Article **Journal-Economic History**
December, 2021 Vol.5 No.9 1-7

Análisis de datos utilizando web scraping para repertorio otomí educativo en dispositivos móviles android

Data analysis using web scraping for educational otomi repertoire on android mobile devices

LOPEZ-GONZALES, Erika*†, ALEJO, Roberto, ANTONIO-VELAZQUEZ, J and AMBRIZ-POLO, J.

ID 1st Author: *Erika, Lopez-Gonzales*

ID 1st Coauthor: *Roberto, Alejo*

ID 2nd Coauthor: *J, Antonio-Velazquez*

ID 3rd Coauthor: *J, Ambriz-Polo*

DOI: 10.35429/JEH.2021.9.5.1.7 Received September 14, 2021; Accepted December 18, 2021

Abstract

Around the world there are about six thousand languages, among the nations most threatened languages Mexico is one of the first places according to the Atlas of Endangered Languages in the World Organization of the United Nations Educational Scientific and Cultural Organization. In the State of Mexico, INEGI registered a total of 97,820 Otomi language speakers, most of whom live in the etnorregión. However one of the current social situations in the country and particularly in the State of Mexico is the loss of identity by new generations on their roots, customs, traditions and culture the interest of young people to preserve this language it is almost nil. On the other hand mobile technology it is becoming a revolution in our society. The use of scraping and data collection will provide the information that is available on the web, facilitating the search for words and integrating an application, resulting in a more efficient translation to support the use, teaching and learning of the Otomi language teens / young Otomi communities north of the State of Mexico.

Otomí, mobile, scraping, analysis

Resumen

Alrededor del mundo existen aproximadamente seis mil lenguas, entre las naciones con más lenguas amenazadas México ocupa uno de los primeros lugares según el Atlas de las Lenguas en Peligro en el Mundo por la Organización de Las Naciones Unidas para la Educación la Ciencia y la Cultura. En el Estado de México, el INEGI registra un total de 97 820 hablantes de lengua otomí, que en su mayoría habitan en la etnorregión. Sin embargo una de las situaciones sociales actuales en el país y particularmente en el Estado de México es la pérdida de identidad por parte de las nuevas generaciones relativas a sus raíces, costumbres, tradiciones y cultura el interés por parte de los jóvenes por conservar dicho lenguaje es casi nulo. Por otro lado la tecnología móvil se está convirtiendo en una revolución dentro de nuestra sociedad. El uso de scraping como recolección de datos suministrará la información que se encuentra disponible en la web, facilitando la búsqueda de palabras e integrándola a una aplicación, dando como resultado una traducción más eficiente para contribuir al uso, enseñanza y aprendizaje de la lengua otomí en los adolescentes/jóvenes de las comunidades otomíes al norte del Estado de México.

Otomí, móvil, scraping, análisis

Citation: LOPEZ-GONZALES, Erika, ALEJO, Roberto, ANTONIO-VELAZQUEZ, J and AMBRIZ-POLO, J. Análisis de datos utilizando web scraping para repertorio otomí educativo en dispositivos móviles android. Journal-Economic History. 2021. 5-9: 1-7

* Correspondence to the Author (e-mail: erika.lopez@tesjo.edu.mx)
† Researcher contributing first author.

Introduction

The Otomi are a native people of Mexico with a presence in several entities of the Republic, especially in the central zone and up to the Gulf of Mexico in the entities of Mexico, Hidalgo, Guanajuato, Querétaro, Puebla and Veracruz, it is one of the ethnic groups more relevant numerically, the number of Otomi speakers places it as the seventh most spoken with a total of 288,052 speakers aged three and over, which represents 4.16 percent of the 6,913,362 speakers of the indigenous language in the country (INEGI, 2010).

In the State of Mexico, the National Institute of Statistics and Geography registers a total of 97,820 speakers of the Otomi language, most of whom live in the ethnoregion. However, one of the current social situations in the country and particularly in the State of Mexico is the loss of identity by the new generations regarding their roots, customs, traditions and culture; In a certain way, the reduction of the Otomí speakers is due to migration from the communities of origin and the urbanization of their ethnic territory, which imposes on them the need to coexist with an exclusively Spanish-speaking population for the most part; as mentioned (Questa, 2006).

He also points out that the elderly and children who attend bilingual education are the ones who speak, understand and use hñäñho (those who speak). There is a group of people who belong to a generation between thirty and forty years old who understand it but do not speak it. Lastly, the largest group is from twelve to thirty years old: they no longer know the language. The members of this last group mostly have non-bilingual primary education.

