Revista Habanera de Ciencias Médicas

AÑO 2019
18(4) JULIO-AGOSTO
ISSN 1729 - 519X

ecimed

CIENCIAS TECNOLÓGICAS
ARTÍCULO ORIGINAL

Software para la categorización semiautomática de asociaciones libres sobre el bienestar de trabajadores habaneros

Software for semi-automatic categorization of free associations to promote the welfare of workers of Havana

Damián Valdés Santiago[1]*, José Carlos Oliva Guerrero[1]

[1] Universidad de La Habana. Facultad de Matemática y Computación. Departamento de Matemática Aplicada. La Habana, Cuba.

***Autor para la correspondencia:** dvs89cs@matcom.uh.cu

Cómo citar este artículo

Valdés Santiago D, Oliva Guerrero JC. Software para la categorización semiautomática de asociaciones libres sobre el bienestar de trabajadores habaneros. Rev haban cienc méd [Internet]. 2019 [citado]; 18(4):678-692. Disponible en: http://www.revhabanera.sld.cu/index.php/rhab/article/view/2529

Recibido: 19 de noviembre del 2018.
Aprobado: 03 de junio del 2019.

RESUMEN

Introducción: Especialistas de la Facultad de Psicología de la Universidad de La Habana propusieron el cuestionario sobre Bienestar Humano Personal, Laboral y Social (BHPLS), que se aplicó a 135 trabajadores cubanos de tres grupos sociolaborales. Dada la variedad de respuestas, se impuso un análisis de contenido (AC) para la Pregunta 1 del cuestionario.

Objetivo: Proponer e implementar un software que permita la categorización semiautomática en un AC para dicha pregunta.

Material y Métodos: Se utilizó el índice de concordancia Kappa para evaluar el acuerdo entre expertos respecto al esquema de categorías. Se implementó un software en el lenguaje de programación Python para cumplir el

objetivo, considerando las funcionalidades de softwares similares.

Resultados: Se implementó, validó y registró un software "BHPLS data processing-UH®" que permite establecer las categorías, cargar los datos, categorizarlos semiautomáticamente y guardar el resultado, entre otras funcionalidades. La categorización manual con estudiantes de Psicología obtuvo un índice de concordancia Kappa negativo (bajo acuerdo entre expertos), mientras que usando el software propuesto, se alcanzó un Kappa global 0.7871 con p=0.00 (alta concordancia y alta significación estadística). Además, se propuso un algoritmo para la unificación de las categorizaciones de expertos y se ejecutó un Análisis de Correspondencias (ANACOR) sobre la combinación de categorizaciones obtenidas.

Conclusiones: Dada la alta concordancia alcanzada, se recomienda el uso del software por su adaptabilidad, facilidad de uso y la "humanización" del AC. El ANACOR permitió observar similitudes entre los grupos sociolaborales. Las funcionalidades del software pueden aplicarse para el procesamiento de asociaciones libres en otros escenarios.

Palabras clave: análisis de contenido, análisis asistido de datos cualitativos, minería de texto, bienestar humano, asociaciones libres, índice de concordancia Kappa.

ABSTRACT

Introduction: Experts of the Faculty of Psychology of the University of Havana proposed the Personal, Labor and Social Human Well-being questionnaire (BHPLS, in Spanish), that was applied to 135 Cuban workers of three social and occupational groups. Given the variety of responses, a content analysis (CA) was used for Question 1 of the mentioned questionnaire.

Objective: To present and implement a software that allows a semi-automatic categorization in a CA used for this question.

Material and Methods: The Kappa index test was used to evaluate experts´ agreement with respect to category schemes. We implemented a software with the Python programming language to achieve our objective, considering other similar software functionalities.

Results: We implemented, validated and registered the software BHPLS data processing-UH® that allows to set up a categories system, load the collected data, categorize associations in a semi-automatic way, and save the results, among other functionalities. This software was validated by Psychology students and, when they performed the manual categorization, a negative Kappa agreement index (low categorization agreement between experts) was obtained whereas using the proposed software, a global Kappa index of 0.7871 with p=0.00 (high and statistically significant categorization agreement between experts) was obtained. Besides, we proposed a unified algorithm for expert's categorizations, and carried out a Correspondence Analysis (ANACOR) on the basis of the categorizations achieved.

Conclusions: According to the high concordance attained, we recommend the software due to its adaptability, ease of use, and "humanization'' of

the process. The CA allowed us to observe similarities in social and occupational groups. The software functionalities can be applied for processing free associations in other scenarios.

Keywords: content analysis, qualitative data analysis, text mining, human welfare, free associations, Kappa index test.

INTRODUCCIÓN

El análisis de datos textuales es un tema importante para la psicología social. Estos datos permiten el estudio del lenguaje con suficiente distancia entre el investigador y lo investigado. Hace unos años atrás la única posible vía para analizar estos datos era un largo y exhaustivo trabajo de clasificación manual para realizar un análisis de contenido (AC),[1] seguido del análisis estadístico de las clases o categorías obtenidas. Actualmente, el volumen de datos textuales presentes en diversos contextos ha aumentado drásticamente debido al desarrollo de los medios de comunicación y los diversos formatos de publicación. Sin embargo, la informática permite (semi)automatizar y controlar este proceso mediante software.

Jenny[2] agrupa el software para el procesamiento de texto en cuatro grandes grupos: lexicométrico, sociosemántico, redes de asociaciones de palabras y análisis del discurso proposicional y predictivo. El enfoque más usado en el procesamiento de representaciones sociales es el lexicométrico.[3,4] Dicho enfoque consiste en comparar perfiles léxicos (distribuciones relativas de ocurrencias léxicas) dentro de un corpus o entre corpora.[2]

El principio del AC está, sobre todo, basado en la concepción de ciertas categorías globales o clases que tienen su propio significado, y luego en la distribución de varios elementos de los textos considerados en dichas clases con un propósito descriptivo y comparativo. La etapa final del análisis es un cálculo que puede apelar a la estadística descriptiva o a la inferencial. Esto se corresponde con el enfoque sociosemántico descrito por Jenny.[2]

Según este investigador, los programas de este tipo segmentan el corpus en unidades de significado relevantes, logran una categorización multidimensional de acuerdo con el entramado conceptual específico a cada investigación (i.e. una codificación a posteriori en la que el investigador lee el texto, anota y codifica las unidades de significado en el corpus), y pueden usar métodos estadísticos.[2]

Para el software basado en el enfoque lexicométrico ocurre a la inversa. Todo comienza con el cómputo que genera clases o grupos de elementos en los textos. La última fase consiste en darle un significado por parte del investigador a las clases obtenidas. Los dos enfoques son radicalmente diferentes: en el primer caso se computa partiendo del significado, en el otro, se encuentra el significado mediante el cómputo.

