Articulo **Revista de Tecnologías de la Información**
Septiembre 2016 Vo.3 No.8 72-79

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

LÓPEZ-GONZÁLEZ, Erika*†, ALEJO, Roberto, ANTONIO-VEL-AZQUEZ, J y AMBRIZ-POLO, J.

Recibido Julio 20, 2016; Aceptado Septiembre 26, 2016

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

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

Citación: LÓPEZ-GONZÁLEZ, Erika, ALEJO, Roberto, ANTONIO-VEL-AZQUEZ, J y AMBRIZ-POLO, J. Análisis de datos utilizando web scraping para repertorio otomí educativo en dispositivos móviles android. Revista de Tecnologías de la Información 2016. 3-8: 72-79

*Correspondencia al Autor (Correo Electrónico: erika.lopez@tesjo.edu.mx)
† Investigador contribuyendo como primer autor.

Introducción

Los otomíes son un pueblo originario de México con presencia en varias entidades de la República, sobre todo de la zona centro y hasta el Golfo de México en las entidades de México, Hidalgo, Guanajuato, Querétaro, Puebla y Veracruz, es una de las etnias más relevantes numéricamente, la cantidad de hablantes de otomí la ubica como la séptima más hablada con un total de 288 052 hablantes de tres años y más, lo que representa 4.16 por ciento de los 6 913 362 hablantes de lengua indígena que hay en el país (INEGI, 2010).

En el Estado de México, el Instituto Nacional de Estadística y Geografía 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; en cierta forma la reducción de los hablantes de otomí se debe a la migración desde las comunidades de origen y a la urbanización de su territorio étnico, que les impone la necesidad de convivir con una población exclusivamente hispanófona en su mayoría; como lo menciona (Questa, 2006). También desataca que las personas de edad avanzada y los niños que asisten a la educación bilingüe son quienes hablan, entienden y utilizan el hñäñho (los que hablan). Hay un grupo de personas que pertenecen a una generación de entre treinta y cuarenta años que lo entiende pero no lo habla. Por último, el grupo más extenso es el de doce a treinta años: ellos ya no conocen el idioma. Los integrantes de este último grupo cuentan en su mayoría con educación primaria no bilingüe.

El Estado de México la población indígena no se concentra principalmente en localidades rurales, sino que tiene una fuerte presencia en las urbanas, lo que involucra en sus diferentes actividades según Montoya.(Montoya-Casasola & Sandoval-Forero, 2013). Para ayudar a salvaguardar esta lengua se han hecho algunos trabajos, como el investigador de la universidad Autónoma de Querétaro(Hekking, 2010), éste publicó el diseño de un programa por internet y en multimedia, para la enseñanza del lenguaje otomí que incluso cuenta con aplicaciones para telefonía móvil.

Por otro lado el uso de internet como base para el desarrollo, abre el panorama para generar el contacto entre los hablantes nativos y las ventajas que la tecnología proporciona, rompiendo las barreras de comunicación e impulsando el progreso en ellos como individuos y comunidad, asegurando el legado de la misma por medio de las herramientas auxiliares. La tecnología a través de diversos dispositivos electrónicos y sistemáticos es una opción para contrarrestar la problemática de la perdida de lengua, aprovechando el uso frecuente y masivo para impulsar el valor de aprender pero sobre todo para difundir una lengua.

Un estudio de perspectivas, estrategia de desarrollo y difusión de aplicaciones móviles en México, presentado por la asociación mexicana de la industria de tecnologías de información (AMITI) en conjunto con el Fondo de Información y Documentación para la Industria (INFOTEC) destaca que en 2012 el sector de desarrollo de apps presentó un crecimiento del 100% con respecto al 2011, registrando un estimado de 500 empresas en México.

ISSN 2410-4000

La presencia de la tecnología está promoviendo una transformación decisiva en la manera de ver el mundo, el desarrollo de una aplicación para dispositivos Android utilizando Apache Cordova como plataforma de desarrollo, aplicando Web Scraping para la recolección de información a un repositorio otomí, permitirá ayudar a que los jóvenes y personas cultiven la lengua o tengan un acceso fácil y cómodo para su traducción.

