A geographic data sources classifier using a supervised learning model

José Augusto Sapienza Ramos
COPPE - Programa de Engenharia de Sistemas e
Computação da Universidade Federal do Rio de Janeiro
Rio de Janeiro, Rio de Janeiro, Brazil 21945-970
sapienza@cos.ufrj.br

***Abstract*—This work presents a classifier of plain text documents that describe geographic data sources such as web pages, blogs, encyclopedia articles, news stories, tweets and travel reports, among other textual contents. Unlike other proposals, this work does not analyze the data source directly or search by geographic references in text, but applies a Support Vector Machine and a classified geographic data source list. Several tests are performed, where the accuracy rate of 93.2% using 128 more relevant terms to train a SVM. Aside from that, other results show that some improvements may be necessary.**

Keywords—geographic information retrieval; web documents;

I. INTRODUCTION

Throughout the history of man, some techniques and technologies have greatly increased the capacity of storage and production of data such as writing, the Gutenberg Press, computers and the Internet. On the other hand, the new production scale increases the challenge of organizing and sharing the available data across their users.

Since the data infrastructures are essential tools in the production of knowledge, governance and capital in modern societies [1], a great effort has been employed for increase collectable, discoverable, available, accessible, and interchangeably data. The aim is that the information should be easily retrieved in order to be useful [2], being certainly the Internet an important and big repository of raw and essentially unavailable data without a proper formatting.

In special case, Geo-referenced data acquisition and sharing are now easier and more widespread than ever, helping us map and understand 500 million square kilometers of the Earth's surface. Geographic data can be defined as the way in which the knowledge of the Earth's surface is represented [3] – from more traditional formats such as maps and globes to texts, tables and other kinds of formats. Advancements in technology as well as the new paradigm in geographical information science and systems also allow the production, visualization, analysis and sharing of large amounts of geographic data suitable to different kinds of audiences. In other words, the geographic data production and its usage have been decentralized from governments and companies to include the population in general, giving birth to terms such as "neocartographers" [4] and "Volunteered Geographic Information" (VGI) [5].

Especially with the advent of the Internet, a considerable amount of information is not in structured databases, but in format semi-structured or unstructured – some works suggests that unstructured data are growing 15 times the rate of structure data [6]. This way, a relevant part of digital geographic information produced or referenced lies in home pages, news, social media and others kinds of semi-structured or unstructured data, where its collection and analysis demands the application of specific strategies.

In this context, several works are debating the Geographic Information Retrieval (GIR) in recent years – see [7,8] for a survey. For instance, works as [9,10,11] propose methods to detect geographic locations in texts, while works as [12,13] perform search engines spatially aware.

This work is more closely related to geographic data mining proposal as [14], where metadata are extracted analyzing linked technical documents or XML records, for instance. However, to the best of our knowledge, there is no other proposal that detects web documents which relate about a geographic data sources – but not is the data source or spatial references directly.

II. OBJETIVE

This paper applies Support Vector Machine, i.e, a supervised machine learning technique, to implement a classifier which returns whether a plain text document describes a geographic data source. At end, a straightforward implementation in Python 2.7 language is applied as proof of concept, thus some results are shown and debated.

III. MATERIALS AND METODOLOGY

A. Preparing corpus for training and test

Assuming that there is a corpus with documents classified as geographic data source (GDS) or non-geographic-data-source (~GDS), the following steps define the initial and preparing procedure:

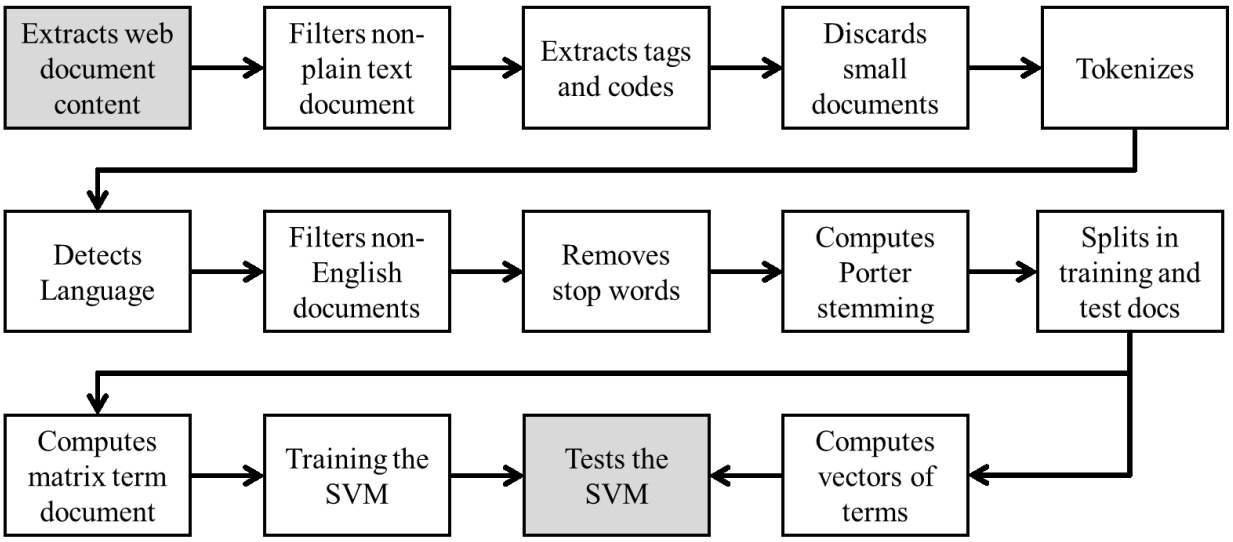

Fig. 1. A flow to represent the methodology's steps of this works. The gray boxes are the beginning and the end of flow.

1. To extract each document content, discarding non-plain text documents (e.g. PDF, JPG or DOC formats);
2. To rip out the XML or HTML tags scripts and CSS styles.
3. To discard documents with less than 50 words because they have little content to be analyzed in classification.
4. To tokenize using unigrams.
5. To apply the language detection algorithm based in stop words analysis as described in [15], filtering documents only in desire language.
6. To remove stop words in all documents.
7. To compute the Porter stemming algorithm [16] for each unigram token (i.e. each word) in each document.
8. To split the corpus in training set and test set, respectively 80% and 20% of probability in a random draw.

At end, we compute two Bags of Words (BoW) with GPS and ~GPS documents: one for training (Ψ) and other for test (Φ).

B. Computing the term document matrix

In this second step, Ψ are analyzed in order to compute the term document matrix (M_{td}) using as weight given by (1) – the term frequency divided by the inverse document frequency – see [2] for details. In this approach, each row i and column j in M_{td} represent respectively one term and one document, where a cell (i,j) gives a measure that regards the importance (or weight) of the term within the document as the term frequency in it [17] normalized by a measure of specificity of term to differentiate its document from others in corpus as the inverse document frequency [18]. This work prefers the following formula to compute *tf/idf*.

$$w_{td} = f_{td} \cdot \log(N/n_t) \quad (1)$$

where w_{td} is weight of the term t in the document d, f_{td} is the frequency of the term t in the document d, N is the total number of documents in corpus, and n_t is the number of documents where the term t appears.

In its turns, a vector $v_d = \{w_{td} \mid t \in M_{td}\}$ of terms in M_{td} is computed to each document d in Φ, where the weight is given as a binary value: $v_{td} = \{1, \text{if } t \in d; 0, \text{otherwise}\}$.

C. Training and testing the Support Vector Machine

The final step is to train and validate a linear SVM using M_{td}, where the document classification (GDS or ~GDS) given as label to classifier.

Since each Bag of Words can easily have thousands of terms, each term is a dimension in SVM's analysis, and only few terms have a relevant tf/idf for the classification, it may be important to reduce the number of terms applying a measure of relevance. To that end, this work adds the td/idf values for each term in all documents, i.e, sums each row of M_{td}, thus selecting the k greater values.