Los software para Análisis de Datos Cualitativos (SADC) se utilizan en investigaciones en ciencias sociales. Brindan una amplia variedad de herramientas cuya asistencia posibilita optimizar el procesamiento e interpretación de grandes volúmenes de datos, contribuir a la validez de los

hallazgos y facilitar el intercambio de opiniones entre varios investigadores que no necesariamente comparten el mismo espacio-tiempo de trabajo.

Más allá de las ventajas operativas y analíticas, la incorporación de los SADC también ha suscitado una serie de objeciones. Algunas voces reconocen que dichos recursos informáticos pueden extremar el fraccionamiento de la información y perder una visión integrada, distanciar al investigador de los datos o incluso retrotraer los avances logrados en el análisis cualitativo a épocas pasadas.

Se pueden utilizar diferentes SADC, entre los más utilizados están los siguientes: ATLAS.ti[5] es un SADC propietario alemán que facilita el desarrollo de las tareas propias de cualquier análisis cualitativo de datos en soporte textual y multimedia y es especialmente apropiado para proyectos de investigación que involucran grandes volúmenes de datos. Permite la segmentación del texto en citas; la codificación de los documentos analizados en función de un sistema de categorías; la recuperación selectiva de datos en función de las necesidades del investigador; la elaboración de comentarios y anotaciones; la generación de familias de documentos, códigos y memos; y la representación gráfica de las relaciones teóricas identificadas y construidas durante el análisis.

NVivo[6] es un software propietario desarrollado por QSR International que permite organizar, analizar y encontrar perspectivas en datos no estructurados como entrevistas, respuestas de encuestas con preguntas abiertas, artículos, contenido de redes sociales y web, que pueden aparecer digitales tanto en texto, audio, video o imagen. Además, brinda herramientas para la codificación de datos, la descripción de contenido, la preparación de relaciones entre códigos a través de un sistema de nodos y árboles, y la rapidez en la búsqueda y presentación de la información.

MAXQDA es un software propietario que puede organizar, categorizar, buscar y recuperar información de cualquier tipo de archivos multimedia. Entre sus funcionalidades están la categorización de segmentos de datos, el análisis rápido de preguntas de encuestas que pueden ser importadas desde Excel, el análisis estadístico de los datos cualitativos, la creación de comentarios, notas y memos; el resumen de contenido haciendo uso de su modo Paraphrase y la interconexión de datos.

Las facultades de Psicología y Matemática y Computación de la Universidad de La Habana colaboran en el proyecto institucional donde se diseñó, por un equipo de psicólogos, el cuestionario sobre Bienestar Humano Personal, Laboral y Social (BHPLS).[7,8] El cuestionario consta de preguntas relativas a la noción de bienestar de los sujetos estudiados. Se entiende el bienestar humano como un constructo multidimensional que consiste en juicios de satisfacción sobre los diferentes dominios o esferas de la vida del sujeto.

Dada la variedad de respuestas y la cantidad de sujetos participantes es necesario reducir la información textual. Esto se logra mediante la creación de un sistema de categorías o clases y la categorización de las respuestas en dichas clases mediante un AC.[9] Luego de la categorización, por

al menos dos expertos de la misma calificación, se obtiene una categorización única que refleja el consenso estos, según el índice de concordancia Kappa.[10] Con dicha categorización se analizan las similitudes de las respuestas por estratos de la muestra, determinados por variables sociodemográficas como el sexo, el grupo de edad y el grupo sociolaboral. Para esto se utiliza el Análisis de Correspondencias (ANACOR).[11]

El AC se realiza típicamente de forma manual, lo que provoca errores en el proceso y la consecuente demora. Aunque existen software para realizar AC en ciencias sociales, como los SADC mencionados anteriormente, no es factible adaptarlos para su uso en el procesamiento de datos del cuestionario BHPLS, puesto que la mayoría tienen como entrada textos largos, el cuestionario usado tiene estructura propia y los textos son relativamente cortos. Además, la categorización por triangulación (codificada mediante el vector 3D) es necesaria para una visión integradora de las respuestas y constituye un aporte al procesamiento del cuestionario, por lo que no se ajusta al modelo de cómputo propuesto por los programas previamente analizados.

Por ello, esta investigación tiene como **objetivo** diseñar, implementar y validar un software para semiautomatizar el AC para datos provenientes de la Pregunta 1 del cuestionario mencionado.

MATERIAL Y MÉTODOS

Obtención de los datos

Se aplicó el cuestionario sobre Bienestar Humano Personal, Laboral y Social (BHPLS)[7,8] a una muestra no probabilística tomando como población a trabajadores cooperativistas, cuentapropistas y estatales. Otros criterios de inclusión fueron la voluntariedad y la pertenencia al sector económico de la gastronomía. Se escogió este sector debido a la particularidad de presentar tres de las modalidades organizativas presentes en la actualización del modelo de la economía cubana. Se incluyeron entidades de tamaño pequeño y mediano. La muestra quedó conformada por 135 individuos. El registro de los datos se realizó mediante Microsoft Excel 2013.

Dada la naturaleza textual de las respuestas a las preguntas, se realizó un preprocesamiento de los datos para corregir faltas de ortografía y descartar individuos con muchos datos faltantes (no respuestas).

En esta investigación se procesó solo la Pregunta 1 del cuestionario BHPS. Esa pregunta pide a los individuos responder mediante asociaciones libres (frases escritas cortas) tres preguntas referidas a su percepción de bienestar, de quién depende este y qué está haciendo para alcanzarlo. El encuestado puede dar hasta cinco respuestas a cada una de estas preguntas y se considera que cada una está relacionada según su orden.