Descripción del método

La metodología que se propone para el desarrollo de éste se observa en la figura 1.

Diseñar una técnica para aplicar el scraping en sitios con temática otomí.
Extracción de datos con scraping
Crear una base de datos para almacenar el vocabulario obtenido.
Generar una interfaz visualmente atractiva y fácil de utilizar.
Implementar la aplicación para verificar el funcionamiento.

Figura . Metodología propuesta

Diseño de técnica

Un programa en python permitirá obtener toda la información de una página web, la primera a la que se tuvo acceso para obtener el inicio del vocabulario de la lengua otomí fue:

http://portal2.edomex.gob.mx/cedipiem/pueblosindigenas/otomi/palabrasenotomi/index.htm

El programa trabajo con el módulo urllib2 y la función “urlopen()” la cuál recibe como parámetro la URL de la página a la que se le ha hecho la petición http, el resultado se puede visualizar en contenido HTML del sitio, figura 2.

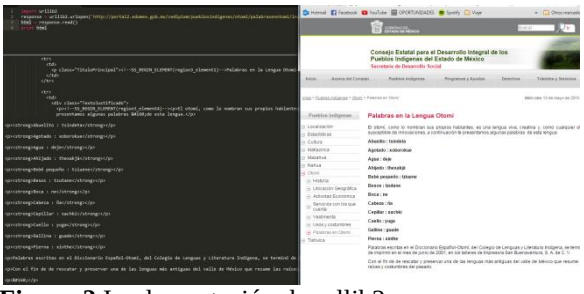

Figura 2 Implementación de urllib2

Una vez almacenado el documento en una variable, se continuó con el proceso de análisis al contenido del sitio web, usando Beautiful Soup importando el módulo “from bs4 import BeautifulSoup”. La figura 3 muestra el uso de la función “BeautifulSoup()” a la cuál se le pasó como argumento la variable que tiene el HTML, de este modo se pueden usar las diferentes funciones que brinda BeautifulSoup. Posteriormente se usó la función “find_all” para encontrar todas las etiquetas de enlace que se analizarían.

```
import urllib2
from bs4 import BeautifulSoup
response = urllib2.urlopen('http://portal2.edomex.gob.mx/cedipiem/pueblosindigenas/otomi/palabrasenoto
html = response.read()
soup=BeautifulSoup(html)
for link in soup.find_all('a'):
    print(link.get('href'))
```

```
None
http://www.edomex.gob.mx
/cedipiem/inicio/index.htm
/cedipiem/acercadelconsejo/index.htm
/cedipiem/pueblosindigenas/index.htm
/cedipiem/programasyayudas/index.htm
/cedipiem/derechos/index.htm
/cedipiem/tramitesyservicios/index.htm
mailto:cedipiem_ve@yahoo.com.mx
javascript:nodelink('4650');
javascript:nodelink('4651');
[Finished in 0.4s]
```

Figura 3 Uso de Beautiful Soup.

Específicamente se analizó el texto de las etiquetas buscando coincidencias que cumplieran con el patrón de la expresión regular figura 4.

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

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

Figura 4 Expresión Regular.

ISSN 2410-4000

Extracción de datos

Cumpliendo con los criterios de búsqueda se hizo la recolección de vocabulario, iniciando con un sitio web y de ahí analizando los enlaces de cada página, generando un vocabulario de 5906 palabras de la lengua otomí, seleccionando 4541 palabras almacenadas en un archivo de texto plano figura 5.

La técnica Web Scraping mediante urllib2 y Beautiful Soup con un entorno de desarrollo Python, permitieron analizar el contenido de 52 sitios web con temática otomí.

