations to be stop words. The second approach uses a corpus to identify those homonymic words that occur above a certain frequency and tags them as stop words.

Use of the lexical context was initially used (e.g. using "city" to indicate the existence of a place name), however the authors decided to forgo this approach due to observed negligible benefits and the recognition that POS tagging and other language dependent rules would be required to obtain a more impacting outcome.

The minimum kilometric distance was also used to disambiguate place names. The distance was measured from non-ambiguous places to ambiguous places. This process helped to better target ambiguous references to geographic names.

The authors built their own test set of articles in multiple languages due to the absence of a proper corpus. 162 newspaper stories were selected that were in different languages. The system built had a high precision score of 91 for English with the lowest score for precision being 68 for French. Italian had the highest score for recall, that of 85. German had a low of 68. The highest and lowest scores of F-measure were recorded for English and French with scores of 84 and 70 respectively.

## Sentiment Analysis

[Jeonghee Yi† [1](#_ENREF_1)] sought to use a more direct approach to parsing natural language. Most of the approaches taken attempt to summarize the polarity of the document instead of finding polarity of individual sentiments. That approach possesses an inherit weakness since such summarized polarities fail to capture individual specifics that would often be useful for users of the information. The association of a sentiment to a topic has also been recognized as a difficult problem. At the time of their research, none of the work reviewed had been ground breaking in that regard. Most of these methods have either depended on a classifier to classify the document or used statistical methods based on terms used in the same context.

Their approach primarily focuses on a topic specific feature term for extraction. Processes of sentiment extraction and subject-sentiment association are performed. The approach uses a sentiment lexicon and a sentiment pattern database. A subject can be either a topic of interest or a feature of the topic. For example; a “camera” may be the topic of interest and “battery” may be a feature of the topic.

The Sentiment Analyzer they designed extracts topic specific features, sentiments of each phrase that contains a sentiment and finally makes relationships between the topic and the sentiment.

A feature term was defined as a term that is either:

* A part of the topic (e.g. the battery of a camera)
* An attribute of the topic (e.g. the cost of the camera)
* An attribute of a known feature of the topic (e.g. the life of the camera’s battery)

Authors used a mixture model and likelihood test methods to extract features from the articles. The likelihood tests outscored the mixture model tests. The authors built their sentiment word collection using data from General Inquirer, Dictionary of Affect of Language and WordNet. An observation was made that feature terms are nouns. Hence, they proceeded to extract noun phrases from document. The researchers properly outlined the research process and the components of the sentiment analysis algorithm.

For testing, the authors used a corpus that was created from different web pages containing reviews about cameras. Half of the cases were predetermined to be favourable. Over 2000 reviews about cameras were used to test the system for commercial grade applications. Even though only one subject was selected for testing, the authors sought to explain failures or present justifications for unexpected results that were achieved. In their results, the system was able to achieve a high precision of 95% with a low recall of 24%.

# Research Design and Methodology

In order to extract the names of Jamaican geographic locations from natural language, a solution will be designed using the Natural Language toolkit for Python. Python is an interpreted that has a gentle learning curve and has excellent learning resources ([Madnani [6](#_ENREF_6)]). Python is also very modular and has a very strong set of built-in modules and other third party plugins. There are many modules available to process geographic datasets. The Natural Language Toolkit for Python is a well-documented module that houses many techniques for doing Natural Language analysis. The online tutorials are very well written and provide a very solid base for doing simple and more advanced tasks using NLP. Researchers have also written several reviews that provide gentle introductions to beginning work with NLTK. The NLTK toolkit for Python has been selected in order to develop an algorithm that will extract Jamaican geographic locations from unstructured text. The Python language is also a familiar domain and thus represents a natural choice for the development of the solution.

## Building a corpus

All research reviewed has underlined the importance of training the system properly so that context rules may be derived and the system may be relevant for the domain of text it will process. In particular [David A. Smith [4](#_ENREF_4)] notes that proper training is necessary for the derivation of context rules. Currently, there is no known corpus that contains confirmed references to geographic locations in Jamaica. Therefore, the first step in the research will be to construct a corpus that will contain such references. A local corpus is needed so that oddities and differences for the Jamaican context can be properly understand and integrated into the algorithm for recognizing places. Primarily, a Python script will be built to extract articles from the following websites:

* Jamaica Gleaner
* Jamaica Gleaner Archives

The aim is to obtain enough articles that will contain up to 1000 references to different geographic names across the island. These geographic features are expected to cover parishes and communities. After the information is downloaded using a Python script, each of these articles will be manually geo-tagged. This corpus will be used to train the system. Statistical methods will be used to store the different variations used in referencing these places. Careful attention will be placed on honorifics, capitalization or other cases used in references, name declensions and aliases.