Categorización semiautomática

Dada la variedad de respuestas de los sujetos es necesario codificarlas para poder realizar el análisis, para ello y, basándose en la frecuencia de n-gramas y las correspondientes nubes de palabras, se construyó un esquema de 18 categorías. Para proceder a asignar respuestas en cada categoría se realizó el siguiente

procedimiento.

A cada sujeto se le asigna un número consecutivo o ID para su identificación en la base de datos. Cada sujeto tiene que responder a tres preguntas: ¿Qué características considera usted que definen un estado óptimo de bienestar?, ¿De quién depende el que usted alcance dicha característica? y ¿Qué es lo que usted está haciendo para alcanzar o satisfacer dichas características? con cinco posibilidades de respuesta.

Por cada respuesta se obtiene un vector 3D formado por los textos evocados por el sujeto a cada una de las tres preguntas previas. Cada vector 3D es identificado como <ID del sujeto>-<número de la respuesta> (e.g. el vector 3-4 se refiere a la cuarta respuesta del sujeto 3 a las preguntas descritas).

Los resultados de la categorización hecha por cada psicólogo (de forma manual o con software) se guardan en un archivo Excel cuyas columnas representan a cada categoría del esquema construido, incluyendo una columna para vectores 3D que no se ajustan a este. Luego, los psicólogos participantes en la categorización colocan cada vector 3D en la columna perteneciente a la categoría que estos le asignan.

El proceso de categorización manual realizado por los psicólogos consistía en: crear el archivo donde guardar los ID categorizados, leer de la base de datos de las encuestas las respuestas de un sujeto, identificar los IDs de los vectores asociados a este, categorizarlos uno a uno manualmente, guardar los resultados en el Excel creado, y repetir para todos los sujetos encuestados.

Luego de obtener todas las categorizaciones de los expertos, se procede a verificar si estos concuerdan respecto a la asignación de cada vector 3D a cada categoría. Para ello se computa el índice Kappa.

Índice de concordancia Kappa

El índice Kappa permite medir el acuerdo de un grupo de expertos frente a la categorización de ciertos objetos en un sistema predeterminado de clases o categorías. Existen varios escenarios de uso como el caso de dos categorías y dos jueces, dos categorías y más de dos jueces, y tres o más categorías y más de dos jueces. Precisamente, este fue el caso en la categorización implementada en el software presentado en este artículo.

Kappa de la categoría j e intervalo de confianza

El índice Kappa de la categoría j [12] se computa de la siguiente forma:

$$\widehat{k_j} = 1 - \frac{\sum_{i=1}^{n} x_{ij}(m - x_{ij})}{n\,m\,(m-1)\,\overline{p_j}\,\overline{q_j}}$$

El intervalo de confianza *jackknife* para el Kappa de la categoría j con nivel de confianza $(1-\alpha)\%$ [10]:

$$\left(J(k_j) - t_{n-1,1-\frac{\alpha}{2}}\,S_j, J(k_j) + t_{n-1,1-\frac{\alpha}{2}}\,S_j\right)$$

Prueba de significación para el Kappa de la categoría j

El estadístico para contrastar la hipótesis $H_0: k_j = 0$ frente a $H_1: k_j \neq 0$ es:

$$z = \frac{\widehat{k_j}}{EE_{0(\widehat{k_j})}} \sim N(0,1)$$

El error estándar de Kappa para el contraste:

$$EE_{0(\widehat{k_j})} = \sqrt{\frac{2}{nm(m-1)}}$$

donde:

- n es el número de sujetos,
- m es el número de observadores por sujeto,
- k es el número de categorías,
- x_{ij} es el número de clasificaciones del sujeto i en la categoría j y $i = 1, \cdots, n$, $j = 1, \cdots, k$,
- $\bar{p_j} = \frac{1}{nm}\sum_{i=1}^{n} x_{ij}$ es la proporción global de clasificaciones en la categoría j, $j = 1, \cdots, k$ y $\bar{q_j} = 1 - \bar{p_j}$,
- $J(k_j)$ es la estimación *jackknife* de Kappa y S_j es el error estándar de $J(k_j)$,
- $t_{n-\,1,1-\frac{\alpha}{2}}$ es el percentil de la distribución t de Student con $n-1$ grados de libertad que deja a la izquierda una cola de probabilidad $1-\frac{\alpha}{2}$, y
- $1-\alpha$ es el nivel de confianza.

Kappa global e intervalo de confianza

El índice Kappa global[12] se calcula de la siguiente forma:

$$\hat{k} = \frac{\sum_{j=1}^{k} \bar{p_j}\left(1-\bar{p_j}\right)\widehat{k_j}}{\sum_{j=1}^{k} \bar{p_j}\left(1-\bar{p_j}\right)}$$

Su intervalo de confianza *jackknife* para el Kappa global con nivel de confianza $(1-\alpha)\%$(10) se computa según:

$$\left(J(k_j) - t_{n-\,1,1-\frac{\alpha}{2}}\, S, J(k_j) + t_{n-\,1,1-\frac{\alpha}{2}}\, S\right)$$

Prueba de significación para el Kappa global

El estadístico para contrastar $H_0: k = 0$ frente a $H_1: k \neq 0$ es:

$$z = \frac{\hat{k}}{EE_{0(\hat{k})}} \sim N(0,1)$$

El error estándar de Kappa para el contraste se estima mediante la expresión:

$$EE_{0(\hat{k})} = \frac{\sqrt{2}}{\sum_{j=1}^{k} \bar{p_j}\bar{q_j}\left(nm(m-1)\right)} \cdot \sqrt{\left(\sum_{j=1}^{k} \bar{p_j}\bar{q_j}\right)^2 - \sum_{j=1}^{k} \bar{p_j}\bar{q_j}\left(\bar{q_j} - \bar{p_j}\right)}$$

donde $\widehat{k_j}$ es el Kappa de la categoría j con $j = 1, \cdots, k$, y $J(k)$ es la estimación jackknife de Kappa y S es el error estándar de $J(k)$.

