

# Análisis de la Disponibilidad Léxica a través Clustering

Darío Rojas D.  
Universidad de Concepción  
Facultad de Educación  
Edmundo Larenas 335  
56-41 2203489  
dariorojas@udec.cl

Carolina Zambrano M.  
Universidad de Concepción  
Facultad de Educación  
Edmundo Larenas 335  
56-41 2203489  
carozambrano@udec.cl

Pedro Salcedo L.  
Universidad de Concepción  
Facultad de Educación  
Edmundo Larenas 335  
56-41 2203489  
psalcedo@udec.cl

## ABSTRACT

In any teaching-learning environment it is produced a communication process that involves retrieving from memory a specific vocabulary with which participants can perform the interaction. In this paper, the analysis of lexical availability for students from two teaching programs from two Chilean universities is presented, proposing a clustering approach for the lexicon analysis of students by applying the k-medoids algorithm and using the Levenshtein distance as a measure. The results show the feasibility of performing a clustering analysis using the proposed approach, obtaining observation elements directly from the characteristics between the lexicons of students, checking in this way the relationship between the lexical availability of students and other descriptive features as the number of university years, university of origin and gender.

## RESUMEN

En todo ambiente de enseñanza-aprendizaje se produce un proceso de comunicación que involucra recuperar de la memoria un vocabulario específico con el cual los participantes puedan llevar a cabo la interacción. En este trabajo, se presenta el análisis de disponibilidad léxica para alumnos de dos carreras de pedagogía de dos universidades chilenas, proponiendo un enfoque de clustering para el análisis de los lexicones de los estudiantes mediante la aplicación del algoritmo k-medoids y utilizando como medida la distancia de Levenshtein. Los resultados muestran la factibilidad de realizar un análisis de clustering mediante el enfoque propuesto, obteniendo elementos de observación directamente de las características entre los lexicones de los estudiantes, comprobando de ésta forma la relación que existe entre la disponibilidad léxica de los alumnos y otros rasgos descriptivos como la cantidad de años en la universidad, universidad a la que pertenece y género.

## CCS Concepts

• **Information systems~Clustering and classification**  
• *Information systems~Document representation* • **Applied computing~Education** • *Applied computing~Document analysis* • *Computing methodologies~Cluster analysis* • **Social and professional topics~Student assessment**

## Keywords

Lexical Analysis, Clustering, k-medoids, Levenshtein Distance.

## 1. INTRODUCCIÓN

En cualquier ambiente de enseñanza-aprendizaje se produce un proceso de comunicación alumno-profesor llamado interacción pedagógica. Este proceso es de especial relevancia, ya que permite a un docente o sistema interactivo de enseñanza, actuar como elemento clave y aclaratorio del aprendizaje [1,2].

Evidentemente, los estudiantes manejan de forma natural cierto “lenguaje” que les permite comunicarse con otras personas, pero debido a que la interacción pedagógica ocurre en un contexto mucho más específico de comunicación, es posible que el alumno necesite un “lenguaje” mucho más especializado para aprender de forma adecuada, ya sea desde materiales de enseñanza o directamente de los profesores. En este sentido, si el lenguaje o vocabulario del material, profesor o alumno es deficiente, la interacción pedagógica podría no ocurrir y por lo tanto afectar negativamente el aprendizaje.

En este mismo contexto, una forma eficiente de obtener una caracterización del vocabulario disponible de los alumnos para llevar a cabo el proceso de comunicación en un área específica del saber, es a través del análisis de la disponibilidad léxica de los alumnos [4]. Los estudios de la disponibilidad léxica nacieron de un proyecto de la UNESCO en la década de los cincuenta, donde se perseguía facilitar el aprendizaje de la lengua francesa a los habitantes no nativos de Francia, buscando así, facilitar la integración social a partir de la simplificación de una lengua común base [5]. La disponibilidad léxica refleja el caudal léxico utilizado en una situación comunicativa, donde ciertas palabras muy usadas en una lengua están estrechamente relacionadas con conceptos aparecidos en las interacciones de comunicación. Para obtener el léxico disponible se realizan pruebas de asociación con estímulos o centros de interés, lo que da como resultado lo que se supone es el vocabulario activo de los sujetos en torno a esos estímulos. Bajo estas premisas, el estudio del léxico disponible supone que son más disponibles aquellas palabras que primero se recuperan de la memoria ante un estímulo. Es así como una de las formas más comunes de obtener el léxico disponible de un conjunto de personas entorno a un centro de interés, puede ser llevado a cabo solicitando a las personas escribir en orden las palabras que primero asocien con un estímulo, considerando comúnmente un tiempo límite para realizar tal tarea. Ese diccionario mental de palabras asociadas a un concepto es denominado lexicón mental del individuo.

