Análisis de textos mediante un modelo de n-grama para evaluar la confiabilidad de una rúbrica. Una aplicación en la Evaluación del Desempeño de Especialistas de Educación del Perú

Giuliana Madeley Vidal Aguilera
Ministerio de Educación del Perú
gvidal@minedu.gob.pe
Luz Mery Pumacayo Manuelo
Ministerio de Educación del Perú
lpumacayo@minedu.gob.pe

Área temática: Evaluación educativa
Línea temática: Aportes metodológicos a la evaluación educativa
Tipo de ponencia: Reporte parciales o final de investigación

Resumen

El estudio tiene como objetivo generar evidencias de confiabilidad de las puntuaciones de rúbricas a partir de patrones textuales extraídos de los sustentos redactados por los comités de la Evaluación del desempeño de Especialistas en educación del Perú en el 2022. Se analizaron 1370 textos redactados por 169 comités, cada texto contenía los argumentos del nivel asignado a un evaluado. Se analizaron los textos de cuatro aspectos de la competencia Orientación a resultados: Planificación, Implementación y seguimiento, Uso de recursos y Compromiso con la Instancia de Gestión Educativa Descentralizada (IGED) donde labora. La metodología inició con la limpieza de los textos, la elección del n-grama adecuado y el análisis de los patrones de escritura por nivel de los 1-grama y se implementó el modelo 2-grama para estimar las secuencias de palabras para finalmente visualizarlas. Se removió el 46.8% de palabras del texto original, el n-grama adecuado fue para n=2 el cual fue identificado luego analizar la frecuencia promedio de los gramas. Se encontró que los niveles I y II estaban caracterizados por el incumplimiento total y parcial de las tareas respectivamente, en los niveles altos (nivel III y nivel IV) las secuencias de palabras fueron similares pero se diferenciaron por el valor agregado que el especialista le agregó al cumplimiento de la tarea. Finalmente, a partir de la gradación mostrada en los patrones extraídos por nivel, se encontraron evidencias de confiabilidad de las puntuaciones de las rúbricas en todos los criterios evaluados.

Palabras clave: Desempeño profesional, rúbricas, confiabilidad, modelos, análisis de textos

Introducción

En el campo de la evaluación educativa los datos de tipo texto también pueden proveer información acerca del evaluado e incluso del funcionamiento de un modelo de evaluación. Un procedimiento común para sintetizar textos es la codificación, es decir la agrupación de

respuestas similares para formar categorías; sin embargo, requiere tiempo para ser realizado. En las últimas décadas, la minería de textos se ha posicionado como un campo interdisciplinario de actividades entre la minería de datos, la lingüística, la estadística computacional y las ciencias de la computación (Feinerer et. al 2018) que provee métodos para procesar y extraer patrones de respuesta en textos.

En situaciones donde se requiere extraer información de fenómenos complejos como un texto de lenguaje natural, Feldman y Sanger (2007) señalan que es útil modelarlos como una forma de proceso aleatorio, por lo que los modelos probabilísticos son muy usados. Los modelos que asignan probabilidades a una secuencia de palabras son llamados Modelos de Lenguaje. El modelo de lenguaje más simple que predice la próxima palabra de una secuencia sucedida hasta el momento es el modelo de n-grama, donde “n” es el número de palabras en la secuencia histórica, si n=1 recibe el nombre 1-grama, si n=2 modelo de 2-grama, si n=2 modelo de 3-grama, etc. Si se considera memoria reciente es decir “n” es pequeño el modelo de n-grama puede ser definido como una cadena de Markov. Estos modelos son muy útiles para extraer características comunes de grandes conjuntos de textos, al estimar secuencias de palabras en base a probabilidades condicionadas a palabras previas, lo que nos permite hacer una lectura secuencial del patrón extraído y facilita su interpretación.

En el modelo de evaluación del desempeño de especialistas en educación en el Perú en el 2022, la calificación de la mayoría de competencias estuvo a cargo de un comité de evaluación y para guiar su proceso utilizaron rúbricas, cada rúbrica enlistó los criterios y una escala de cumplimiento: nivel I (Muy deficiente), nivel II (En proceso), nivel III (Suficiente) y nivel IV (Destacado). Además de calificar, el comité debía argumentar la puntuación asignada en forma escrita detallando las razones de la calificación esperándose que cada argumento esté alineado al nivel asignado, lo que brindaría una evidencia de confiabilidad de la calificación.

La confiabilidad de las puntuaciones de un instrumento es uno de los aspectos imprescindibles para que este pueda ser tomado en cuenta al momento de hacer una medición (Gómez-Benito & Hidalgo-Montesinos, 2003). Si bien la naturaleza de una rúbrica es cualitativa no está exenta de esta propiedad (Moskal y Leydens, 2000; Goodrich 2014). En una rúbrica la confiabilidad está relacionada a la consistencia de la calificación de los evaluadores y la verificación de la gradación de los niveles de puntuación otorgados. La gradación es uno de los aspectos más importantes en la construcción de una rúbrica y está caracterizada por descripciones progresivas de lo que se espera del evaluado en cada nivel de puntuación y deben ser distinguibles, por lo que es importante la aplicación de métodos como que nos permitan generar evidencias de que esta propiedad se está cumpliendo a partir de datos textuales.

Debido a lo anterior surge la siguiente problemática ¿Cómo extraer características comunes o patrones textuales de los sustentos redactados por los comités para generar evidencias de confiabilidad de las puntuaciones asignadas? Para verificar la confiabilidad de la calificación se analizó la gradación de los niveles en términos de los patrones textuales derivado, esperándose

que cada argumento esté alineado al nivel asignado. Los patrones textuales fueron extraídos utilizando el modelo de n-grama.

La aplicación fue realizada con los sustentos de la competencia Orientación a Resultados el cual evaluó cuatro aspectos: Planificación, implementación y seguimiento de tareas, el uso eficiente de recursos asignados y el compromiso con la Instancia de Gestión Educativa Descentralizada (IGED) donde labora (Minedu, 2022).

Desarrollo

Los datos fueron los 1370 textos redactados por 169 comités de evaluación, cada comité redactó un texto para cada uno de los cuatro aspectos medidos en la competencia de Orientación resultados: Planificación, Implementación y seguimiento, Uso de recursos y Compromiso con la IGED donde labora. Por instrucción los comités debían redactar los sustentos en función del cumplimiento de requisitos para encontrarse en el nivel. Se siguió la siguiente metodología: 1) Preprocesamiento, 2) Extracción de n-gramas y 3) Análisis de los patrones de escritura por nivel, donde a manera exploratoria se analizaron los 1-grama y se implementó el modelo 2-grama basado en la teoría y métricas que se describen a continuación:

- **Modelo n-grama** (Jurafsky y Martin, 2018):

La probabilidad conjunta de una secuencia de palabras está definida por:

$$P(X_1=w_1,X_2=w_2,X_3=w_3...X_n=w_n)=$$

Donde X es una variable aleatoria que puede ser una palabra cualquiera.