Algoritmo para la unificación de categorizaciones

El algoritmo para la obtención de una sola categorización, sobre la misma muestra y el mismo esquema de categorías, que resuma los criterios de diversos expertos con alto índice Kappa, recibe como entrada archivos Excel con las categorizaciones realizadas (Ver formato de la Figura 3) y obtiene una categorización única mediante la heurística majority voting. Esta categorización resume la opinión de todos los expertos que categorizaron y permite ejecutar un ANACOR para establecer relaciones entre las categorías y diversos estratos de la muestra.

En la heurística majority voting se itera por todos los vectores 3D y por cada uno se guardan las categorías que le fueron asignadas en los archivos dados, finalmente, se le elige la categoría con más ocurrencias, en caso de haber más de una categoría con la mayor cantidad de ocurrencias se toma entre ellas la primera que fue encontrada.

RESULTADOS

Tomando en cuenta la revisión del estado del arte, se implementó un software en el lenguaje de programación Python 3.6[13] y sus módulos (e.g. PyQt5;[14] openpyxl[15]), que permitió la categorización semiautomática de los vectores 3D.

A través del software se posibilita la creación del diccionario de categorías, así como su edición y eliminación (Figura 1). El carácter flotante de la ventana del diccionario hace que los psicólogos inmersos en la categorización puedan consultar el diccionario desde la misma aplicación, evitando tenerlo impreso o en formato Microsoft Word, como se hacía en el proceso tradicional.

El software BHPLS data processing-UH®,(16,17) implementado y presentado en este artículo, siguió el enfoque lexicométrico, puesto que permite que los psicólogos elaboren sus categorías y les posibilita la asignación de las asociaciones libres de los sujetos a cada categoría.

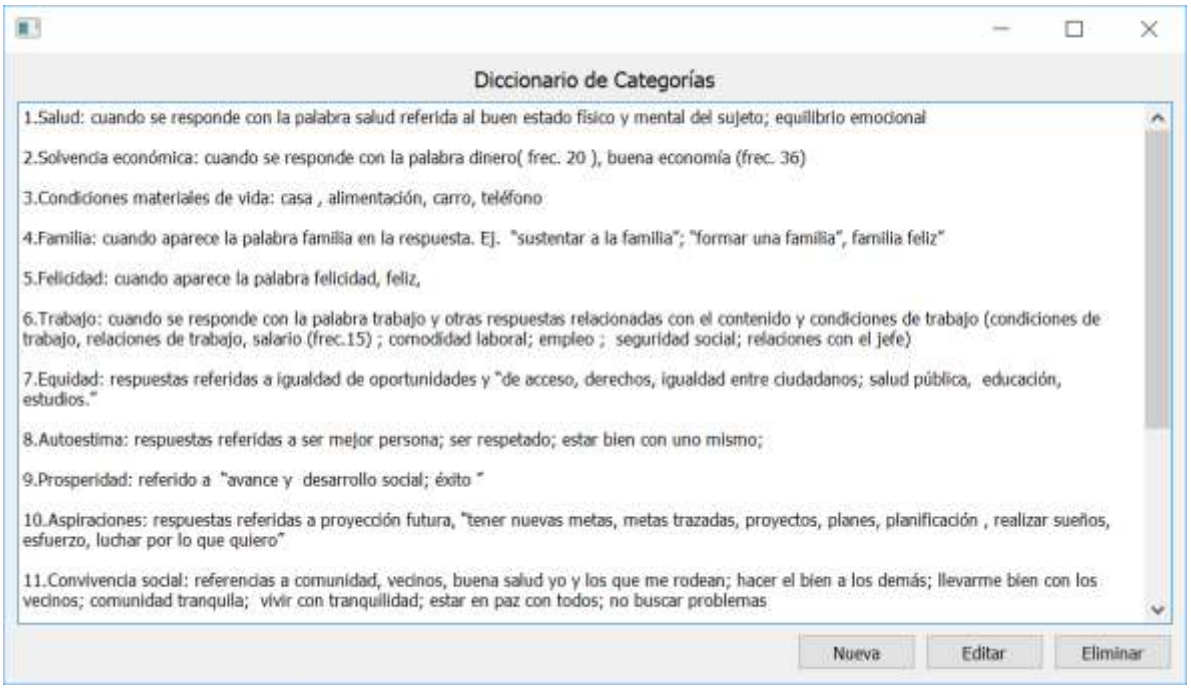
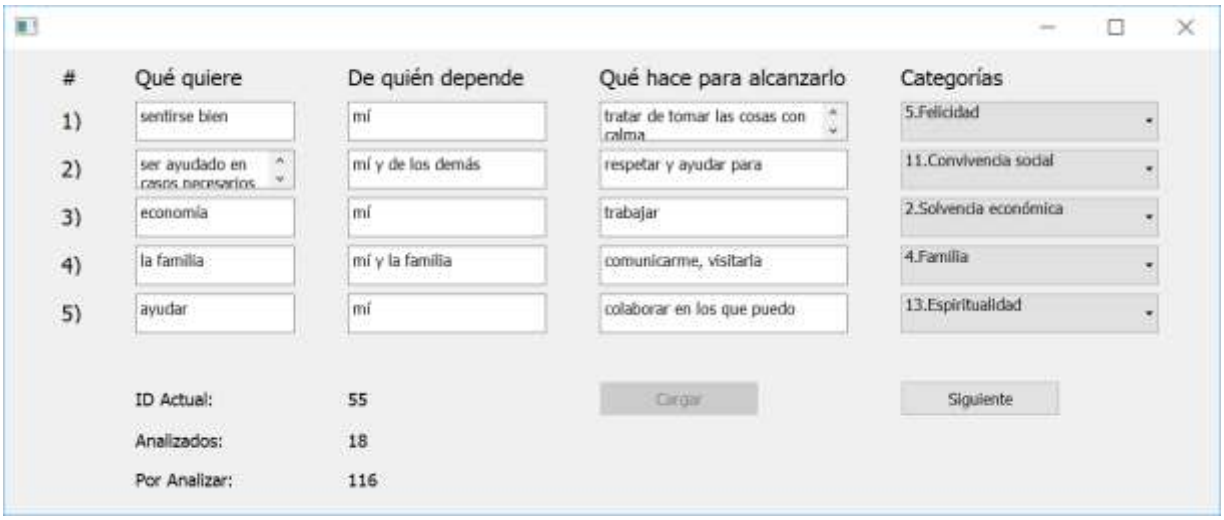