El presente trabajo presenta un estudio sobre el análisis de la disponibilidad léxica de alumnos de la carrera de pedagogía en matemática de dos universidades chilenas, proponiendo el uso de herramientas de análisis de clusters para caracterizar la relación de los lexicones mentales de los estudiantes.

## 2. ANÁLISIS DE DISPONIBILIDAD LÉXICA

El análisis de disponibilidad léxica se centra en la obtención de los lexicones (conjunto de palabras más disponibles de un individuo sobre un centro de interés específico) [3,10]

Para nuestro estudio, se realizó la prueba de disponibilidad léxica para 126 alumnos pertenecientes a las carreras de Pedagogía en Matemática y Computación de la Universidad de Concepción (UdeC) y a la carrera de Pedagogía en Matemática de la Universidad del Bío-Bío (UBB), considerando en ambos casos alumnos de primero a quinto año. El instrumento de recolección de datos corresponde al Test de Disponibilidad léxica empleado por Valencia y Echeverría [12]. Los centros de interés seleccionados corresponden a la agrupación por temática de los 21 estándares disciplinares que determina el Ministerio de Educación de Chile para la enseñanza de las matemáticas en educación media [6].

### 3. ENFOQUE DE ANÁLISIS DE CLUSTERS

Desde el punto de vista de análisis de datos, el estudio y análisis de la disponibilidad léxica trata mayormente sobre el estudio de las características grupales de la información, y por lo tanto corresponde a un tipo de análisis de clusters de lexicones y vocablos, lo que hace relevante para estos estudios el poder caracterizar a los grupos. Sin embargo, la definición de los grupos generalmente se ha realizado a través de la determinación de estadígrafos, formando los grupos tomando en consideración alguna característica adicional como el sexo, lugar de procedencia, etc. Debido a que los lexicones están compuestos de palabras, y que el mismo lexicon puede ser considerado como una palabra donde cada vocablo representa un símbolo, es posible llevar a cabo la propuesta, utilizado la distancia de Levenshtein [14] y el algoritmo de clustering k-medoids [13]. Sin embargo, este enfoque hace difícil la evaluación de la calidad de los resultados, debido a que índices de validación de clustering tradicionales tales como los utilizados en [8], no pueden ser aplicados a los lexicones. Es por eso que para esta investigación se ha utilizado el índice de cohesión (IC) utilizado comúnmente para medir la homogeneidad del léxico disponible de un grupo con tal de determinar la calidad de las agrupaciones.

### 4. RESULTADOS

A continuación, se muestran los resultados del análisis de disponibilidad léxica mediante el modelo análisis de clústeres propuesto. Cabe notar que cada vocablo o palabra que aparece como parte de un lexicon no contiene tildes, ya que estos han sido eliminados en el procesamiento de la información. Además, algunas palabras han sido truncadas para hacerlas corresponder con el espacio disponible para desplegarlas al interior de gráficos y tablas.

#### 4.1 Estadígrafos

En la Tabla 1, se muestran los estadígrafos por cada centro de interés. Como se puede apreciar, el centro de interés Estructuras Algebraicas es el que tiene una menor cantidad de vocablos por lexicon en promedio (XR), al contrario, Geometría es el que más vocablos en promedio posee, y ambos se ven reflejados en la cantidad de palabras distintas que posee cada uno (NPD). Por otro lado, comparando Datos y Azar con Geometría, se puede ver que a pesar de que Datos y Azar tiene una mayor cantidad de palabras distintas, Geometría igualmente tiene un XR superior, haciendo sentido al índice de cohesión (IC) de este centro que es el doble de los otros centros de interés, indicando que, en comparación a los otros centros de interés, este centro está altamente cohesionado y existe una homogeneidad en el léxico disponible de los alumnos respecto a este tema.

**Tabla 1. Estadígrafos por centro de interés**

<i>Centro de Interés</i>	<i>N</i>	<i>XR</i>	<i>NPD</i>	<i>IC</i>
Datos y Azar	126	12.77	483	0.0264
Cálculo	126	13.78	529	0.0260
Estructuras Algebraicas	123	10.16	372	0.0273
Geometría	126	19.94	423	0.0471
Sistemas Numéricos	126	13.05	455	0.0287