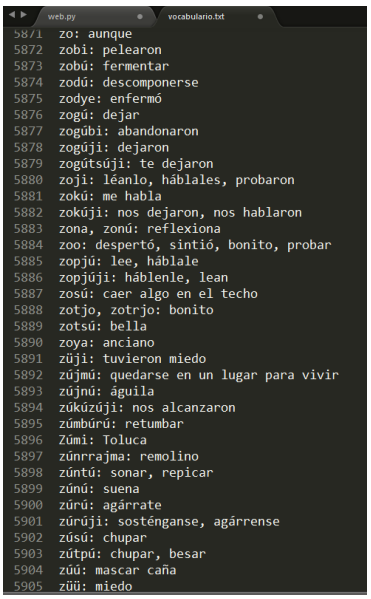

Figura 5 Colección de palabras.

El almacenamiento se creó en el cliente. En primer lugar porque permite que una aplicación funcione cuando el usuario no está conectado, posiblemente sincronizando datos cuando vuelve a establecer conexión. En segundo lugar, aumenta el rendimiento, por lo que se puede mostrar una gran cantidad de datos en cuanto el usuario hace clic en el sitio en lugar de esperar a que vuelvan a descargarse. En tercer lugar, es un modelo de programación que no requiere infraestructura de servidor.

Creación de la base de datos

Web SQL Database una base de datos SQL que a diferencia de la mayoría de los navegadores la implementan usando SQLite, cuyo dialecto de SQL es bastante completo. La información almacenada sobrevive a reinicios de aplicación y es almacenada por el navegador que esté usando PhoneGap/Cordova.

El diseño de la base de datos figura 6, consta de dos tablas, la tabla "CAT_PALABRAS" en la cual se almacena todas las palabras recibidas, esta tabla cuenta con un campo "id" que sirve como identificador para cada elemento insertado, el campo "esp" que es donde se almacena la palabra en español y el campo "otomí" que almacena la palabra en otomí. La tabla "HISTORICO_BUSQUEDA" almacena un historial de las búsquedas que el usuario realice, cuenta con el campo "id" que sirve como identificador de la búsqueda, el campo "id_palabra" almacena el id de la palabra buscada y el campo "fecha" almacena la fecha y hora en que se realizó la búsqueda.

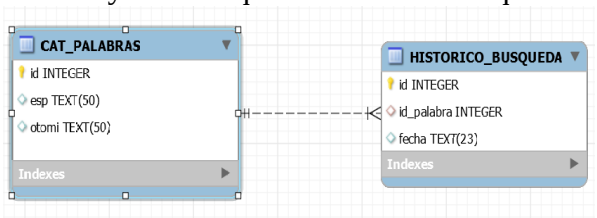

Figura 6 Diagrama Entidad-Relación.

El API para manejar la base de datos necesita conectarse a la base de datos o crear una nueva usando la función "openDatabase". Si se intenta abrir una base de datos que no existe, la API la creara sobre la marcha, también no hay que preocuparse por cerrar la base de datos, para crear y abrir una base de datos, se usa el siguiente código:

```
var db = openDatabase('mydb','1.0','my first database',2*1024*1024);
```

ISSN 2410-4000

Una vez que se tiene Database, se pueden ejecutar transacciones sobre la base de datos usando el método “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
});
```

Posteriormente se manda llamar 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)');
});
```

Se creara una simple tabla llamada “foo” en la base de datos llamada “mybd”. Hay que notar que si la base de datos ya existe la transacción fallaría, para esto se puede utilizar otra transacción es decir crear una tabla si ésta no existe y posterior hacer una inserción a la tabla.

```
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")');
});
```

Si la aplicación es abierta por primera vez, se llama a una función llamada “guarda_dic ()” que es quien va a crear la tabla ‘palabras’ al mismo tiempo insertará su contenido. El vocabulario que estaba almacenado en un archivo de texto plano se guardó en un arreglo de JavaScript identificado como ‘pal’ como se muestra en la figura 7. La carga es de 4541 palabras recibidas hacia la base de datos “otomí”.

```
var pal = [
    "a: a, n’a noya un n’a ha ra hmunts’a nsihni españämfo.",
    "a (hacia): ha",
    "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’ami",
    "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",
```

Figura 7 Array Javascript.

Generación de la interfaz

Lungo está basado y pensado para aprovechar las características de tecnologías más avanzadas de estándares Web como HTML5, CSS3 y JavaScript, ofreciendo un entorno de desarrollo homogéneo para dispositivos móviles, televisores o dispositivos de escritorio.