La probabilidad total de una secuencia entera de palabras o se puede expresar en términos de probabilidades condicionales utilizando la regla de la cadena:

$$P(w_{1:n})=P(w_1)P(w_2 |w_1)P(w_3 |w_{1:2})...P(w_n |w_{1:n-1})=\prod_{k=1}^{n}P(w_k |w_{1:k-1})$$

Donde $P(w_n |w_{1:n-1})$ es la probabilidad de la palabra dado una secuencia de n-1 palabras anteriores. Cuando las probabilidades condicionales se aproximan solo con algunas palabras previas $P(w_{1:n})$ es el modelo n-grama, es por esta característica que se define como una cadena de Markov. El caso más simple es cuando n=2, se denomina modelo 2-grama o bigrama y solo se considera la palabra anterior:

$$P(w_n |w_{1:n-1})\approx P(w_n |w_{n-1})$$

Donde $P(w_n |w_{n-1})$ es la probabilidad de la palabra dado la palabra anterior.

- **Métrica TF-IDF:** La métrica TF-IDF (Term Frequency Inverse Document Frequency) determina la frecuencia relativa de las palabras de un documento en específico comparado con la proporción inversa de una palabra sobre un conjunto entero de documentos (corpus). Este cálculo determina cuan relevante es una palabra en un documento en particular. Las

palabras que están en un solo documento o en un pequeño grupo de documentos tienden a tener valores altos de TF-IDF (Ramos, 2003).

- **Medida de similitud:** Para cuantificar en qué medida los textos redactados en los criterios tienen alguna similitud en cada nivel de logro, se calculó el coeficiente de correlación de Pearson de la frecuencia de los bigrama de los criterios según los niveles de logro.

Los cálculos se realizaron con el software R y los paquetes usados fueron: quanteda (Benoit et. al, 2018) para el preprocesamiento y visualización de la secuencia de palabras se utilizaron las librerías igraph (Csárdi, 2007) y ggraph (Pedersen 2017).

Conclusiones

- En el proceso de limpieza se removieron 31,177 palabras que representó el 46.8% de las palabras iniciales. Se conservaron las palabras de negación como “no”, de oposición como “aunque”, “sin embargo”, “no obstante” y los que dan cuenta de frecuencia como “casi siempre”, “pocas veces” entre otros, pues fueron claves para expresar el nivel de cumplimiento entre niveles y analizar su gradación. Los autores recomiendan no realizar una remoción automática de stopwords sino personalizarlo en base a las necesidades de información.
- Fue necesario tokenizar los textos para distintos valores de n de 1 hasta 10 y calcular el promedio de las frecuencias de ocurrencia de los gramas en cada n para plantear el modelo de n-grama adecuado. Debido a que a partir de n=2 el promedio de las frecuencias de ocurrencia de los gramas se hacían mínimas en todos los criterios en análisis, se decidió trabajar con el modelo 2-grama para la estimar las probabilidades condicionales y con el 1-grama hacer el análisis exploratorio.
- Hubo una mayor cantidad de textos redactados para los niveles III y IV porque estas fueron las calificaciones más predominantes en los evaluados. Del análisis de los 1-grama, se vio que estos niveles altos tienen la mayor cantidad de palabras mientras que los niveles bajos (I y II) tiene el mayor porcentaje de palabras distintas y mayor uniformidad en las palabras usadas. La longitud promedio de palabras estuvo entre 8 y 8.5. En los niveles I y II predominó la palabra “no” dando cuenta de un sentido negativo en el cumplimiento de los textos redactados en este nivel.
- El análisis exploratorio de los 2-gramas (antes de plantear el modelo) se realizó con la métrica TF-IDF, este tuvo valores altos en los niveles I y II pues eran los niveles con la menor cantidad de textos redactados, menor variabilidad y por ende mayor frecuencia de ocurrencia, este resultado se encontraría alineado a lo mencionado por Ramos (2003) pues menciona que las palabras en un pequeño grupo de documentos tienden a tener números TF-IDF más altos.
- Los patrones textuales extraídos para cada nivel utilizando el modelo de 2-grama mostraron la gradación esperada, dando evidencia de confiabilidad de los puntajes asignados. El nivel

I estuvo caracterizado por secuencias de palabras que daban cuenta del incumplimiento de las tareas con secuencias del tipo “no cumple”, “no evidencia”. El nivel II por un cumplimiento parcial y donde predominaban las secuencias “sin embargo”, “propuestas inviables”. El nivel III y IV presentaron patrones similares, pero la diferencia estuvo en el valor agregado. Por ejemplo, mientras en el nivel III se proponen alternativas en el nivel IV se ejecutan.

- La similitud de los patrones de los niveles III y IV fue evaluada con la correlación de las frecuencias entre los bigramas, el cual tuvo un valor alto alrededor de 0.8 en todos los criterios. Por otro lado, las correlaciones de los bigramas del nivel I con el resto si fueron pequeñas indicando de que se tratan de patrones distintos.
- Las instituciones que usan los resultados de una evaluación de desempeño para tomar decisiones de alto impacto se enfrentan al desafío de mostrar que la evidencia derivada de este tipo de evaluación es válida y confiable. Desde que es realizada con la guía de rúbricas, el diseño efectivo, la comprensión y el uso competente de las rúbricas es crucial (Jonson y Svingby, 2007). Con el análisis realizado se han utilizado métodos de la minería de textos y del modelo de lenguaje natural en el análisis de psicométrico de un instrumento, aplicación no vista en la literatura.

Tablas y figuras

Tabla 1. Cantidad de palabras registradas en sustentos antes y después de la limpieza del texto

Cantidad de palabras	Cantidad de stopwords	Cantidad de palabras sin stopwords
64 559	31 177	33 382

Figura 1. Nube de palabras registradas en los sustentos antes y después de la limpieza del texto