In the State of Mexico, the indigenous population is not concentrated mainly in rural localities, but has a strong presence in urban ones, which is involved in its different activities according to Montoya (Montoya-Casasola & Sandoval-Forero, 2013). To help safeguard this language, some work has been done, such as the researcher from the Autonomous University of Querétaro (Hekking, 2010), who published the design of a program on the Internet and in multimedia, for the teaching of the Otomi language that even has mobile phone applications.

On the other hand, the use of the Internet as a basis for development opens the panorama to generate contact between native speakers and the advantages that technology provides, breaking communication barriers and promoting progress in them as individuals and as a community, ensuring the legacy of the same by means of auxiliary tools. Technology through various electronic and systematic devices is an option to counteract the problem of language loss, taking advantage of frequent and massive use to promote the value of learning but above all to spread a language.

A study of perspectives, development strategy and dissemination of mobile applications in Mexico, presented by the Mexican association of the information technology industry (AMITI) in conjunction with the Information and Documentation Fund for Industry (INFOTEC) highlights that in 2012 the app development sector presented a growth of 100% compared to 2011, registering an estimated 500 companies in Mexico.

The presence of technology is promoting a decisive transformation in the way of seeing the world, the development of an application for Android devices using Apache Cordova as a development platform, applying Web Scraping for the collection of information to an Otomi repository, will help that young people and people cultivate the language or have easy and comfortable access for its translation.

Method Description

The methodology proposed for its development is shown in figure 1.

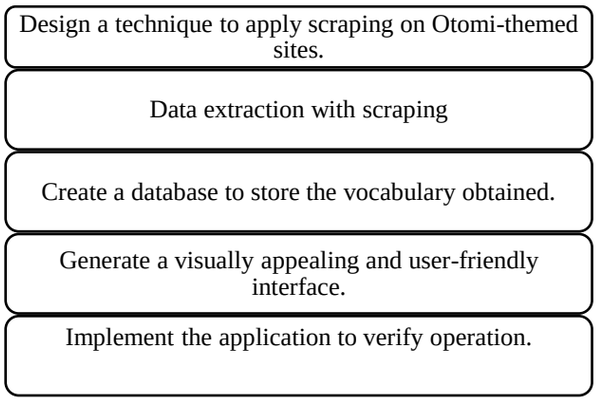

Figure 1 Proposed methodology

ISSN 2524-2059

Technique design

A program in python will allow obtaining all the information from a web page, the first one that was accessed to obtain the beginning of the vocabulary of the Otomi language was: http://portal2.edomex.gob.mx/cedipiem/indigenous peoples/otomi/palabrasenotomi/index. html The program works with the urllib2 module and the "urlopen()" function, which receives as a parameter the URL of the page to which the http request has been made, the result can be displayed in the HTML content of the site, figure 2.

Figure 2 Implementation of urllib2

Once the document was stored in a variable, the website content analysis process continued, using Beautiful Soup by importing the “from bs4 import BeautifulSoup” module. Figure 3 shows the use of the “BeautifulSoup()” function to which the variable that has the HTML was passed as an argument, in this way the different functions that BeautifulSoup offers can be used. The “find_all” function was then used to find all link tags to be analyzed.

Figure 3 Use of Beautiful Soup

Specifically, the text of the tags was analyzed looking for matches that met the pattern of the regular expression figure 4.

$$VAR = [\ [a\text{-}zA\text{-}Z]^{+}\ [a\text{-}z]^{*}\ [\ \lambda]\ ]^{+}$$

$$VAR^{3}[:]\ VAR^{5}$$

Figure 4 Regular phrase

Data extraction

Fulfilling the search criteria, the vocabulary was collected, starting with a website and from there analyzing the links on each page, generating a vocabulary of 5,906 words of the Otomi language, selecting 4,541 words stored in a plain text file. 5.

The Web Scraping technique using urllib2 and Beautiful Soup with a Python development environment allowed the content of 52 Otomi-themed websites to be analyzed.

Figure 5 Coword lesson

The storage was created on the client. First, because it allows an application to work when the user is offline, possibly synchronizing data when connected again. Second, it increases performance, so a large amount of data can be displayed as soon as the user clicks through to the site instead of waiting for it to re-download. Third, it is a programming model that does not require server infrastructure.

ISSN 2524-2059

Creation of the database

Web SQL Database an SQL database that unlike most browsers implements it using SQLite, whoseor dialect of SQLandIt's pretty complete. The stored information survives application restarts and is stored by the browser that PhoneGap/Cordova is using.