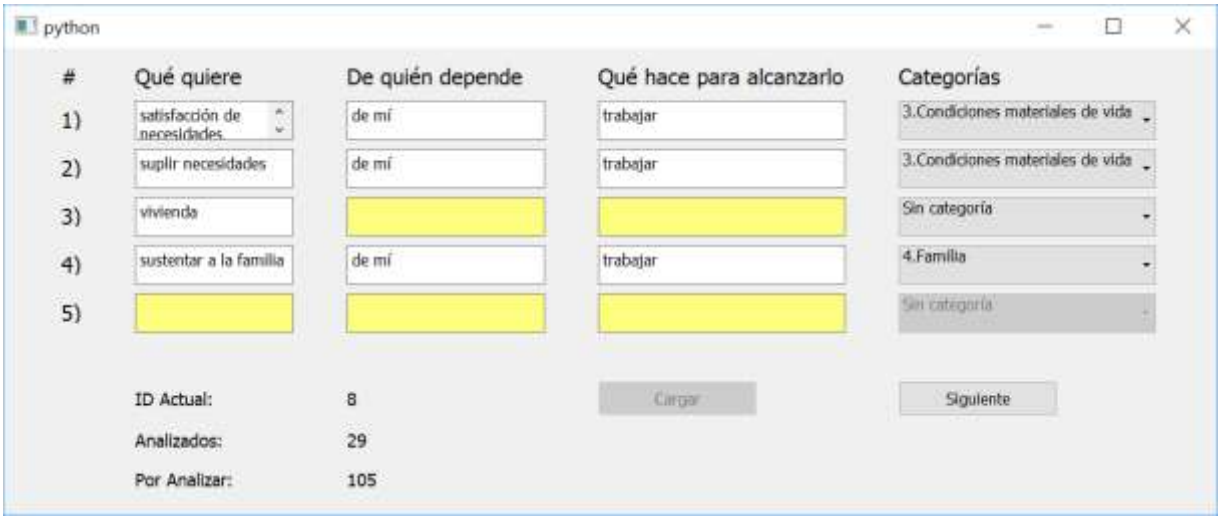
Figura 1. Ventana flotante donde se muestra el esquema de categorías desde la misma aplicación

Entre las funcionalidades de este software está la lectura de las respuestas del cuestionario desde un archivo Excel con una estructura específica. Luego, se muestran los vectores 3D de un sujeto elegido de manera aleatoria y el usuario asigna una categoría del esquema construido a cada vector (Figura 2a). En la Figura 2b puede notarse el sombreado de los datos faltantes. Estos datos pueden estar en todo el vector 3D o en alguno de sus componentes y pueden provocar que la información no sea suficiente para categorizarla. Este proceso se reitera hasta categorizar todos

los vectores 3D de los sujetos en la muestra. Una vez terminada la categorización de todos los vectores 3D, el software crea un archivo Excel donde se guardan los resultados del proceso en una ruta especificada por el usuario (Figura 3). Cabe destacar que los vectores en el Excel creado se muestran ordenados por cada categoría. A lo largo de la categorización se muestran el ID del sujeto analizado, los datos faltantes, cuántos sujetos se han analizados y cuántos faltan por analizar

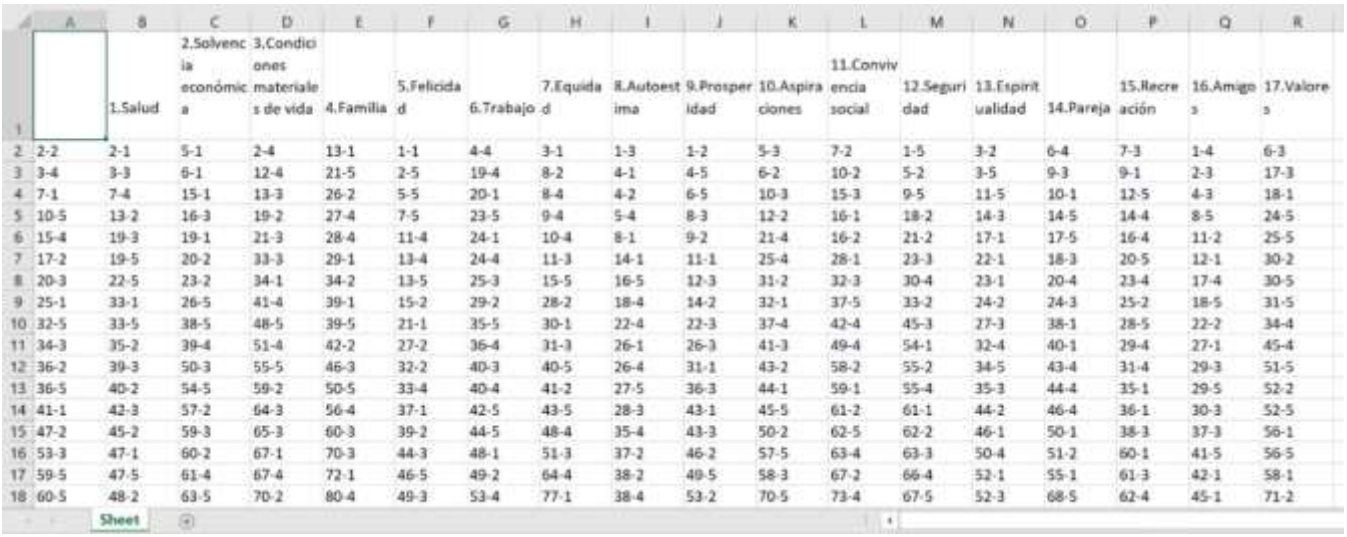


(a)



(b)