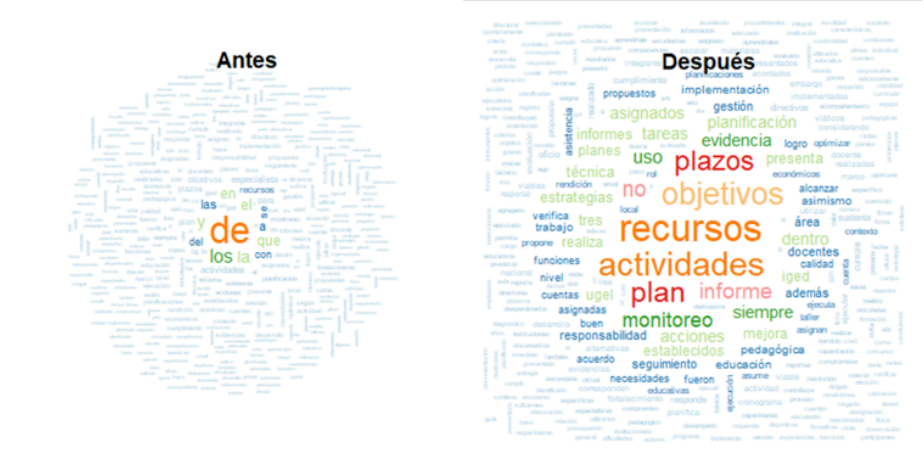


Figura 2. Frecuencia promedio de n-gramas, según criterio y nivel de logro

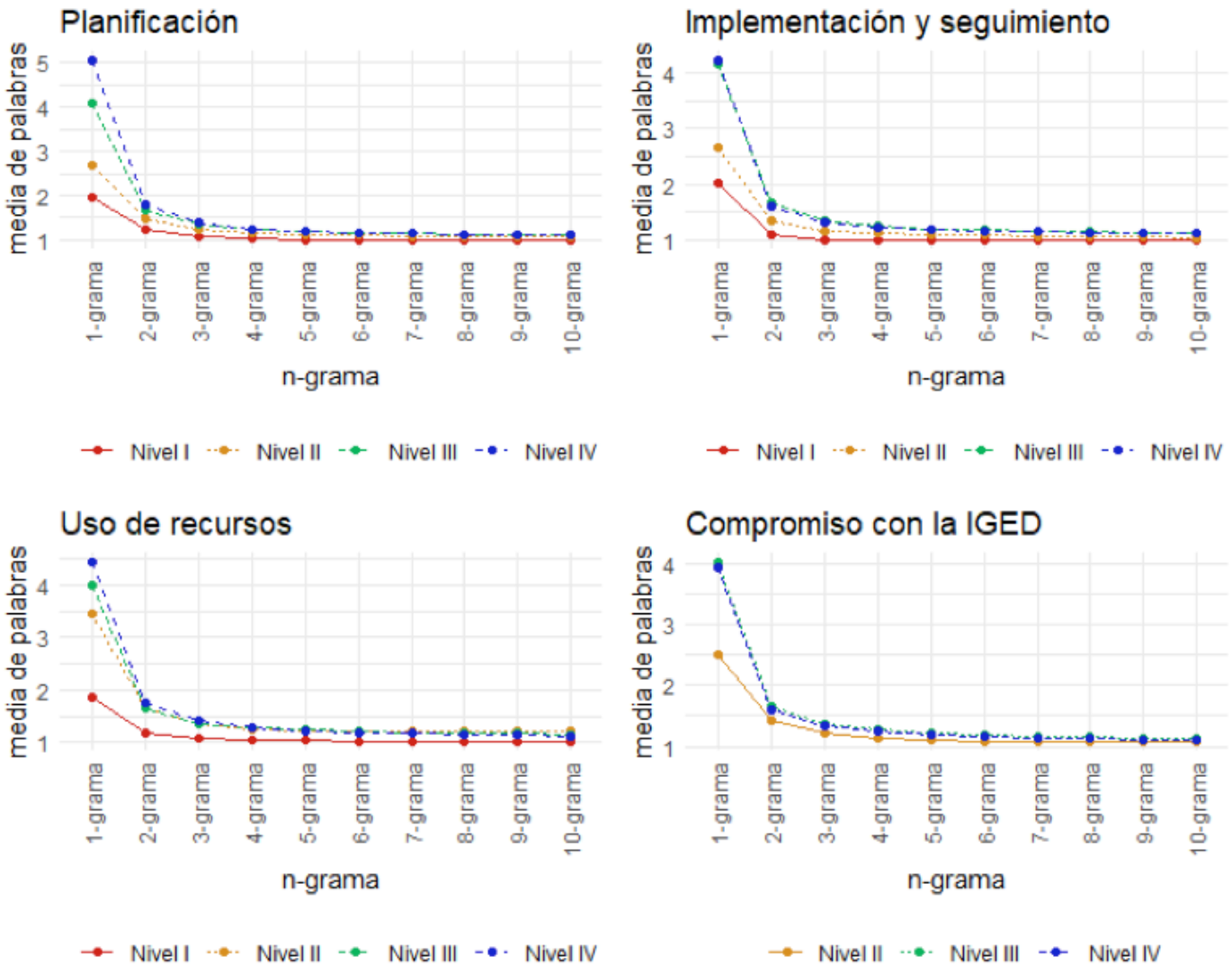


Tabla 2. Cantidad de textos, palabras y palabras distintas registradas

Criterios de la competencia Orientación a Resultados	Total de textos	Total de palabras	Palabras distintas	% Palabras distintas
Planificación	**301**	**9 192**	**1 451**	**15,8%**
Nivel I	7	183	93	50,8%
Nivel II	22	523	194	37,1%
Nivel III	98	2 676	657	24,6%
Nivel IV	174	5 810	1 150	19,8%
Implementación y seguimiento	**311**	**8 217**	**1 399**	**17,0%**
Nivel I	11	183	91	49,7%
Nivel II	33	797	298	37,4%
Nivel III	146	3 613	865	23,9%
Nivel IV	121	3 624	856	23,6%
Uso de recursos	**306**	**8 563**	**1 411**	**16,5%**
Nivel I	6	172	92	53,5%
Nivel II	54	1 255	364	29,0%
Nivel III	115	3 160	792	25,1%
Nivel IV	131	3 976	899	22,6%
Compromiso con la IGED	**307**	**7 410**	**1 346**	**18,2%**
Nivel I	-	-	-	-
Nivel II	32	622	247	39,7%
Nivel III	130	2 965	738	24,9%
Nivel IV	145	3 823	972	25,4%

Tabla 3. Cantidad promedio de palabras por texto en cada nivel

Criterios	Total de textos	Mínimo de palabras	Promedio de palabras	Máximo de palabras	Coeficiente de Variación
Planificación	**301**	**12**	**30,5**	**167**	**0,73**
Nivel I	7	14	26,1	50	0,49
Nivel II	22	15	23,8	43	0,37
Nivel III	98	12	27,3	122	0,55
Nivel IV	174	13	33,4	167	0,79
Implementación y seguimiento	**311**	**11**	**26,4**	**147**	**0,70**
Nivel I	11	11	16,6	19	0,15
Nivel II	33	13	24,2	64	0,51
Nivel III	146	12	24,7	105	0,62
Nivel IV	121	12	30,0	147	0,77
Uso de recursos	**306**	**9**	**28,0**	**155**	**0,65**
Nivel I	6	11	28,7	54	0,56
Nivel II	54	15	23,2	71	0,46
Nivel III	115	9	27,5	104	0,62
Nivel IV	131	13	30,4	155	0,71
Compromiso con la IGED	**307**	**3**	**24,1**	**132**	**0,59**
Nivel I	-	-	-	-	-
Nivel II	32	3	19,4	54	0,39
Nivel III	130	11	22,8	110	0,58
Nivel IV	145	8	26,4	132	0,60

Tabla 4. Longitud promedio de las palabras según criterio y nivel

Criterios	Total de textos	Mínimo de letras	Promedio de letras	Máximo de letras	Coeficiente de Variación
Planificación	**301**	**5,8**	**8,2**	**10,5**	**0,08**
Nivel I	7	6,2	7,3	8,9	0,13
Nivel II	22	7,0	7,7	9,1	0,08
Nivel III	98	5,8	8,3	10,5	0,08
Nivel IV	174	6,3	8,3	9,8	0,07
Implementación y seguimiento	**311**	**6,7**	**8,5**	**10,8**	**0,07**
Nivel I	11	6,7	7,8	8,4	0,06
Nivel II	33	7,2	8,1	10,8	0,09
Nivel III	146	7,2	8,7	10,8	0,07
Nivel IV	121	7,1	8,5	10,3	0,07
Uso de recursos	**306**	**6,2**	**8,1**	**14,2**	**0,09**
Nivel I	6	6,2	7,0	7,8	0,09
Nivel II	54	6,3	7,8	12,1	0,11
Nivel III	115	6,6	8,2	14,2	0,10
Nivel IV	131	6,8	8,1	9,9	0,07
Compromiso con la IGED	**307**	**6,3**	**8,0**	**14,8**	**0,10**
Nivel I	-	-	-	-	-
Nivel II	32	6,3	7,6	10,3	0,11
Nivel III	130	6,5	8,0	10,0	0,09
Nivel IV	145	6,4	8,0	14,8	0,11