The design of the database figure 6, consists of two tables, the table "CAT_PALABRAS" in which all the words received are stored, this table has an "id" field that serves as an identifier for each element inserted, the field “esp” which is where the Spanish word is stored and the field “otomí” which stores the word in Otomí. The "HISTORICO_BUSQUEDA" table stores a history of the searches that the user performs, it has the "id" field that serves as the search identifier, the "id_word" field stores the id of the word searched for and the "date" field stores the date and time the search was performed.

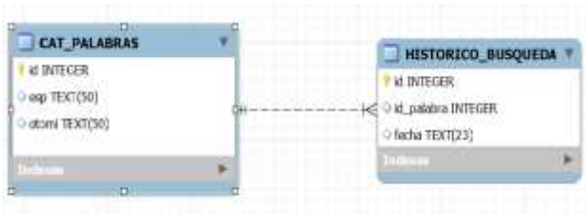

Figure 6 Entity relationship diagram

The API to manage the database needs to connect to the database or create a new one using the “openDatabase” function. If you try to open a database that doesn't exist, the API will create it on the fly, also you don't have to worry about closing the database, to create and open a database, use the following code: var db = openDatabase('mydb','1.0','my first database',2*1024*1024);

Once you have Database, you can execute transactions on the database using the method “db.transaction (...)”. var db = openDatabase('mydb','1.0','my first database',2*1024*1024);
db.transaction(function(tx){
// here be the transaction
// do SQL magic here using the tx object
});

Subsequently, a call is sent for a “executeSql” y ejecutar código SQL. var db = openDatabase('mydb','1.0','my first database',2*1024*1024);
db.transaction(function(tx){
tx.executeSql('CREATE TABLE foo (id unique, text)');
});

A simple table called "foo" will be created in the database called "mybd". Note that if the database already exists the transaction would fail, for this you can use another transaction, i.e. create a table if it does not exist and then make an insert to the table. var db = openDatabase('mydb','1.0','my first database',2*1024*1024);
db.transaction(function(tx){
tx.executeSql('CREATE TABLE IF NOT EXISTS foo (id unique, text)');
tx.executeSql('INSERT INTO foo (id, text) VALUES (1, "synergies")');
});

If the application is opened for the first time, a function called “save_dic()” is called, which is the one that will create the 'words' table at the same time it will insert its content. The vocabulary that was stored in a plain text file was saved in a JavaScript array identified as 'pal' as shown in Figure 7. The payload is 4541 words received into the “otomí” database.

```
var pal = [
    "a: a, n'a noya un n'a ha ra hmunts'a nsihni españämfo.",
    "a (hacia): hë",
    "a escondidas: ngu ma ägi",
    "a lo mejor: ua",
    "a poco: xige, hange",
    "a veces: r'abü",
    "abajo: ngati",
    "abandonado: xotsogi",
    "abandonar: tsogi, hëpü, hëgi",
    "abanicar: ts'üdi",
    "abanico: fuki, nthiti",
    "abaratar: k'omi",
    "abdicar: hiëgi",
    "abdomen: debi",
    "abecedario: hmunts'a nsihni",
    "abedul: täxiza",
    "abeja: sefi",
    "abejorro: gäni, hmini",
    "abierto: xogi",
    "abismo: hñe, ndengi, moho",
    "ablandar: tuki",
    "abnegar: jingi nhesë",
    "abochornado: thendi",
    "abogado: ñänte, fötsi",
    "abogar: ñäni, fötsi",
    "abominar: tsäni, ütsa",
    "abonar: lama, däb'i",
    "abonar (dinero): kjüti",
    "abono (estiércol): däb'i",
    "abordar: pëtse",
    "aborigen: mingu",
    "aborrecer: ütsa",
    "abortar: yaxki",
    "aborto: ñaxki",
    "abotonar: të'te",
    "abrazar: hüfi",
    "abrazo: hñüfi, nthüfi",
```

Figure 7 Array Javascript

Interface Generation

Lungo is based on and designed to take advantage of the most advanced technologies of Web standards such as HTML5, CSS3 and JavaScript, offering a homogeneous development environment for mobile devices, televisions or desktop devices.

Lyou pillars from Lungo fromfrom his birth are based on:

– Optimize the framework using the current features of HTML5.

– Focus on mobile development, leaving side functionalities and libreferences intended for desktop environments, which are not make sense in mobile applications.

– Provide a clear and easy-to-understand JavaScript API.

– Designed for current and future browsers.

– Ivector images, offering resolution independence.

– Creation of interfaces through semantic markup in HTML5.

– Possibility of extending the framework through plugins (known as sugars).

The main premise is to create a semantic structure, starting with the HTML markup language, continuing with a well-organized CSS, and ending with the JavaScript API. The minimum structure of the Lungo application body must contain at least:

– Section: the main container.