Thus, each vector of terms is classified by SVM, where the measured label (GPS or ~GPS) is compared with predicted label returned by the SVM, computing the hits and the misses. The final accuracy measure *accur* is given by (2).

$$\text{accur} = \text{hits} / (\text{hits} + \text{misses}) \quad (1)$$

IV. TESTS AND RESULTS

A straightforward implementation was written in Python 2.7 language as proof of concept and to compute results. The following modules are applied to implement the main features:

- bs4 or Beautiful Soup in version 4.4.1 to rip out tags, scripts or css styles from contents of plain text documents.
- numpy 1.11.0 to perform representation and operation with vectors and matrixes.
- sklearn 0.17 to implement the SVMs.
- nltk 3.1 to tokenize, list stop words and execute the Porter stemming in English and in Portuguese.

A GitHub's site was created to host the source code [19]. See it for more details about our implementation.

The Labgis System – Center of Geotecnologies of State University of Rio de Janeiro has a site [20] with links from pages that describe geographic data sources (GDS) in several languages. This dataset was classified by a group of specialists and has continuous feedback by visitors to inform new data sources or broken links, for instance. This list with 579 links to documents is provided kindly to this work, where are selected documents in English.

In order to collect the same number of documents that are not geographic data sources (~GDS), random links are extracted from site called URouLette Links [21]. It is noted that a link randomly collected can be a geographic data source, however this work assumes that this probability is small enough to be disregarded. In other words, these unclassified documents are classified as ~GDS.

This way, a corpus and two Bags of Words (BoW) was created, namely: BoW for training Ψ and for test Φ. Some statistics about these Bags of Words are shown in Table 1.

Thus a linear SVM was trained and tested using Ψ and Φ. The tests perform the SVM several times varying the value of k most relevant terms: k = $\{2^0, 2^1, 2^2, \ldots, 2^{12}, 2^{13}\}$ – see the accuracy results in Fig. 2.

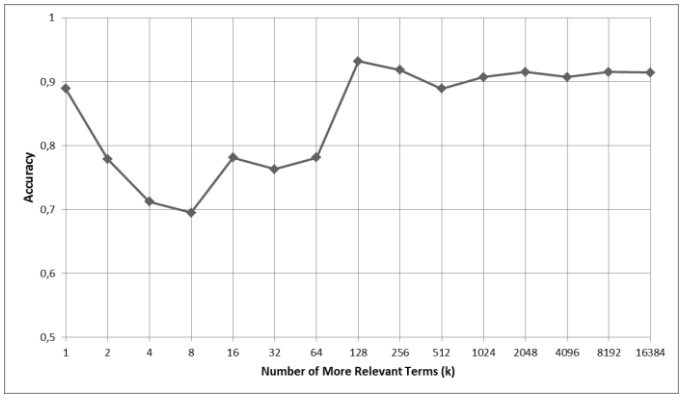

Fig. 1. The graph plots accuracy versus k more relevant terms, where it is noted that accuracy peak occurs with 128 terms.

The Table 1 shows a low rate of repeated terms, resulting in a low *tf/idf* overall. However, the accuracy rate is maintained high and relatively stable from 128 terms. Fig. 1 plots in x-axis only up 16384 terms because Φ has a limit 18316 unique terms, thus the next step exceed it.

TABLE I. STATISTICS ABOUT THE BAGS OF WORDS

BoW	Table Column Head					
	Number of Documents			*Number of Terms*		
	GDS	*~GDS*	*Total*	*Total*	*Unique*	*% Uniqueness*
Ψ	226	208	434	507391	353533	69.8%
Φ	50	68	118	25548	18316	71.9%

Aside from that, the list of the 20 most relevant terms in descendent sort showed an important characteristic, namely: data, map, download, usg, student, state, file, gi, global, univers, februari, dataset, read, servic, offic, program, nation, research, world, and land. Some terms such as map, gi and land are proper to a document that describes a GDS. On the other hand, other terms such as data, download, file and program also can be associated with other types of data sources as, for instance, tabular data.

V. CONCLUSIONS

This work applies Support Vector Machines in order to classify plain text web documents as refers to a geographic data source (GDS) or not. A classified dataset with GDS and random linked represented non-GDS was extracted and analyzed to training and test a linear SVM.