Figura 3. Top 10 de palabras empleadas del criterio Planificación, según nivel de logro

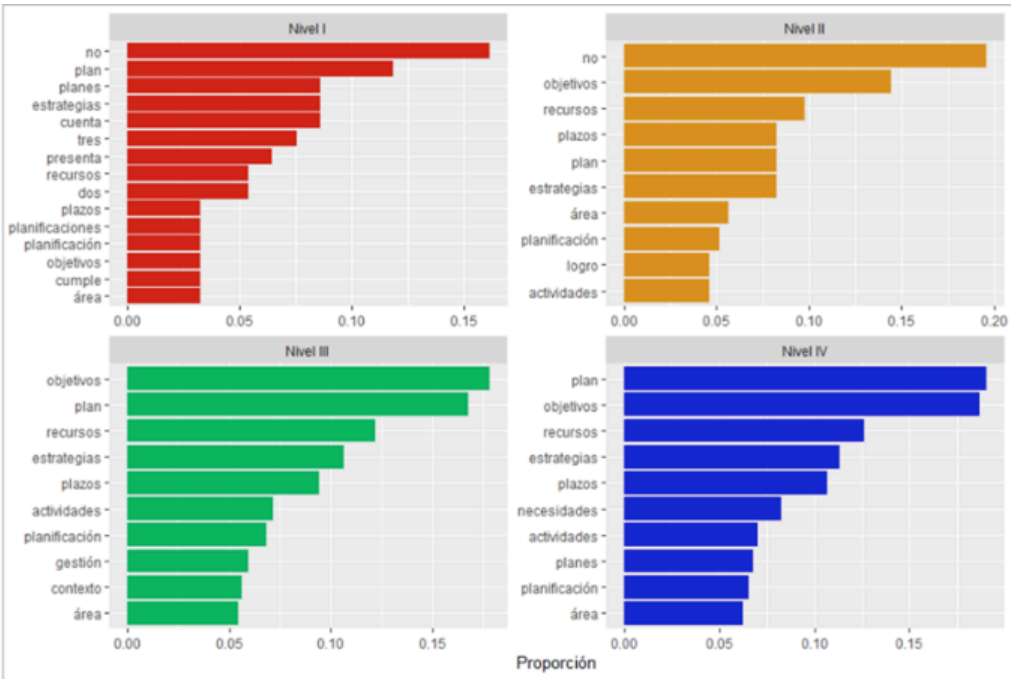


Figura 4. Top 10 de palabras empleadas del criterio Implementación y seguimiento, según nivel de logro

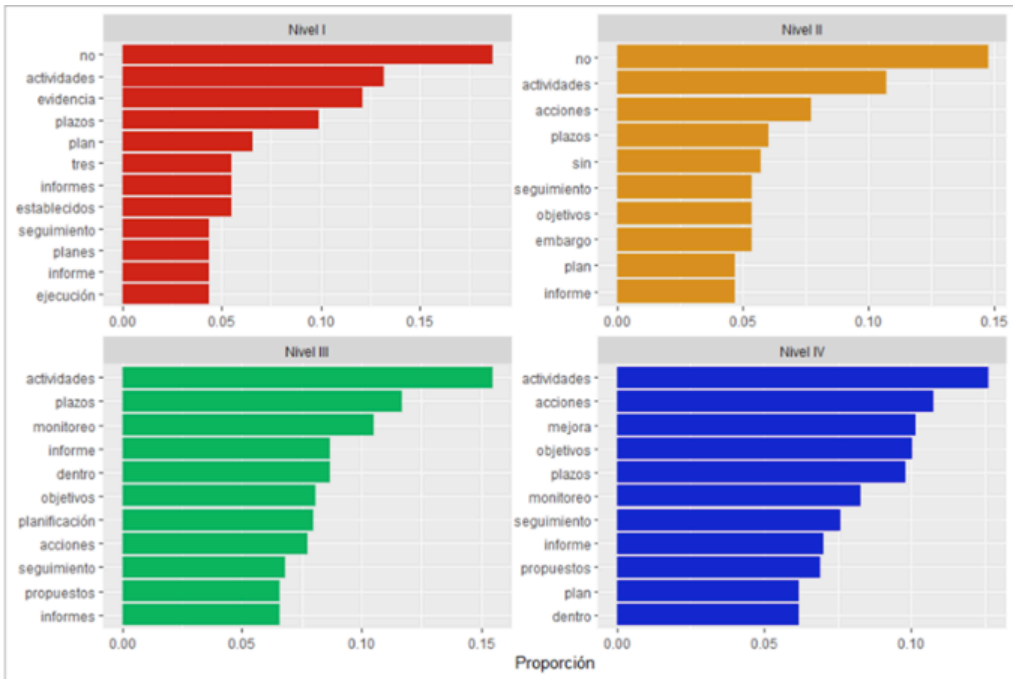


Figura 5. Top 10 de palabras empleadas del criterio Uso de recursos, según nivel de logro

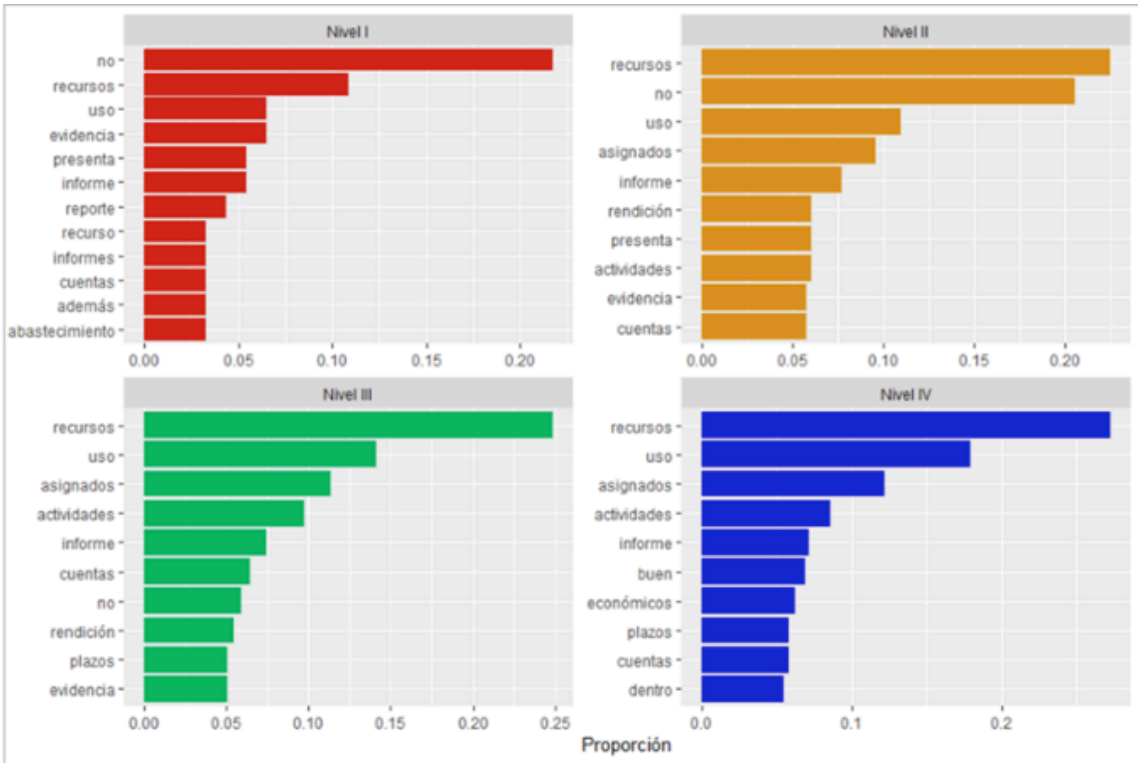


Figura 6. Top 10 de palabras empleadas del criterio Compromiso con la IGED, según nivel de logro