#### 4.2 Índices de Disponibilidad Léxica

Las Tablas 2 y 3 muestra el listado de las primeras 20 palabras con índice de disponibilidad léxica (IDL) más alto ordenadas de mayor a menor. Como se puede observar, en el centro de interés Datos y Azar, existen vocablos de alta disponibilidad como 'dado', 'juego' y 'suerte' que pueden asociarse más a juegos de azar que al lenguaje técnico propiamente de dicha temática. Por ejemplo, el vocablo 'carta' aparece como una de las 20 más disponibles, haciendo claro que los estudiantes tienen altamente disponibles palabras asociadas a juegos de azar y por lo tanto sería plausible utilizar dicho tipo de ejemplos para poder maximizar la eficiencia del proceso comunicativo en estos temas. Por otro lado, se puede ver en los demás centros de interés, que el lenguaje es más propio de cada área y es común ver que al menos uno de los vocablos que componen el nombre del centro de interés tales como 'azar', 'calculo', 'algebra', 'geometría' y 'numero', aparezcan con un índice de alta disponibilidad. Por otro lado, vocablos como 'probabilidad', 'función', 'anillo', 'reales', etc., tienen directa relación con los contenidos y el vocabulario específico de cada área.

**Tabla 2. Índices de Disponibilidad Léxica por cada Centro de Interés.**

<b>Datos y Azar</b>	<b>IDL</b>	<b>Cálculo</b>	<b>IDL</b>	<b>Estruct. Alg.</b>	<b>IDL</b>
probabilidad	0,6531	derivada	0,5113	Anillo	0,3939
Estadística	0,3405	Limite	0,4611	Grupo	0,2943
Dado	0,2792	integral	0,4126	Algebra	0,2856
Dato	0,2522	Función	0,2085	Cuerpo	0,2047
Moda	0,2369	Numero	0,2066	Estructura	0,1651
Mediana	0,2055	Infinito	0,1165	Conjunto	0,1392
Azar	0,2054	Teorema	0,1153	Demostración	0,1387
Media	0,1968	calculo	0,1095	Letras	0,1027
porcentaje	0,1369	suma	0,1065	matriz	0,0988
Juego	0,1290	continuidad	0,1045	abeliano	0,0916
Promedio	0,1267	diferencial	0,1020	campo	0,0891
Grafico	0,1118	multiplicacion	0,0931	numeros	0,0798
Razón	0,1088	variable	0,0906	grupoide	0,0792
Muestra	0,1083	division	0,0828	conmutativida	0,0775
Moneda	0,1009	area	0,0784	propiedad	0,0773
frecuencia	0,0996	resta	0,0775	operacion	0,0760
Varianza	0,0942	analisis	0,0772	axioma	0,0759
Suerte	0,0923	adicion	0,0746	orden	0,0750
Numero	0,0914	volumen	0,0641	ecuacion	0,0700
Carta	0,0890	demostracion	0,0630	teorema	0,0688

**Tabla 3. Índices de Disponibilidad Léxica por cada Centro de Interés (continuación)**

Geometría	IDL	Sis. Numéricos	IDL
Triangulo	0,4900	numero	0,6213
Angulo	0,3387	ecuacion	0,3529
circunferencia	0,3183	incognita	0,2060
Recta	0,2797	suma	0,1719
Cuadrado	0,2763	reales	0,1472
Área	0,2434	letra	0,1466
Figura	0,1947	resta	0,1395
Perímetro	0,1725	multiplicacion	0,1335
Rectángulo	0,1656	algebra	0,1239
Euclides	0,1465	sistema	0,1230
Punto	0,1420	naturales	0,1193
Lado	0,1288	complejos	0,1113
Volumen	0,1231	conjunto	0,0997
Geometría	0,1183	division	0,0967
Circulo	0,1176	enteros	0,0928
Semejanza	0,1160	variable	0,0896
Plano	0,1150	simbolo	0,0892
congruencia	0,1148	matriz	0,0891
Teorema	0,1044	operacion	0,0883
Cuerpo	0,1025	racionales	0,0820

### 4.3 Clustering de Lexicones

Para esta parte del estudio se realizó un análisis de clusters por cada centro de interés. A cada uno de ellos, se le aplicó el algoritmo k-medoids utilizando como medida de disimilitud la distancia de Levenshtein entre los lexicones. Los resultados se muestran en la Tabla 4, donde por cada centro de interés se han obtenido de 2 a 5 agrupaciones (k) en conjunto con el índice de cohesión promedio de todos los clusters (IC Prom.) y la cantidad de elementos contenidos en cada clúster (N).