Figura 2: 2a- Vista de la ventana de categorización. **2b-** Instante de una categorización con datos faltantes destacados en amarillo



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1		1.Salud	2.Solvencia económica	3.Condiciones materiales de vida	4.Familia	5.Felicidad	6.Trabajo	7.Equidad	8.Autoestima	9.Prosperidad	10.Aspiraciones	11.Convivencia social	12.Seguridad	13.Espiritualidad	14.Pareja	15.Recreación	16.Amigos	17.Valores
2	2-2	2-1	5-1	2-4	13-1	1-1	4-4	3-1	1-3	1-2	5-3	7-2	1-5	3-2	6-4	7-3	1-4	6-3
3	3-4	3-3	6-1	12-4	21-5	2-5	19-4	8-2	4-1	4-5	6-2	10-2	5-2	3-5	9-3	9-1	2-3	17-3
4	7-1	7-4	15-1	13-3	26-2	5-5	20-1	8-4	4-2	6-5	10-3	15-3	9-5	11-5	10-1	12-5	4-3	18-1
5	10-5	13-2	16-3	19-2	27-4	7-5	23-5	9-4	5-4	8-3	12-2	16-1	18-2	14-3	14-5	14-4	8-5	24-5
6	15-4	19-3	19-1	21-3	28-4	11-4	24-1	10-4	8-1	9-2	21-4	16-2	21-2	17-1	17-5	16-4	11-2	25-5
7	17-2	19-5	20-2	33-3	29-1	13-4	24-4	11-3	14-1	11-1	25-4	28-1	23-3	22-1	18-3	20-5	12-1	30-2
8	20-3	22-5	23-2	34-1	34-2	13-5	25-3	15-5	16-5	12-3	31-2	32-3	30-4	23-1	20-4	23-4	17-4	30-5
9	25-1	33-1	26-5	41-4	39-1	15-2	29-2	28-2	18-4	14-2	32-1	37-5	33-2	24-2	24-3	25-2	18-5	31-5
10	32-5	33-5	38-5	48-5	39-5	21-1	35-5	30-1	22-4	22-3	37-4	42-4	45-3	27-3	38-1	28-5	22-2	34-4
11	34-3	35-2	39-4	51-4	42-2	27-2	36-4	31-3	26-1	26-3	41-3	49-4	54-1	32-4	40-1	29-4	27-1	45-4
12	36-2	39-3	50-3	55-5	46-3	32-2	40-3	40-5	26-4	31-1	43-2	58-2	55-2	34-5	43-4	31-4	29-3	51-5
13	36-5	40-2	54-5	59-2	50-5	33-4	40-4	41-2	27-5	36-3	44-1	59-1	55-4	35-3	44-4	35-1	29-5	52-2
14	41-1	42-3	57-2	64-3	56-4	37-1	42-5	43-5	28-3	43-1	45-5	61-2	61-1	44-2	46-4	36-1	30-3	52-5
15	47-2	45-2	59-3	65-3	60-3	39-2	44-5	48-4	35-4	43-3	50-2	62-5	62-2	46-1	50-1	38-3	37-3	56-1
16	53-3	47-1	60-2	67-1	70-3	44-3	48-1	51-3	37-2	46-2	57-5	63-4	63-3	50-4	51-2	60-1	41-5	56-5
17	59-5	47-5	61-4	67-4	72-1	46-5	49-2	64-4	38-2	49-5	58-3	67-2	66-4	52-1	55-1	61-3	42-1	58-1
18	60-5	48-2	63-5	70-2	80-4	49-3	53-4	77-1	38-4	53-2	70-5	73-4	67-5	52-3	68-5	62-4	45-1	71-2

Sheet

Figura 3: Archivo de Excel que constituye la salida del software al terminar una categorización

Validación del software por expertos

Estudiantes de la Facultad de Psicología pusieron a prueba el software categorizando los vectores 3D de 135 sujetos, terminaron la tarea en aproximadamente 1 hora. Los mismos estudiantes habían realizado con anterioridad el proceso manualmente, finalizaron en 3 o 4 días. La semiautomatización de la categorización con el empleo del software reduce significativamente el tiempo del proceso y el usuario no necesita lidiar con las dificultades propias de la lectura de los datos de las encuestas ni de la creación del archivo resultante de la categorización. Estas nuevas facilidades mejoran la experiencia y los resultados del proceso.

Además, la calidad de la categorización aumentó ampliamente al utilizar el software presentado. Se realizaron dos iteraciones de la categorización con estudiantes de psicología obteniendo un índice Kappa negativo en cada ocasión. Al utilizar el software con 675 vectores 3D, 19 categorías y 5 observadores se logró un Kappa global 0.78, error estándar 0.01, IC = [0.75, 0.81] y p = 0.00. (Tabla 1).

Tabla 1: Índices Kappa global y por categorías. Notar que en todos los casos se obtiene una alta concordancia y una alta significación estadística

Categorías	Kappa	IC (95 %)		Estadístico *z*	Valor *p*
Sin categoría	0,79	0,72	0,85	65,11	0,00
Salud	0,89	0,86	0,93	73,93	0,00
Solvencia económica	0,88	0,84	0,91	72,38	0,00
Condiciones materiales de vida	0,83	0,78	0,88	68,70	0,00
Familia	0,91	0,88	0,95	75,27	0,00
Felicidad	0,72	0,57	0,86	59,38	0,00
Trabajo	0,80	0,76	0,85	66,45	0,00
Equidad	0,43	0,27	0,59	35,47	0,00
Autoestima	0,44	0,18	0,69	36,33	0,00
Prosperidad	0,36	0,21	0,51	29,94	0,00
Competente	0,69	0,57	0,81	57,11	0,00
Aspiraciones	0,45	0,28	0,63	37,66	0,00
Convivencia social	0,67	0,60	0,74	55,20	0,00
Seguridad	0,45	-0,32	1,24	37,49	0,00
Espiritualidad	0,57	0,25	0,89	47,39	0,00
Pareja	0,89	0,83	0,96	73,75	0,00
Recreación	0,72	0,58	0,86	59,69	0,00
Amigos	0,88	0,78	0,99	73,04	0,00
Valores	0,54	0,45	0,64	45,13	0,00
Kappa global	0,78	0,76	0,81	205,11	0,00

En la Tabla 2 se muestran las 19 categorías obtenidas, luego de la unificación de categorías, y se desglosa por categoría la cantidad de asociaciones brindadas por los sujetos dentro de cada grupo laboral. Las columnas finales presentan el resultado de aplicar la prueba de independencia[12] y los valores correspondientes. Esta prueba estadística se hace por cada categoría respecto a los tres grupos laborales considerados.