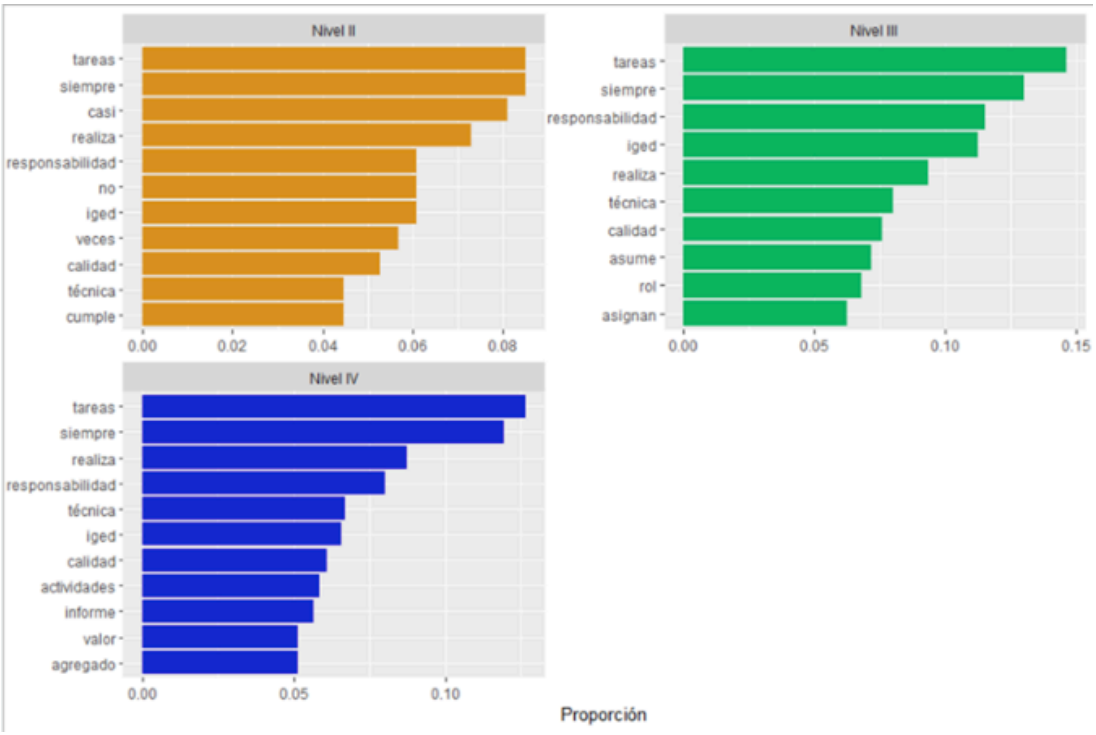


Figura 7. Top 3 de métrica TF-IDF del criterio Planificación, según nivel de logro

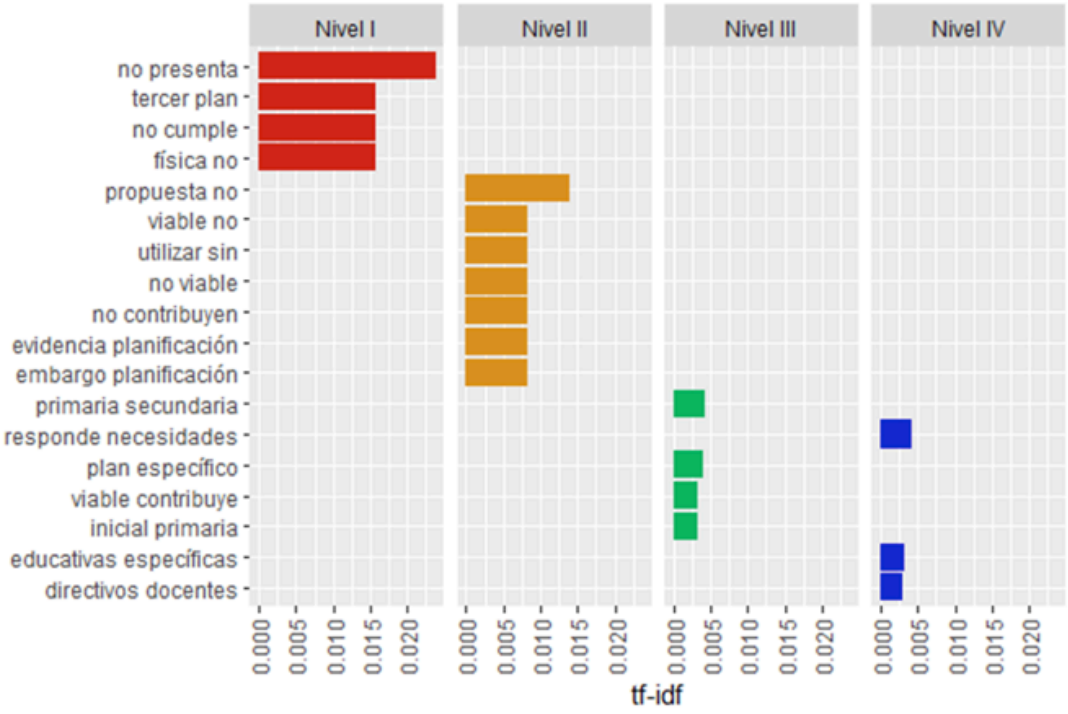


Figura 8. Top 3 de métrica TF-IDF del criterio Implementación y seguimiento, según nivel de logro