**Tabla 4. Resultados de Clustering para cada Centro de Interés**

	k	IC Prom.	N1	N2	N3	N4	N5
Azar	2	0,1251	116	10			
	3	0,1342	104	10	12		
	4	0,1390	14	91	12	9	
	5	0,1514	25	46	37	7	11
Calc.	2	0,1063	39	87			
	3	0,1110	75	37	14		
	4	0,1194	16	5	2	103	
	5	0,1264	17	78	5	22	4
Est.	2	0,1071	104	19			
	3	0,1230	59	56	8		
	4	0,1274	11	34	28	50	
	5	0,1318	3	15	11	33	61
Geo.	2	0,1760	42	84			
	3	0,1832	57	62	7		
	4	0,1908	53	35	28	10	

	5	0,1953	85	9	7	22	3
Num.	2	0,1125	108	18			
	3	0,1228	107	15	4		
	4	0,1240	92	8	2	24	
	5	0,1409	84	5	26	3	8

Como se aprecia en la Tabla 4, todas las mejores configuraciones, considerando a IC como medida de cohesión interna de los clústers, contemplan la generación de la máxima cantidad de clústeres ( $k=5$ ), siendo los centros de interés Datos y Azar, y Geometría los con mayores índices de cohesión interna. Esto hace mucho sentido, debido a que cuando se intentan determinar una menor cantidad de agrupación, el análisis encuentra que son pocos los lexicones que se asemejan entre ellos, dando lugar a configuraciones con grupos de pocos elementos versus uno o dos grupos con muchos elementos. Lo anterior, puede ser debido a la alta heterogeneidad de las respuestas de los alumnos, encontrando una concordancia en los resultados con respecto a los estadígrafos mostrados en la Tablas 2, donde se aprecia una alta diferencia en la cohesión del centro de interés de Geometría, el que se ve reflejado en la cohesión de los clusters para todos los valores de  $k$ .

Por otro lado, en la Tabla 5, se puede apreciar una caracterización de las configuraciones encontrada para cada centro de interés incluyendo información de los grupos como el Promedio de Notas (Prom. Nota) que corresponde al promedio de notas de un alumno sobre todas las asignaturas que tratan los ejes temáticos correspondientes a los todos los centros de interés. Además, se incluye una contabilización de la cantidad de hombres (Hombres), cantidad de mujeres (Mujeres), promedio de años en la universidad (Prom. Años), cantidad de alumnos de la Universidad del Bio-Bio (UBB) y cantidad de alumnos de la Universidad de Concepción (UdeC).

A continuación, se presenta un resumen del análisis por cada centro de interés considerando  $k=5$  según los resultados de la Tabla 5:

- *Datos y Azar:* Se puede observar una clara diferencia en la pertenencia a distintas universidades, donde el cluster 2 y 4 tienen una alta concentración de alumnos de la UdeC con los promedios de notas más altos, salvo el cluster 3 que tiene el promedio de notas más alto de todos, pero con pocos estudiantes, por lo que se puede interpretar como que los alumnos de la UdeC tienen un mayor promedio de notas que los alumnos de la UBB. Por otro lado, respecto al promedio de años en la universidad se puede ver que los clusters 2 y 3 también marcan una clara diferencia en este indicador y están en concordancia a la pertenencia a las distintas universidades. Por otro lado, en el contexto del género de los alumnos, no se aprecia diferencias muy grandes en las agrupaciones.
- *Cálculo:* En este centro de interés se ve un comportamiento similar en el cluster 2 al mostrado en general para el centro de interés Datos y Azar, encontrando agrupaciones que principalmente diferencian la pertenencia a la universidad.
- *Estructuras Alg.:* En el cluster 2 de este centro de interés se puede apreciar también una concentración de alumnos de la UdeC, salvo que en este caso se aprecia una diferencia de género más notable en conjunto a una menor diferencia en el promedio de notas de los grupos. Además, se puede observar que el grupo 4 tienen una gran diferencia en el promedio de años en la universidad respecto a los otros grupos.

- **Geometría:** En este caso, se puede observar un comportamiento acorde a los otros centros de interés con una concentración de alumnos UdeC en el cluster 2. Sin embargo, el cluster 3 muestra una tendencia contraria, con una alta concentración de alumnos UBB. Además, se puede observar, que el cluster 4 está conformado sólo por dos alumnos, lo que se debe a una alta similitud entre sus respuestas y una gran cantidad de vocablos en cada uno alejándolos del promedio de vocablos por respuesta de los demás estudiantes.
- **Sis. Numéricos:** En este centro de interés, salvo el cluster 3 que se diferencia por tener una baja cantidad de estudiantes, no se observan grandes diferencias en sus otras características.