Se hallaron diferencias significativas entre la noción de bienestar de los tres grupos laborales respecto a las categorías solvencia económica, trabajo, condiciones materiales de vida y tranquilidad, la diferencia en las dos últimas categorías es la más acentuada.

Tabla 2: Categorización de las asociaciones libres para la Pregunta 1

Categorización	Cooperativista	Cuentapropista	Estatal	Total	$\chi^2(gl = 2)$	p
Salud	23	26	20	69	0.37	0.82
Solvencia económica	28	18	12	58	4.03	0.13**
Condiciones materiales de vida	26	12	12	50	6.06	0.04*
Familia	10	15	11	36	1.33	0.51
Autoestima	10	10	9	29	0.18	0.91
Felicidad	9	12	6	27	0.99	0.60
Trabajo	8	5	11	24	4.68	0.09**
Tranquilidad	2	12	9	23	8.14	0.01*
Competente	10	5	6	21	1.61	0.44
Equidad	10	6	4	20	1.65	0.43
Amor	4	9	6	19	2.15	0.34
Convivencia social	6	6	5	17	0.02	0.98
Paz	4	9	4	17	2.39	0.30
Prosperidad	7	5	2	14	1.63	0.44
Aspiración	4	6	1	11	2.51	0.28
Seguridad	2	3	3	8	0.59	0.74
Amigos	3	2	3	8	0.55	0.75
Creencia espiritual	1	2	2	5	0.68	0.70

*p<0.05, **p<0.2

En la Figura 4 se muestra el resultado del ANACOR[11] ejecutado en el software Statistica[17] en los datos de la Tabla 2. La proyección en dos dimensiones mostró ser suficiente dado que se explicó el 95.94 % de la variabilidad presente en los datos donde $\chi^2 = 42.90, df = 36, p = 0.1993$.

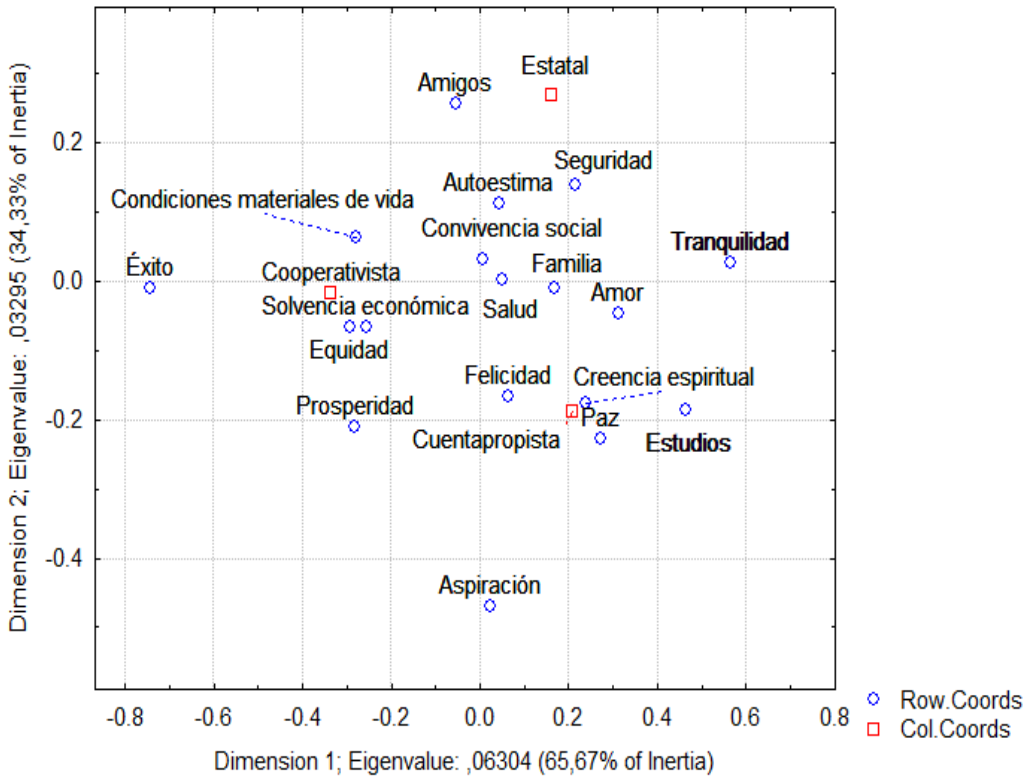


Figura 4: Análisis de correspondencias de los datos de la categorización.

Los resultados que muestra la Figura 4 ofrecen datos que han sido analizados como parte del estudio y otros cuyo análisis desborda los objetivos de este artículo y que brindan nuevas perspectivas para investigaciones futuras.

Los trabajadores cooperativistas se posicionan muy cerca de las categorías solvencia económica, condiciones materiales de vida y equidad. Entre las categorías más lejanas se hallan tranquilidad y estudios. Los cuentapropistas se colocan junto a las categorías felicidad, paz y creencia espiritual. Entre las categorías más lejanas se observan éxito y amigos. Los trabajadores estatales se ubican cerca de las categorías amigos y seguridad. Entre las categorías más lejanas se observan aspiración y éxito. Las categorías salud, familia y convivencia social muestran su importancia para los tres grupos laborales.

Limitaciones del estudio

El software propuesto está implementado a la medida de la estructura de la Pregunta 1 del cuestionario BHPLS, aunque es flexible de adaptarse a otros cuestionarios con preguntas abiertas. El algoritmo para el cómputo del índice Kappa y el método para unificar las categorizaciones dependen de una estructura común para los archivos Excel donde estas se almacenan.