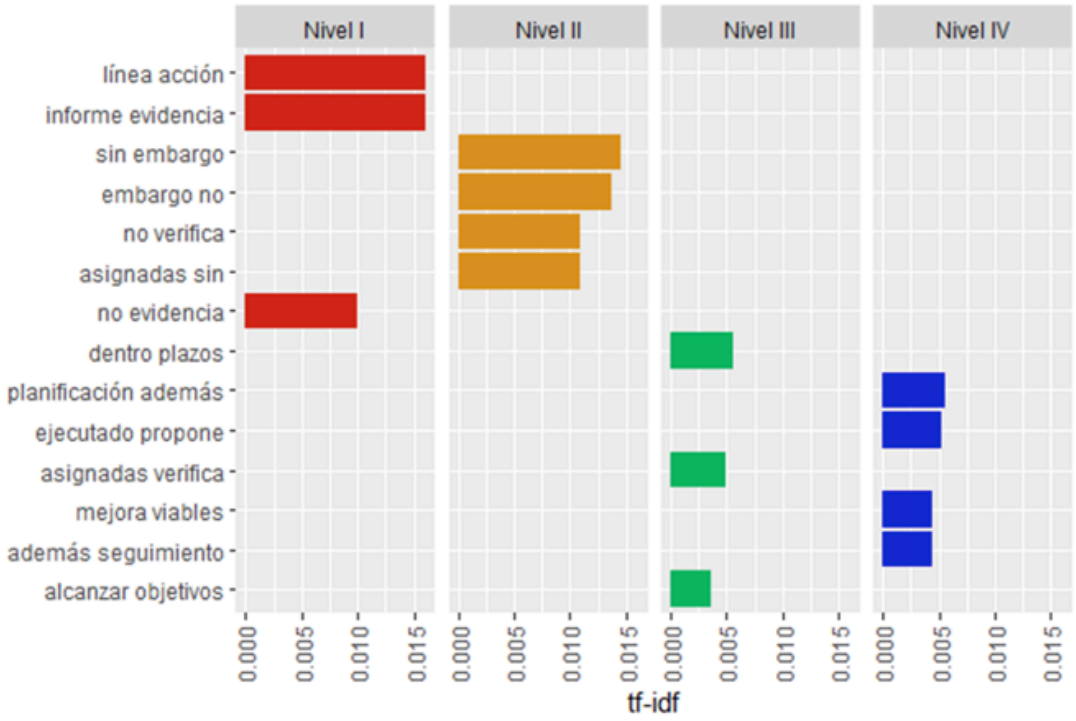


Figura 9. Top 3 de métrica TF-IDF del criterio Uso de recursos, según nivel de logro

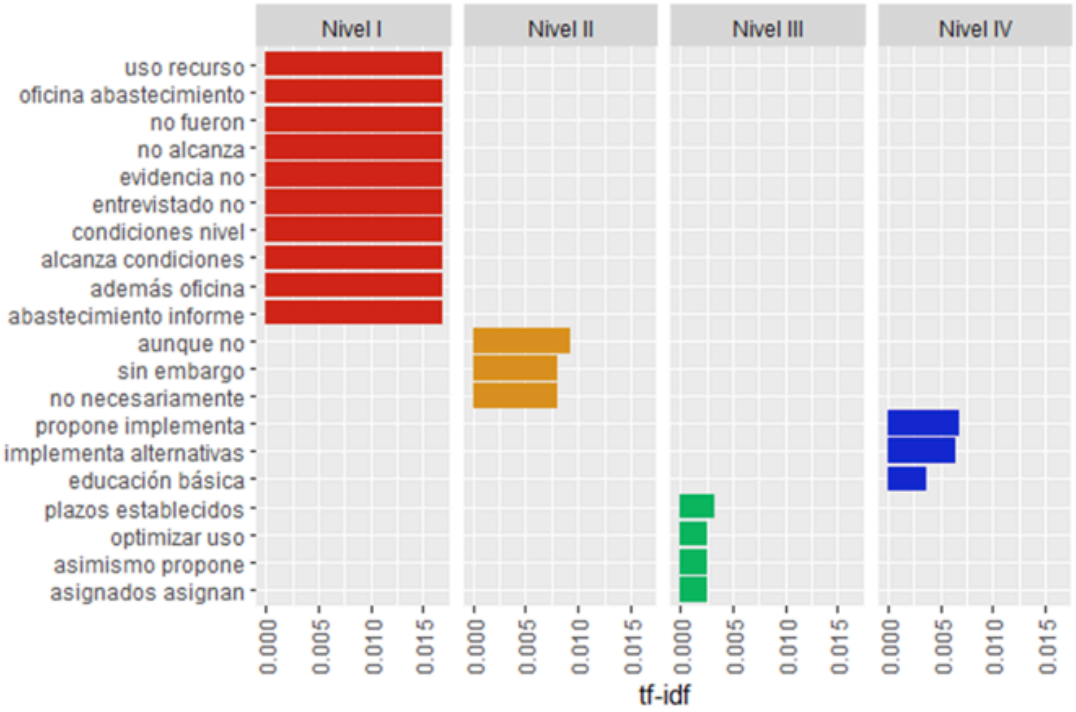


Figura 10. Top 3 de métrica TF-IDF del criterio Compromiso de la IGED, según nivel de logro

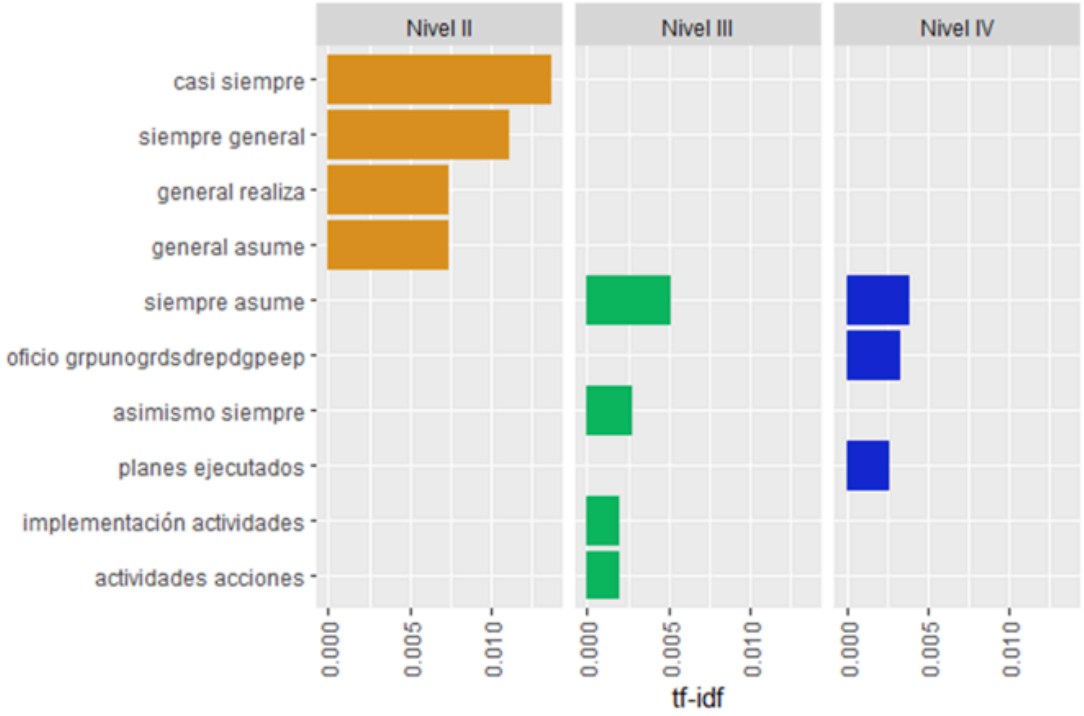


Figura 11. Grafo direccionado de los bigramas comunes del criterio Planificación según nivel