**Tabla 5. Caracterización de las configuraciones de clusters por centro de interés con  $k=5$ .**

	Cluster	Prom Nota	Mujeres	Hombres	Prom. Años	UdeC	UBB
Azar	1	4,69	3	4	2,14	4	3
	2	5,05	5	7	3,58	11	1
	3	5,48	3	1	3,25	3	1
	4	5,08	10	16	2,42	21	5
	5	4,71	44	33	2,00	40	37
Cálculo	1	4,77	46	41	2,06	52	35
	2	5,09	15	14	2,76	24	5
	3	4,72	2	3	2,60	1	4
	4	5,05	2	0	4,00	2	0
	5	4,70	2	1	2,67	0	3
Estructuras	1	4,87	35	30	2,26	44	21
	2	5,10	8	14	3,55	18	4
	3	5,15	1	1	2,00	1	1
	4	4,57	17	14	1,58	12	19
	5	5,17	1	2	2,33	3	0
Geometría	1	4,76	4	1	2,60	3	2
	2	4,95	44	39	2,34	63	20
	3	4,55	15	12	2,11	4	23
	4	4,50	1	1	1,00	1	1
	5	4,86	3	6	2,44	8	1
Sis.	1	4,99	13	17	2,60	19	11
	2	4,72	29	20	2,18	26	23
	3	4,90	2	1	1,67	1	2
	4	4,86	20	16	2,14	26	10
	5	4,96	3	5	2,63	7	1

Respecto a la disponibilidad de los vocablos por clusters, la Tabla 6 muestra cada centro de interés y una lista de las cinco palabras de más alto IDL ordenadas de mayor a menor para cada cluster (con  $k=5$ ). En esta tabla se puede apreciar de otra forma la diferenciación en algunas de las configuraciones. A continuación, un breve análisis por cada centro de interés:

- **Datos y Azar:** Como se puede observar, los grupos establecen casi los mismos vocablos altamente disponibles del centro de interés como lo son: ‘probabilidad’, ‘estadística’ y ‘dado’. Sin

embargo, existe una diferencia en el cluster 2, donde es más disponible el vocablo ‘dado’, en desmedro de probabilidad y estadística. Al igual, el cluster 3 no posee como vocablo altamente disponible la palabra ‘probabilidad’. Esto coincide con lo presentado en la Tabla 6 donde estos grupos tienen una mayor cantidad de alumnos UdeC, pero sobre todo un promedio de años en la universidad mayor a los otros grupos, lo que lleva a pensar que las primeras palabras de alta disponibilidad pueden ser distintas según la universidad de procedencia.

- **Cálculo:** Aquí se puede ver, en concordancia a la Tabla 7, que el único cluster distinguible en sus características es el cluster 2, el que posee como vocablo altamente disponible a la palabra ‘teorema’, en coincidencia con una mayor pertenencia de alumnos UdeC según Tabla 6.
- **Estructuras Alg.:** En este centro de interés se puede destacar la aparición el vocablo ‘matriz’ en el cluster 4, que se relaciona con las materias del algebra lineal en conjunto a que el grupo es el de promedio más bajo respecto a años en la universidad. Según esto, se puede determinar que el grupo podría estar en su mayoría formado por alumnos de los primeros años.
- **Geometría:** El cluster 3 contiene como vocablo altamente disponible a la palabra ‘espacio’, y en relación a lo presentado en la Tabla 6, este grupo tiene una mayor concentración de alumnos UBB, lo que indicaría una diferencia notable, ya que ‘espacio’ es una palabra de baja disponibilidad general en el centro de interés, pero altamente concentrada en un subgrupo de alumnos de la UBB.
- **Sis. Numéricos:** En este centro de interés se puede destacar la diferencia del tipo de vocablos del cluster 1, los que hacen referencia a las operaciones elementales y con relativa baja disponibilidad respecto a otros vocablos más relacionados con el centro de interés como los son ‘sistemas’, ‘reales’, ‘complejos’, etc. Este es un resultado del que no se puede concluir mucho debido a la homogeneidad de las características presentadas para este cluster en la Tabla 6.

**Tabla 6. Vocablos más disponibles para la mejor configuración de clusters por centro de interés.**

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Datos y Azar	Dado	Dato	Moda	probabilidad	probabilidad
	Probabilidad	Azar	estadística	moda	estadística
	Dato	Probabilidad	Mediana	mediana	dado
	Moneda	Población	Media	estadística	dato
	estadística	Media	Muestra	media	juego
Cálculo	Derivada	Derivada	Derivada	integral	limite
	Limite	Limite	Limite	derivada	analisis
	Integral	Integral	Integral	funcion	derivada
	Numero	Función	multiplica	diferencial	calculo
	Función	Teorema	Épsilon	numero	coronel
Estructuras	Anillo	Grupo	Demostración	algebra	algebra
	Algebra	Anillo	propiedad	funcion	estructura
	Cuerpo	Cuerpo	Plano	matriz	cuerpo
	Grupo	conmutativa	Difícil	estructura	anillo

	conjunto	campo	Polar	conjunto	Myriam
Geometría	triangulo	triangulo	triangulo	geometria	Angulo
	angulo	circunferencia	Figura	triangulo	Triangulo
	rectangulo	angulo	area	angulo	Cuadrado
	euclides	Recta	espacio	cuadrilatero	Punto
	Circulo	cuadrado	plano	calcular	Recta
Sis. Numéricos	numero	ecuacion	reales	numero	Numero
	Suma	numero	ecuacion	incognita	Sistema
	Resta	incognita	trigonome	ecuacion	Base
	multiplicacion	reales	binomio	sistema	Algebra
	ecuacion	complejos	numero	simbolo	Operación