Los pilares de Lungo desde su nacimiento se basan en:

- Optimizar el framework haciendo uso de las características actuales de HTML5.
- Enfocarse en desarrollo móvil, dejando de lado funcionalidades y librerías destinadas a entornos de escritorio, que no tienen sentido en aplicaciones móviles.
- Proveer de una API JavaScript clara y sencilla de entender.

ISSN 2410-4000

- Pensado para navegadores actuales y futuros.

- Imágenes vectoriales, ofreciendo una independencia de resolución.

- Creación de interfaces a través de marcado semántico en HTML5.

- Posibilidad de extender el framework a través de plugins (conocidos como **sugars**).

La principal premisa es crear una estructura semántica, empezando por el lenguaje de marcado HTML, siguiendo con un buen organizado CSS y terminando con el API de JavaScript. La estructura mínima del cuerpo de la aplicación Lungo debe contener al menos:

- Sección: el contenedor principal.

- Artículo: debe estar colocado dentro de la sección y debe tener la clase activa.

- Dependencias: Los JavaScript necesarios son quo.js y lungo.js.

- Función de inicio: la función que inicializa Lungo.

Diseño y creación de interfaz mediante el framework Lungo.js La figura 8, diseño creado para dispositivos móviles android.

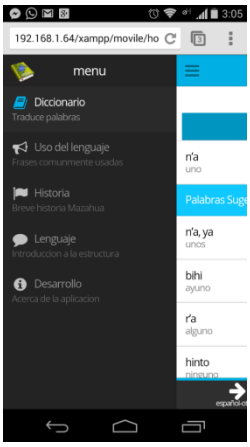

Figura 81 Vista barra de Menú

Apache Cordova permite que una aplicación se desarrolle una sola vez y que el mismo código pueda ser compilado y desplegado en diversos sistemas operativos móviles. En general, las aplicaciones sobre Cordova están creadas sobre HTML5, Javascript, CSS3 y están soportadas por un conjunto de librerías propias que, de acuerdo al sistema operativo, permiten acceder a los recursos del dispositivo, como cámara, acelerómetro entre otros.

Creación del archivo apk mediante la compilación de Apache Cordova, que como parte final fue copiado al dispositivo android para su instalación interna dentro del Smartphone, en la figura 9 se percibe el icono y el nombre de la app ya instalada, así como el diseño y ejecución de su interfaz.

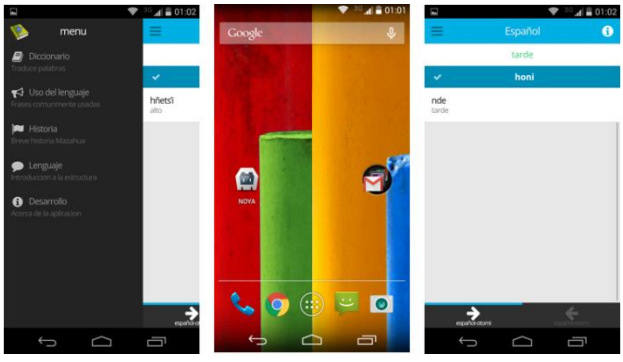

Figura 92 Instalación de APK.

Resultados

En las tecnologías móviles a diferencia de otras (Web, escritorio, TV digital), la usabilidad es un problema más significativo, esto es debido a la movilidad que permiten estos dispositivos, las pruebas de usabilidad en un ambiente real de uso son difíciles de realizar. En primer lugar, puede ser complicado establecer estudios realistas que reflejan la riqueza del contexto de uso descrito anteriormente. En segundo lugar, está lejos de ser trivial aplicar técnicas de evaluación clásicas, cuando la prueba se realiza en condiciones reales de uso. Sin embargo se puntualizan el uso de ciertas métricas objetivas para la aplicación como son:

Tiempo requerido para introducir los datos: Esta métrica mide el tiempo tomado por el usuario para ingresar datos de entrada.

Tiempo necesario para responder: Esta métrica mide el tiempo empleado por la aplicación para responder a la entrada del usuario.