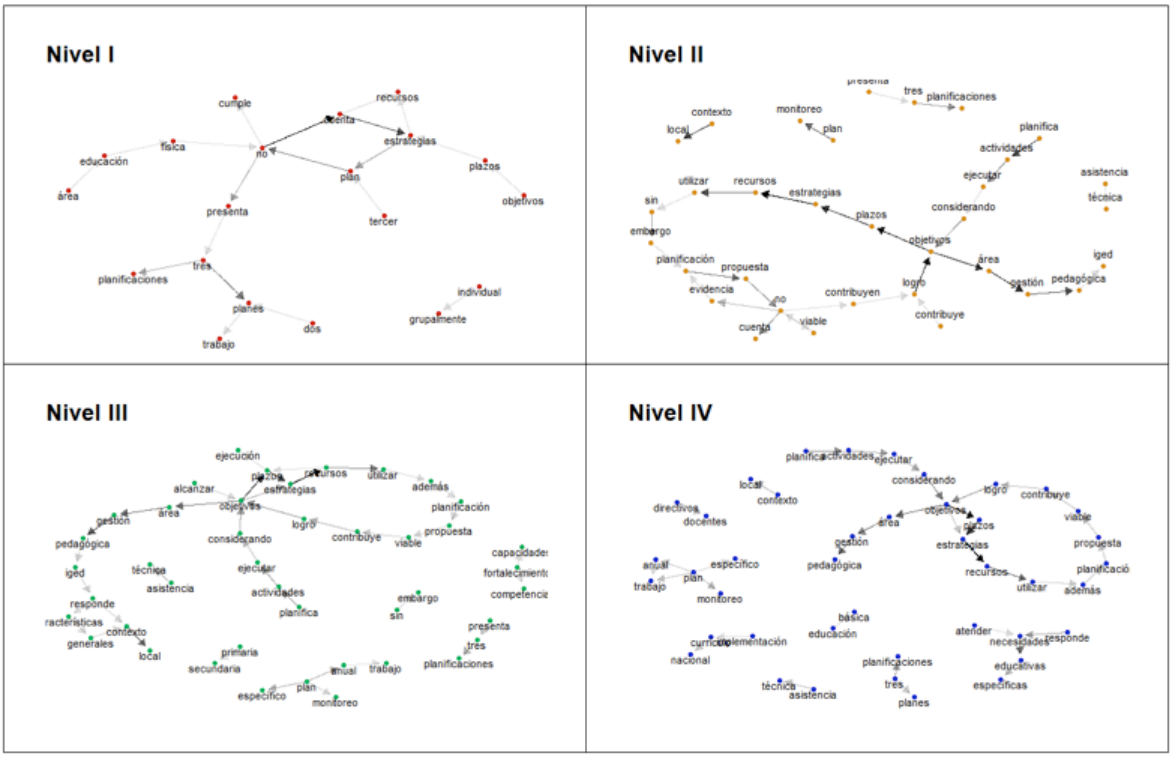


Figura 12. Grafo direccionado de los bigramas comunes del criterio Implementación y seguimiento de la IGED, según nivel de logro

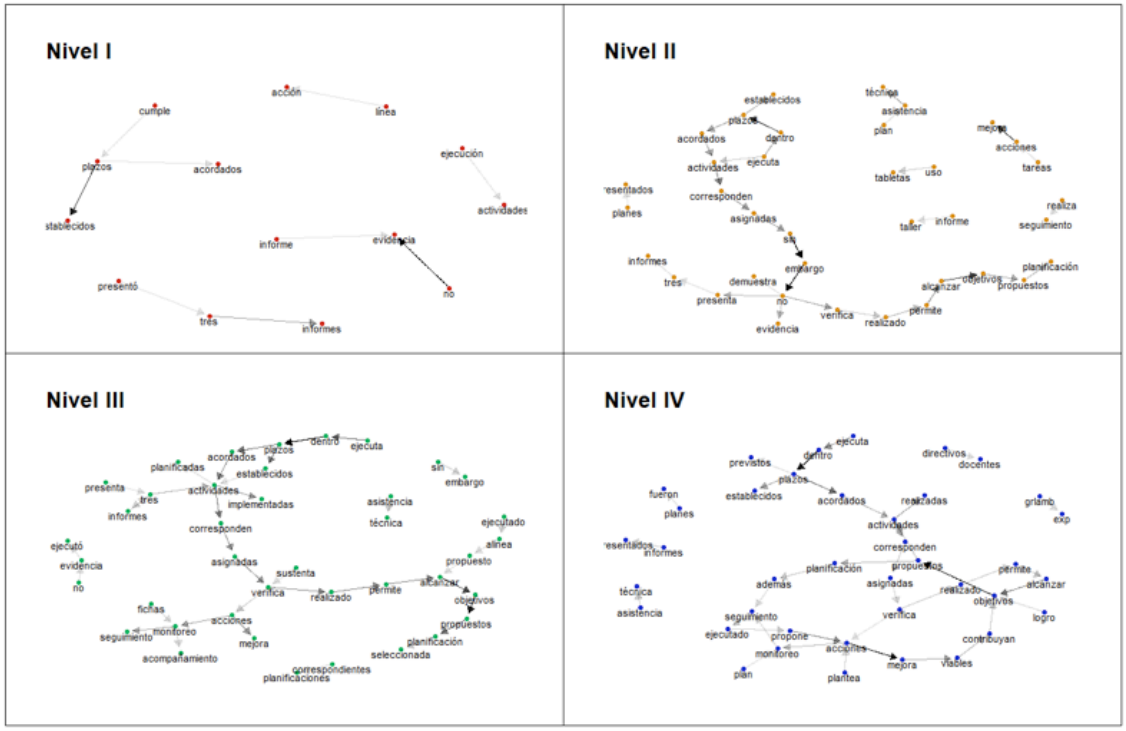


Figura 13. Grafo direccionado de los bigramas comunes del criterio Uso de recursos según nivel