CONCLUSIONES

A pesar de que existe variedad de SADC, se hizo imprescindible la creación de uno que se ajustara al cuestionario BHPLS y al trabajo realizado por los investigadores de la Facultad de Psicología. A partir de la semiautomatización del proceso aportada por el software presentado y el algoritmo propuesto para la unificación de categorías, los investigadores, con muy buenas impresiones en su uso, han llegado a mejores resultados en el estudio que desarrollan.

AGRADECIMIENTOS

Los autores desean agradecer la participación de las profesoras Dra.C. Maiky Díaz Pérez, Dra.C. Daybel Pañellas Álvarez y MSc. Marta Martínez Rodríguez, y los estudiantes Amanda Noriega Rodríguez, Gabriela Díaz Pérez, Surashy de la Caridad García Milán, Andy Luis Marrero Vega y Anabel Rodríguez Conde de la Facultad de Psicología de la Universidad de La Habana por su colaboración en la categorización manual y semiautomática mediante el software propuesto en este artículo. Además, deseamos agradecer al revisor anónimo de este trabajo por sus acertados comentarios y revisión profunda del texto

REREFERENCIAS BIBLIOGRÁFICAS

1. Neuendorf KA. The content analysis guidebook [Internet]. Second edition. 2017 [cited 2019 Mar 8]. 438 p. Available from: https://us.sagepub.com/en-us/nam/the-content-analysis-guidebook/book234078

2. Jenny J. Méthodes et pratiques formalisées d'analyse de contenu et de discours dans la recherche sociologique française contemporaine. Etat des lieux et essai de classification. Bull Méthodologie Sociol. 1997;54:64–112.

3. Lahlou S. A method to extract social representations from linguistic corpora. Japanese J Exp Soc Psychol. 1996;35(3):278–391.

4. de Rosa AS. The associative network: A technique for detecting structure, contents, polarity and stereotyping indexes of the semantic fields. Eur Rev Appl Psychol. 2002;3–4:81–200.

5. Woolf NH, Silver C. Qualitative analysis using ATLAS.ti: the five-level QDA method [Internet]. Routledge; [cited 2019 Mar 8]. 194 p. Available from: https://www.routledge.com/Qualitative-Analysis-Using-ATLASti-The-Five-Level-QDA-Method/Woolf-Silver/p/book/9781138743656

6. Woolf NH, Silver C. Qualitative analysis using nvivo: The five-level QDA® method. Qualitative Analysis Using NVivo: The Five-Level QDA Method. London: Routledge, 2017 [cited 2019 Mar 8]. Available from: https://www.routledge.com/Qualitative-Analysis-Using-NVivo-The-Five-Level-QDA-Method/Woolf-Silver/p/book/9781138743670

7. Noriega Rodríguez A. Cuestionario que estudia el bienestar humano como constructo multidimensional. Caso de Estudio: Relación entre Bienestar y Trabajo. Trabajo de Diploma presentado en opción al título de Licenciado en Psicología. Tutores: Maiky Díaz Pérez, Marta Martínez Rodríguez y Damian Valdés Santiago. Universidad de La Habana; 2018.

8. Díaz-Pérez M, Valdés-Santiago D. Desarrollo del potencial humano y bienestar en el trabajo. In:

Ferreira M, editor. Intensificação, Precarização, Esvaziamento do Trabalho e Margens de Enfrentamento, "GT Trabalho e Saúde'''." Associação Nacional de Pesquisa e Pós-Graduação (ANPEPP), Brasil; 2018.

9. Krippendorff KH. Content Analysis: An Introduction to Its Methodology. 4th ed. Los Angeles: SAGE, 2019. Available from: https://us.sagepub.com/en-us/nam/content-analysis/book258450

10. Tang W, Hu J, Zhang H, Wu P, He H. Kappa coefficient: a popular measure of rater agreement. Shanghai Arch psychiatry [Internet]. 2015 Feb 25 [cited 2019 Mar 8];27(1):62–7. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4372765

11. Greenacre M. Correspondence Analysis in Practice. 3rd edition. Boca Raton: CRC Press, Taylor & Francis Group, 2017. [cited 2019 Mar 8] Available from: https://www.crcpress.com/Correspondence-Analysis-in-Practice/Greenacre/p/book/9781498731775

12. Fleiss JL, Levin B, Paik MC, Fleiss J. Statistical Methods for Rates & Proportions. 3rd ed. New York: Wiley-Interscience; 2014. [cited 2019 Mar 8] Available from: https://www.wiley.com/en-cu/Statistical+Methods+for+Rates+and+Proportions%2C+3rd+Edition-p-9780471526292

13. Romano F. Learn Python programming: a beginner's guide to learning the fundamentals of Python language to write efficient, high quality code [Internet]. Second edition. 2018 [cited 2019 Mar 8]. Available from: https://www.packtpub.com/application-development/learn-python-programming-second-edition

14. Summerfield M. Rapid GUI programming with Python and Qt : the definitive guide to PyQt programming [Internet]. [Place of publication not identified]: Prentice Hall; 2015 [cited 2019 Mar 8]. Available from: https://www.oreilly.com/library/view/rapid-gui-programming/9780132354189/

15. Gazoni E, Clark C. openpyxl: A Python library to read/write Excel 2010 xlsx/xlsm files. 2018.

16. Guerrero-Oliva JC, Valdés-Santiago D. Software para el Análisis de Contenido de nociones de bienestar en una aplicación del cuestionario BHPLS. En: X Encuentro de Estudiantes de Psicología. La Habana, Cuba: ISBN 978-959-16-3917-2; 2018.

17. Oliva Guerrero JC, Valdés Santiago D. BHPLS data processing-UH®. Número de registro 2984-09-20, Centro Nacional del Derecho de Autor, La Habana, Cuba, 2018.

18. StatSoft, I. (2017). STATISTICA (data analysis software system), version 8.0.

Conflicto de intereses

Los autores declaran no tener ningún conflicto de intereses.

Contribución de autoría

Todos los autores participamos en la discusión de los resultados y hemos leído, revisado y aprobado el texto final del artículo.