In despite of good accuracy rate using several numbers of terms, the list of more relevant terms shows terms which can be linked to data sources with other kind of data. For example, we reach almost 90% of accuracy using only the term "data" in SVM.

As future work, the SVM may compute the classification in two stages. At first the classifier returns if the document refers to a data source or not. If positive, thus it classifies if the document refers to a geographic data source or other kind of data source. Other potential directions can be test with others n-grams, others languages, other corpus, among other parameters and technical concerning a best known about their importance in final classification.

Along with a crawler, this classifier can continuously scan the web by searching for geographic data sources to feed lists or even search engines spatially aware.

REFERENCES

[1] R. Kitchin, The data revolution: Big data, open data, data infrastructures and their consequences. Sage, 2014.

[2] C. D. Manning, R. Srabhakar, and S. Hinrich. "Introduction to information retrieval." An Introduction To Information Retrieval 151 (2008): 177.

[3] M. F. Goodchild. "Challenges in geographical information science." *Proceedings of the Royal Society of London: Mathematical, Physical*

and Engineering Sciences. Vol. 467. No. 2133. The Royal Society, 2011.

[4] S. B. Liu., and L. Palen. "The new cartographers: Crisis map mashups and the emergence of neogeographic practice." Cartography and Geographic Information Science 37.1 (2010): p. 69-90.

[5] N. R. Budhathoki, and Z. Nedovic-Budic. "Reconceptualizing the role of the user of spatial data infrastructure." GeoJournal 72.3-4 (2008): 149-160.

[6] P. C. Zikopoulos, et al. "Understanding big data." New York et al: McGraw-Hill 5.8 (2012).

[7] C. B. Jones, and R. S. Purves. "Geographical information retrieval." International Journal of Geographical Information Science 22.3 (2008): 219-228.

[8] J. L. Leidner, and M. D. Lieberman. "Detecting geographical references in the form of place names and associated spatial natural language." ACM SIGSPATIAL Special (2011): 5-11.

[9] J. Kim, M. Vasardani, and S. Winter. "Similarity matching for integrating spatial information extracted from place descriptions." International Journal of Geographical Information Science 31.1 (2017): 56-80.

[10] C. Wang, et al. "Detecting geographic locations from web resources." Proceedings of the 2005 workshop on Geographic information retrieval. ACM, 2005.

[11] M. D. Lieberman, H. Samet, and J. Sankaranarayanan. "Geotagging with local lexicons to build indexes for textually-specified spatial data." Data Engineering (ICDE), 2010 IEEE 26th International Conference on. IEEE, 2010.

[12] R. S. Purves, et al. "The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet." International journal of geographical information science 21.7 (2007): 717-745.

[13] L. Backstrom, et al. "Spatial variation in search engine queries." Proceedings of the 17th international conference on World Wide Web. ACM, 2008.

[14] H. J. Miller, and J. Han, eds. Geographic data mining and knowledge discovery. CRC Press, 2009.

[15] N. Hardeniya, J. Perkins, D. Chopra, N. Joshi, I. Mathur. Natural Language Processing: Python and NLTK. Packt Publishing Ltd, 2016. p. 508.

[16] M. F. Porter. "An algorithm for suffix stripping." Program 14.3 (1980): 130-137.

[17] Luhn, Hans Peter (1957). "A Statistical Approach to Mechanized Encoding and Searching of Literary Information". IBM Journal of research and development. IBM. 1 (4): 315. doi:10.1147/rd.14.0309. Retrieved 2 March 2015.

[18] Spärck Jones, K. (1972). "A Statistical Interpretation of Term Specificity and Its Application in Retrieval". Journal of Documentation. 28: 11–21. doi:10.1108/eb026526.

[19] Github. Trabalho Final da disciplina de BRI 2012 2 do PESC/COPPE/UFRJ. https://github.com/jasapienza/. Last access: Set/2017.

[20] Sistema Labgis/UERJ. Lista de fontes de dados geográficos. Link: https://www.labgis.uerj.br/fontes_dados.php . Last access: Set/2017.

[21] World Readable. UROULETTE: Random Link Generator. http://www.uroulette.com/ . Last access: Set/2017.