Por último, la Tabla 7 muestra hasta 10 lexicones, con un máximo de 7 vocablos para cada uno de los clusters generados en el proceso de clustering. Los lexicones fueron obtenidos considerando la primera palabra como de alta disponibilidad según su IDL en el centro de interés. Por extensión, y debido a que las conclusiones sobre el proceso de clustering son generalizables a los otros centros de interés, sólo se muestran los centros de interés Datos y Azar y Estructuras Alg. Se debe aclarar, que el análisis mostrado a continuación, tiene por objetivo validar la agrupación de lexicones similares, y no tiene por objetivo concluir respecto a las razones de su conformación, pudiendo hacer las siguientes observaciones:

- *Datos y Azar:* Como se observa, existe para todos los grupos una alta similitud en el orden en que están los primeros vocablos respondidos por los alumnos (elementos más a la izquierda), pudiendo encontrar concordancias incluso en el 5to. y 6to. vocablo.
- *Estructuras Alg.:* Se puede ver, que al igual que Datos y Azar, hay una alta similitud en los lexicones y su orden por cada cluster. Además, aquí se puede corroborar que el cluster 4, que posee un bajo promedio en años en la universidad, también tienen poco vocabulario disponible, debido a que sus lexicones cuentan con una baja cantidad de vocablos (pocas palabras respondidas por los alumnos). Por otro lado, se puede observar la alta frecuencia del vocablo matriz, en diferencia a los otros clusters donde prácticamente no aparece.

Tabla 7. Ejemplo de miembros de clusters generados con  $k=5$

	C	Elementos del Cluster (un lexicon por fila)						
Datos y Azar	1	probabilidad	carta	dado	moneda	suerte	ocurrir	posible
		dado	estadistico	web	busqueda	buscar	conocer	preguntar
		dado	carta	dado	moneda	suerte	probabilidad	azar
		dado	probabilidad	ocurrencia	azar	integral	tabla	posicion
		estadistica	dato	variable	tabla	grafico	cuantitativa	cuantitativa
	mediana	media	moda	promedio	conjuntoded	grafico	tabladedato	
	probabilidad	dado	estadistica	aleatorio	evento	moneda	lanzar	
	2	poblacion	muestra	dato	probabilidad	frecuencia	intervalodec	media
		dato	azar	modelo	estadigrafo	tabla	variable	dependencia
		dato	azar	estadistica	regresion	lineal	intervalo	formula
dato		azar	probabilidad	fraccion	denominado	numerador	pelota	
dato		azar	juegomenta	logica	razon	abstrccion	pensamientc	
dato	azar	regresion	intervalo	confianza	normalidad	varianza		
probabilidad	estadistica	dato	azar	histograma	graficodebai	muestra		
probabilidad	porcentaje	tabla	numero	dato	frecuencia	razon		
probabilidad	muestra	poblacion	registro	cantidad	frecuencia	relativo		
promedio	desviaciones	varianza	distribucion	distribucion	dado	moneda		