## Building a gazetteer

Approaches have been presented that do not use gazetteers for the process of NER ([David A. Smith [4](#_ENREF_4" \o "David A. Smith, 2001 #7)]), however considering the results of the previous research and the abundance of methods that favoured the use of a gazetteer we have decided to build a gazetteer for the purpose of name recognition. Unfortunately, only international gazetteers currently exist, thus we will use the Global discovery gazetteer and extend it to include geographic locations in Jamaica. A proper source of information for such has not yet been secured. It is expected that a considerable amount of manual effort will need to be placed on manually obtaining lists of the different features in each category then geocoding these references. If we find that the information is not available, the research will be confined to the list of categories that are readily available.

All items stored in our gazetteer will be assigned a class and weightings assigned appropriately. The approach taken here will mirror the approach in [Bruno Pouliquen and Blackler [5](#_ENREF_5)] in assigned classes and weights. We expect to assign classes and weights as dictated below (lowest class first).

|  |  |  |
| --- | --- | --- |
| **Geographic Name Category** | **Class** | **Weighting** |
| Country | 0 | 80 |
| Parish | 1 | 30 |
| Community | 2 | 10 |

**Table 1:** Geographic name categories with class and weight

## Develop an algorithm

An algorithm will be built using the Python NLTK module for the extracting of geographic locations. The algorithm will first seek to construct a list of the lowest classes found in the document which will be deemed the domain(s) of the article. Since the domain of articles is generally expected to be in Jamaica, we expect that actual references to Jamaica may not be present in the article. Hence, the algorithm will start with the lowest class and keep incrementing the class until the domain of the document is determined. We also expect that there may be several references to geographic places sharing the same domain. Hence, the algorithm will start by searching for all references to class zero (0) geographic places, incrementing the class each time no result is found.

Once a domain or list of domains is constructed for the document, the algorithm will continue to search the document for references to higher class geographic names in the document. For each result obtained, a weighting will be calculated for the reference. The weighting will be calculated based on the class weighting, weighting based on statistical similarity to usages in training set and mean geometric distance to the domains and its geographic containment by a domain of the document. A centroid will then be calculated for the references in the document. All geometric features that are more than two standard deviations away from the centroid will be discarded. A new centroid will then be calculated from the remaining set of features.

The algorithm will then loop over each of the non-domain features, keeping a context of up to four geographic references before and after the current one. It will check the corpus to compare the feature’s context with contexts for the same feature in the corpus. A weighting will be assigned based on the similarity to the corpus. After each feature contains its final weighting different heuristics will be applied to determine what should be eliminated.

## Evaluation

For the purposes of evaluation, articles will be selected at random from the Jamaica Gleaner and Jamaica Observer websites (once they are not already contained in the training set). The python solution will be used to properly recognize and classify up to 500 references to geographic names with an F-measure of higher than 0.8. Tests will be done first with a threshold set for the weighting to see what values will be obtained for precision and accuracy. Then we will remove the threshold to see what values of precision and accuracy are required to produce an F-measure of 0.8 or above. Finally, the different weightings will be adjusted to see what effect this will have on the accuracy, precision and by extension, the F-measure.

# References

1. Jeonghee Yi†, T.N., Razvan Bunescu, Wayne Niblack†, *Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques.* Proceedings of the Third IEEE International Conference on Data Mining, 2003.

2. CAROL FRIEDMAN, P., LYUDMILA SHAGINA, MS, YVES LUSSIER, MD, GEORGE HRIPCSAK, MD, MS, *Automated Encoding of Clinical Documents Based on Natural Language Processing.* J Am Med Inform Assoc, 2004. **11**: p. 392-402.

3. Ravin, Y., *Extracting Names from Natural-Language Text.* 1997.

4. David A. Smith, G.C., *Disambiguating Geographic Names in a Historical Digital Library.* P. Constantopoulos and I.T. Sølvberg (Eds.), 2001: p. 127-136.

5. Bruno Pouliquen, M.K., Ralf Steinberger, Camelia Ignat, Tamara Oellinger, Ken and F.F. Blackler, Wajdi Zaghouani, Anna Widiger, Ann-Charlotte Forslund, Clive Best, *Geocoding Multilingual Texts: Recognition, Disambiguation and Visualisation.* In Proceedings of LREC-2006, 2006.

6. Madnani, N., *Getting Started on Natural Language Processing with Python.* ACM Magazine, 2007. **13**(4): p. 5-5.