Métricas Subjetivas

Satisfacción con la salida: Este indicador mide el nivel de satisfacción proporcionada por la aplicación.

Satisfacción con la interfaz: Es también una medida importante porque una buena interfaz atraerá a más usuarios a utilizar la aplicación.

La tabla 1 muestra la evaluación de las métricas descritas anteriormente, probando la aplicación en un ambiente real para su uso.

Métrica a evaluar	**Resultado**
Tiempo requerido para introducir los datos	15 seg.
Tiempo necesario para responder	350ms
Satisfacción de salida	
Satisfacción de la interfaz	

Tabla 1 Métricas Evaluadas

Las pruebas de rendimiento pueden servir para diferentes propósitos. Pueden demostrar que el sistema cumple los criterios de rendimiento. Para ello se analizó la robustez de la aplicación por medio de Apkudo una herramienta totalmente libre que permite probar una aplicación antes de ser distribuida, esta herramienta prueba la app en muchos dispositivos, para evidenciar una respuesta en cuanto al fallo de la instalación de la app. Al mismo tiempo muestra el listado de dispositivos que han fallado y el tipo de error que ha devuelto.

Para llevar a cabo el análisis de la aplicación se accedió al sitio https://www.apkudo.com, registrando la app para la carga del archivo apk, evidencia el registro de la aplicación con el nombre “Noya”. Se mostró que la aplicación fue probada en 37 modelos diferentes de dispositivos móviles, la figura 10 muestra una tabla de resultados con un estado de éxito en 37 dispositivos y 3 fallas.

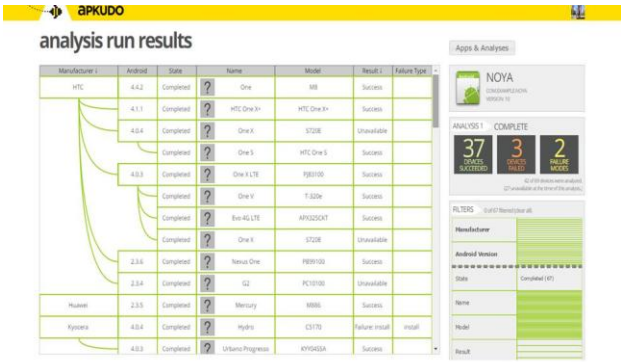

Figura 103 Reporte de dispositivos.

Agradecimiento

Tecnológico de Estudios Superiores de Jocotitlán

Conclusiones

El conocer y estudiar el lenguaje otomí permitió identificar la importancia que tiene esta lengua valiendo la pena para seguirla cultivando y que mejor que el utilizar la tecnología. De igual manera la consulta del vocabulario de la lengua otomí fue satisfactoria consultando 52 sitios que ofrecían esta información, permitiendo el conjunto de 4541 palabras almacenadas en la base de datos adquiriendo de esta manera un diccionario de traducción para la aplicación utilizando la técnica de scraping en cada sitio.
El diseño y creación de la interfaz fue basado en un modelo amigable y óptimo que respeta los criterios de usabilidad del usuario; aplicando las herramientas propuestas como Android, Apache Cordova, ResponsiveDesign.

La aplicación cuenta con una base de datos donde se encuentra el vocabulario, lo cual permite hacer uso de esta sin necesidad de internet, el desempeño de la aplicación haciendo consultas es muy rápido ya que esas consultas las hace de manera local, así que se puede hacer uso de la aplicación en cualquier momento, si se desea actualizar el vocabulario o agregar nuevas palabras, habrá la necesidad de actualizar completamente la aplicación.

Por lo tanto se concluye que la aplicación es funcional, se necesita actualizar los datos periódicamente, solamente cuando sea necesario y este proceso se puede llevar a cabo cuando el dispositivo esté conectado a una red wifi, además de solicitar la autorización del usuario, de este modo no será obligatorio actualizar la aplicación por completo, solamente se modificará la base de datos de la aplicación, con lo cual no se modificará el comportamiento de la aplicación, sólo se optimizará la base de datos.

Referencias

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 2410-4000