Estructuras Alg.	3	estadistica	muestra	tendencia	moda	mediana	promedio	poblacion
		estadistica	inferencia	descriptiva	moda	mediana	media	varianza
		datosazar	media	mediana	moda	estadistica	probabilidad	dfisher
	4	dado	suerte	propabilidad	media	mediana	promedio	cota
		dado	moneda	probabilidad	estadistica	media	promedio	mediana
		dato	azar	probabilidad	dado	moneda	carta	juego
		dato	azar	probabilidad	estadistica	media	moda	varianza
		dato	azar	aleatorio	muestra	fraccion	proporcion	porcentaje
		datoyazar	probabilidad	estadistica	incognita	moda	mediana	promedio
		estadistica	probabilidad	suceso	evento	aleatorio	discreto	continio
estadistica		probabilidad	media	mediana	moda	frecuencia	absoluto	
estadistica		estadisticadi	dado	probabilidad	casino	loto	datosagrupa	
estadistica		probabilidad	permutacion	combinacion	moda	media	mediana	
estadistica	promedio	meda	mediana	moda	probabilidad	dado		
5	dado	juego	casino	moneda	azar	probabilidad	regladetres	
	dado	media	mediana	exel	grafico	mediaaritm	promedio	
	dado	porcentaje	fraccion	posibilidad	pascal	juego	suerte	
	dado	probabilidad	azar	calculo	sorteo	muestra	poblacion	
	dado	probabilidad	dato	frecuencia	absoluta	bola	naipe	
	dato	azar	estadistico	moda	media	promedio	frecuencia	
	dato	azar	probabilidad	estadistica	moda	mediana	frecuencia	
	dato	azar	probabilidad	estadistica	fraccion	percentage	numero	
	dato	azar	probabilidad	secuencia	razon	proporcion	oportunidad	
	dato	azar	probabilidad	suerte	aleatorio	prediccion	casosposible	
dato	cifra	probabilidad	proyeccion	numero	dado	orden		
1	anillo	cerrado	independien	dependiente	algebra	demostracio	cuerpo	
	anillo	cuerpo	dificil	teorema	proposicion	complicada		
	anillo	cuerpo	grupo	orden	conmutativi	neutro		
	anillo	cuerpo	monomio	unitario	distributivid	espacio	vectorial	
	anillo	espacio	cuerpo	grupoide	adicion	sumatoria	vector	
	anillo	grupo	abeliano	subgrupo	cuerpo	grupoide	cerrado	
	anillo	grupo	subgrupo	abeliano	ideal	lci	lce	
	anillo	grupoide	abeliano	cuerpo	clausura	aplicacion	transformac	
	conjunto	cuerpo	anillo	abeliano	subcuerpo	subanillo	modulo	
	conjunto	racional	ponderacion	notacion	auxiliares	inecuacion	determinant	
conjunto	relacion	producto	adicion	asimetria	reflexividad	transitividad		
2	anillo	cuerpo	grupo	teorema	demostracio	ciclico	axioma	
	anillo	grupo	campo	polinomio	real	complejo	pequenoteo	
	grupo	anillo	clase	congruencia	modulo	relacion	orden	
	grupo	anillo	cuerpo	relacion	simetria	reflexividad	propiedad	
	grupo	anillo	cuerpo	matriz	determinant	praxeologia	complejo	
	grupo	anillo	grupoabellai	conjunto	operacion	conmutativi	asociativida	
	grupo	anillo	lci	asociativida	conmutativi	ideal	subgrupo	
	grupo	anillo	teorema	subgrupo	ciclico	cuerpo	axioma	
	grupo	campo	anillo	ciclico	h	abeliano	conmutativi	
	grupo	campo	grupoide	semigrupo	abeliano	vector	anillo	
grupo	subgrupo	cuerpo	ivobosso	subcuerpo	anillo	subanillo		
3	plano	polar	cuadrado	cubo	triangulo	terceradime	propiedad	
	dificil	demostracio	numeros	grupo	anillo	cuerpo	campo	
4	conjunto	algebra	notacion					
	conjunto	matriz	notacion					
	conjunto	real	natural					
	demostracion							
	demostracio	analisis	calculos					
	ecuacion	demostracio	propiedad					
	estructura	algebra	matriz					
	estructura	matriz	funcion	dominio				
	estructura	numeros	funcion	relacion	demostracio	congruencia		
	funcion	notable	teorema					
funcion	trigonometricas							
5	estructura	algebra	miriamv	cnen	horror	cuerpo	ciclico	
	estructura	algebra	polinomio	binomio	trinomio	ecuacion	resultado	
	myriam	vicente	puntos	algebra	cartesiano	anillo	cuerpo	

## 5 CONCLUSIÓN

Se ha desarrollado un análisis de clustering mediante el algoritmo k-medoids y la distancia de Levenshtein. A partir de este enfoque, se ha realizado un análisis de la disponibilidad léxica de dos grupos de estudiantes de pedagogía en dos universidades chilenas. Los resultados muestran que es factible utilizar la representación

de lexicones como conjunto de símbolos ordenados, y a pesar que el análisis realizado tiene el componente subjetivo de necesitar un experto en el área para poder obtener conclusiones de los resultados, es posible obtener algunas medidas cuantitativas mediante la búsqueda no supervisada de grupos representativos de los centros de interés utilizando algoritmos de clustering e índices de cohesión, disminuyendo así la subjetividad de la interpretación.

El análisis de los datos para los centros de interés seleccionados muestra una diferencia notable en el léxico disponible de ambas universidades, así como en la cantidad de años del estudiante en la universidad. Además, existen pequeños grupos que pueden tener diferencias de género y notas.

Como proyección para trabajo futuro y potenciales aplicaciones de este enfoque, es posible caracterizar a un grupo de estudiantes respecto a su disponibilidad léxica en un tema específico, el proceso de clustering permitiría agrupar a los alumnos de un curso respecto al “vocabulario” que manejan respecto a un tema específico. La naturaleza caracterizadora del proceso de clustering podría ayudar a conformar grupos de trabajo (automáticamente) teniendo en cuenta el vocabulario de los grupos (que es similar entre sus integrantes), ayudando al profesor en la confección de ejemplos y explicaciones en forma separada para cada grupo si fuese necesaria. Por ejemplo, para los grupos conformados para el tema “Datos y Azar”, es claro que los alumnos de los grupos 1 y 5 asocian este concepto con los juegos de azar (ver Tabla 7), por lo que podría ser beneficioso una aclaración de las diferencias y similitudes respecto al tema “Datos y Azar” con respecto a “Juegos de Azar” para estos dos grupos.