– Article: must be placed inside the section and must have the active class.

– Dependencies: The required JavaScripts are quo.js and lungo.js.

– Fstart function: the function that initializes Lungo.

Interface design and creation using the Lungo.js framework Figure 8, design created for Android mobile devices.

Figure 8 Menu bar view

Apache Cordova allows an application to be developed once and the same code can be compiled and deployed on multiple mobile operating systems. In general, applications on Cordova are created on HTML5, Javascript, CSS3 and are supported by a set of proprietary libraries that, depending on the operating system, allow access to device resources, such as camera, accelerometer, among others.

Creation of the apk file through the compilation of Apache Cordova, which as a final part was copied to the android device for its internal installation inside the Smartphone, in figure 9 the icon and the name of the already installed app can be seen, as well as the design and execution of its interface.

ISSN 2524-2059

Figure 9 APK installation

Results

In mobile technologies, unlike others (Web, desktop, digital TV), usability is a more significant problem, this is due to the mobility that these devices allow, usability tests in a real environment of use are difficult to carry out . First, it can be difficult to establish realistic studies that reflect the rich context of use described above. Second, it is far from trivial to apply classical evaluation techniques, when the test is carried out under real conditions of use. However, the use of certain objective metrics for the application is specified, such as:

Time required to enter data: This metric measures the time taken by the user to enter input data.

Time taken to respond: This metric measures the time taken by the application to respond to user input.

Subjective Tricks

Satisfaction with the output: This indicator measures the level of satisfaction provided by the application.

Satisfaction with the interface: It is also an important measure because a good interface will attract more users to use the application. Table 1 shows the evaluation of the metrics described above, testing the application in a real environment for its use.

Itriage to evaluate	Result
Ttime required to enter the data	15 sec
Ttime needed to respond	350ms
Exit Satisfaction	
Interface satisfaction	

Table 1 Evaluated Metrics

Thes performance tests can serve different purposes. They can demonstrate that the system meets performance criteria. For this, the robustness of the application was analyzed through Apkudo, a totally free tool that allows testing an application before being distributed, this tool tests the app on many devices, to show a response regarding the failure of the app installation. . At the same time, it shows the list of devices that have failed and the type of error that has been returned.

To carry out the analysis of the application, the site was accessedhttps://www.apkudo.com,registering the app to load the apk file, evidences the registration of the application with the name "Noya". It was shown that the application was tested on 37 different models of mobile devices, Figure 10 shows a table of results with a success status on 37 devices and 3 failures.

Figure 10 Rdevice reporting

Thank you

Tecnológico de Estudios Superiores de Jocotitlán

Conclusions

The knowledge and study of the Otomi language allowed us to identify the importance of this language, making it worthwhile to continue cultivating it, and what better way than using technology. In the same way the consultation of the vocabulary of the Otomi language was satisfactory consulting 52 sites that offered this information, allowing the set of 4541 words stored in the database thus acquiring a translation dictionary for the application using the scraping technique in each site.

ISSN 2524-2059

The design and creation of the interface was based on a friendly and optimal model that respects the user's usability criteria; applying the proposed tools such as Android, Apache Cordova, ResponsiveDesign.

The application has a database where the vocabulary is located, which allows to make use of this without internet, the performance of the application making queries is very fast since these queries are made locally, so you can make use of the application at any time, if you want to update the vocabulary or add new words, there will be the need to completely update the application.

Therefore it is concluded that the application is functional, it is needed to update the data periodically, only when necessary and this process can be carried out when the device is connected to a wifi network, besides requesting the user's authorization, in this way it will not be mandatory to update the application completely, only the database of the application will be modified, thus the behavior of the application will not be modified, only the databa se will be optimized.

References

Hekking, E. (24 de mayo de 2010). *El universal* .Recuperado el 17 de 02 de 2015, de http://www.eluniversal.com.mx/articulos/58766 .html

INEGI. (2010). *Censo de Población y Vivienda 2010,*. Mexico: Instituto Nacional de Estadística.
Montoya-Casasola, M. Á., & Sandoval-Forero, E. A. (2013). Marginación sociodemográfica de los otomíes del Estado de México. *Papeles de Población* , 257-289.

Questa, R. A. (2006). *Oyomies al norte del Estado de México y sur de Queretaro.* México: Comision Nacional para el Desarrollo de los pueblos Indigenas.

Solis, A. (19 de 04 de 2015). *FORBES*. Recuperado el 19 de 04 de 2015, de http://www.forbes.com.mx/las-15-apps-mas-utilizadas-del-mundo/

ISSN 2524-2059