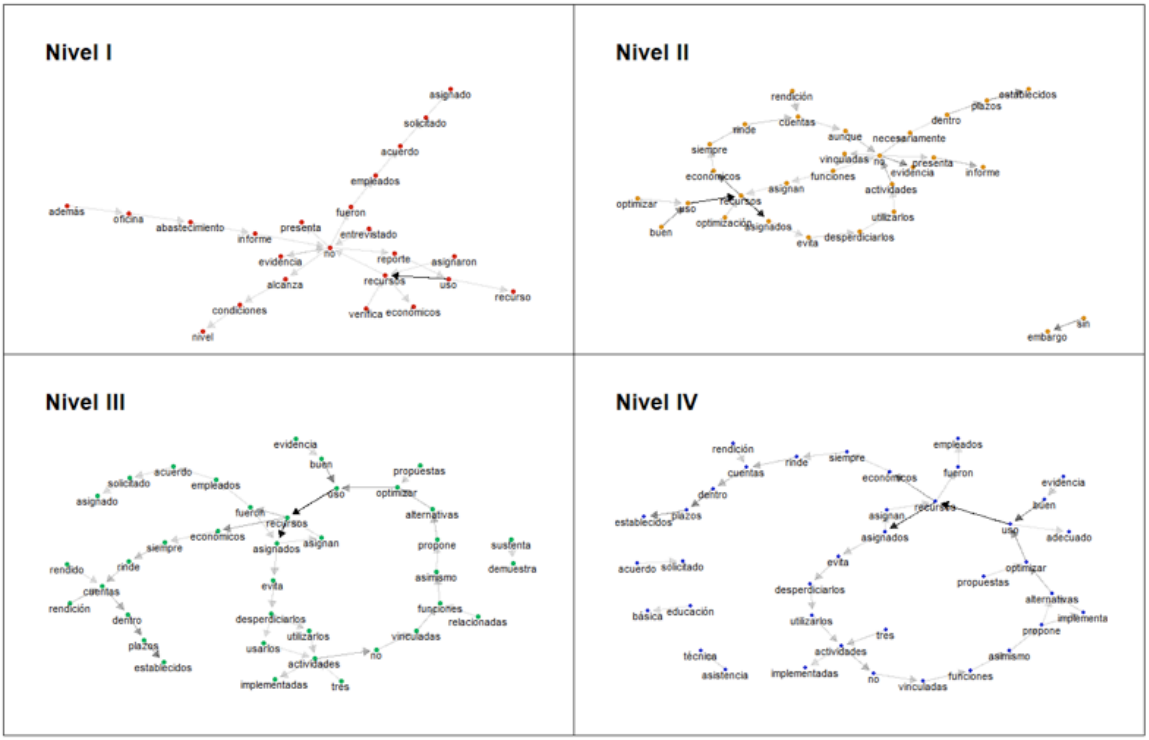


Figura 14. Grafo direccionado de los bigramas comunes del criterio Compromiso con la IGED, según nivel

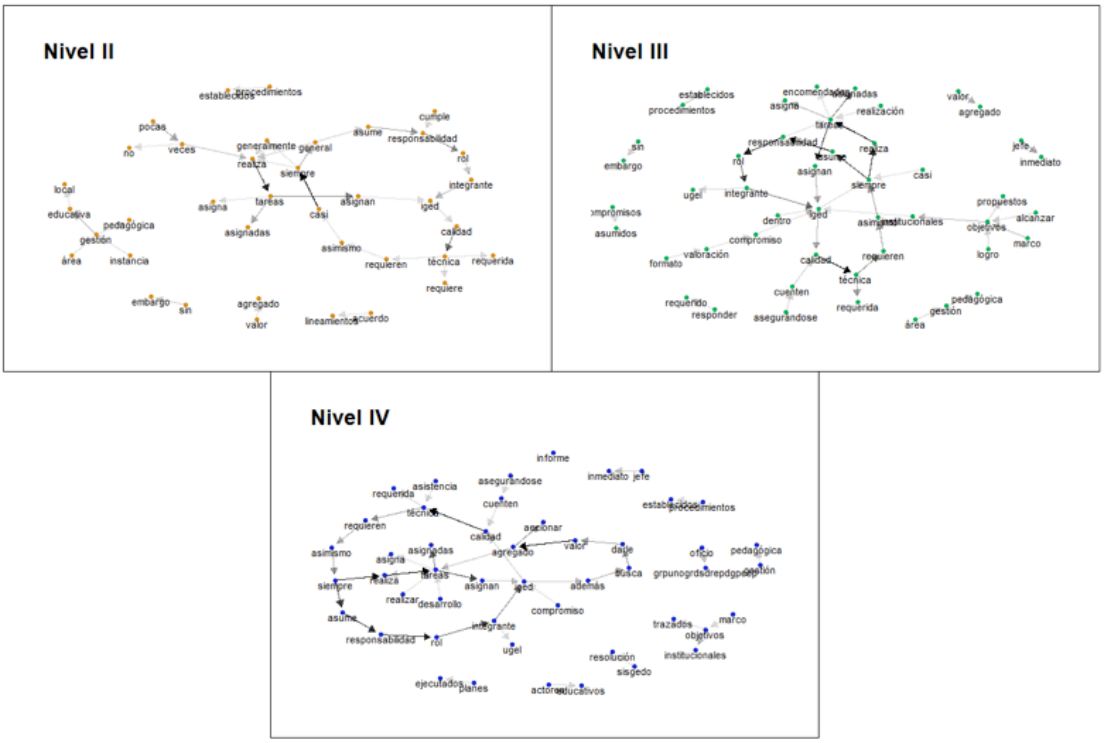
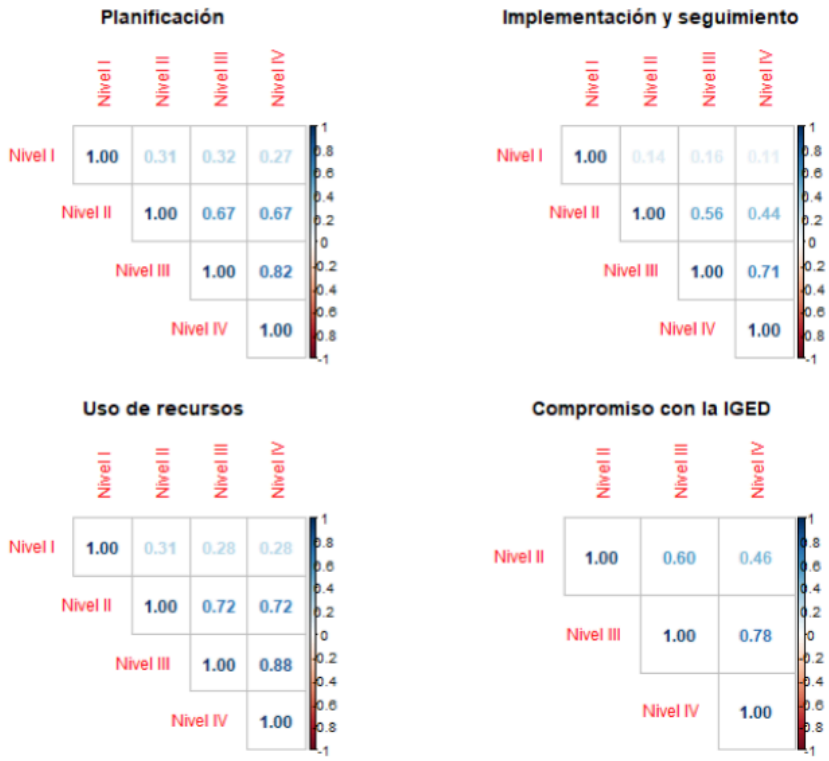


Figura 15. Correlación de bigramas, según criterio y nivel de logro



Referencias

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). *quanteda: An R package for the quantitative analysis of textual data*. Journal of Open Source Software, 3(30), 774-774. Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. Journal of statistical software, 25(5), 1-54.

Csardi, G., & Csardi, M. G. (2007). The igraph package

Feldman, R., & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press.

Jurafsky, D., & Martin, J. H. (2018), *Speech and Language Processing* (Third Edition), Prentice Hall.

Jonsson, A., & Svingby, G. (2007). *The use of scoring rubrics: Reliability, validity and educational consequences*. Educational research review, 2(2), 130-144.

Ministerio de Educación de Perú (2022). *Norma que regula la Evaluación del Desempeño en el cargo de Especialista en Educación de las Unidades de Gestión Educativa Local y Direcciones Regionales de Educación - 2022, en el marco de la Carrera Pública Magisterial de la Ley de Reforma Magisterial*

Moskal, B. M. (2000). *Scoring rubrics: What, when and how?. Practical Assessment, Research, and Evaluation*, 7(1), 3.

Moskal, B. M., & Leydens, J. A. (2000). *Scoring rubric development: Validity and reliability. Practical assessment, research, and evaluation*, 7(1), 10.

Ramos, J. (2003, December). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning* (Vol. 242, No. 1, pp. 29-48).

Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach*. O'Reilly Media, Inc.