Para finalizar cabe aclarar que estos resultados corresponden al análisis de los lexicones solamente, y que a pesar de que se podría haber llegado a la misma conclusión observando los datos estadísticos de los centros de interés respecto al género, universidad y demás características, esto hace más interesante el enfoque, debido a que la agrupación y caracterización de la disponibilidad léxica se realizó sin tener en consideración las características adicionales, sino que solamente utilizando como fuente de información las encuestas de disponibilidad léxica de los alumnos, lo que podría mostrar una alta relación existente entre el léxico de los estudiantes y algunas de las características descriptivas más comúnmente utilizadas en la academia.

## 7 AGRADECIMIENTOS

Proyecto de Investigación - Fondecyt 1140457 “Plataforma Adaptativa online para el fortalecimiento de las competencias matemáticas y pedagógicas a partir del estudio léxico semántico de estudiantes y profesores de pedagogía en matemática”, de la Comisión Nacional de Investigación Científica y Tecnológica de Chile.

## 8 REFERENCIAS

- [1] Anderson, L. W.; Krathwohl, D. R. & Bloom, B. S. A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives *Allyn & Bacon*, **2001**.
- [2] Colomb, J. & Yves, C. La Transposition didactique: du savoir savant au savoir enseigné *Revue française de*

*pédagogie, Institut national de recherche pédagogique*, **1986**, 76, 89-91.

- [3] \*Echeverría, M.; Vargas, R.; Urzua, P. & Ferreira, R. DispoGrafo: una nueva herramienta computacional para el análisis de relaciones semánticas en el léxico disponible *RLA. Revista de lingüística teórica y aplicada, scielo.cl*, **2008**, 46, 81 – 91.
- [4] Ferreira, A.; Salcedo, P. & Del Valle, M. Estudio de disponibilidad léxica en el ámbito de las matemáticas *Estudios filológicos*, **2014**, 69-84.
- [5] Michea, R. Mots fréquents et mots disponibles, un aspect nouveau de la statistique du langage *Langues modernes*, **1953**, 47, 338-344
- [6] MINEDUC. (2012). Estándares orientadores para carreras de pedagogía en educación media, <http://portales.mineduc.cl/usuarios/cpeip/File/libroestandaresvale/libromediafinal.pdf>
- [7] Pakhira, M.; Bandyopadhyay, S. & Maulik, U. Validity index for crisp and fuzzy clusters *Pattern recognition, Elsevier*, **2004**, 37, 487-501.
- [8] Rojas, D.; Rueda, L.; Ngom, A.; Hurrutia, H. & Carcamo, G. Image segmentation of biofilm structures using optimal multi-level thresholding. *International journal of data mining and bioinformatics, Inderscience Publishers*, **2011**, 5, 266-286.
- [9] Salcedo, P.; Ferreira, A. & Barrientos, F. A bayesian model for lexical availability of chilean high school students in mathematics. *Natural and Artificial Models in Computation and Biology: 5th International Work-Conference on the Interplay Between Natural and Artificial Computation, IWINAC 2013, Mallorca, Spain, June 10-14, 2013. Proceedings, Part I, Springer Berlin Heidelberg*, **2013**, 245-253.
- [10] \*Salcedo, P.; del Valle, M.; Contreras, R. & Pinninghoff, M. A. LEXMATH - A Tool for the study of available lexicon in mathematics. *Bioinspired Computation in Artificial Systems: International Work-Conference on the Interplay Between Natural and Artificial Computation, IWINAC 2015, Elche, Spain, June 1-5, 2015, Proceedings, Part II, Springer International Publishing*, **2015**, 11-19.
- [11] Urzua, P.; Saez, K. & Echeverría, M. Disponibilidad léxica matemática: análisis cuantitativo y cualitativo. *RLA. Revista de lingüística teórica y aplicada, scielo.cl*, **2006**, 44, 59 – 76.
- [12] Valencia, A. & Echeverría, M. Libro: Disponibilidad léxica en estudiantes chilenos. *Editorial Universidad de Chile*, **1999**.
- [13] Xu, R. & Wunsch, D. Clustering. *John Wiley & Sons*, **2008**, 10.
- [14] Yujian, L. & Bo, L. A Normalized Levenshtein Distance Metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **2007**, 29, 1091-1095.