Capítulo 8

Conclusiones

El objetivo principal que este proyecto perseguía era la construcción de una aplicación computacional de la teoría formulaica, que fuera útil para hacer *clustering* sobre textos no estructurados. Al final de este proyecto se puede asegurar que el objetivo se ha cumplido.

Para la realización de este proyecto se requirió revisar mucha literatura, porque aunque casi no hay nada explícitamente relacionado, eventualmente, se pudo ver que sí hay mucha documentación que respalda la utilidad y la popularidad de los algoritmos empleados para la implementación. Se puede decir que la implementación es útil porque toca muchas áreas de estudio, para las que la funcionalidad es de interés.

Al revisar la literatura, se encontraron muchos trabajos relacionados, donde la identificación de expresiones formulaicas puede ser útil. Para empezar, en nuestro proyecto implementamos una técnica para (*clustering*) que es muy común en minería, y por lo tanto hay un nexo directo con esta área. Después, cuando se analizaron algoritmos en el capítulo de marco teórico, se revisó el de indizado que es muy común en recuperación de información. Éste consiste en asignar pesos a las palabras que forman un texto: aquellas que son claves tendrán mayor peso. De cierta manera esto fue lo que hicimos, por lo menos tienen la misma intención, aunque ya no sólo se trató de palabras, sino también

con frases y fórmulas tanto estructurales como semánticas y situacionales. Entonces, a pesar de que en nuestra implementación introducimos otros conceptos, se puede ver que también hay una relación directa.

En el caso de la lingüística y la lingüística computacional, en este trabajo se aplicaron algoritmos que derivan de estás áreas, como son el algoritmo de Porter y la eliminación de palabras vacías. Además, dentro de la formalización se definieron términos como peso, representatividad y carga semántica, que no son más que parámetros que aplican algunas teorías lingüísticas.

La utilidad de encontrar expresiones formulaicas en textos quedó demostrada con lo anterior. Aunque no hay casi ningún trabajo que aplique directamente la teoría formulaica, se puede palpar su validez, porque estaba siendo parcialmente utilizada en muchos trabajos incluso antes de que fuera presentada formalmente.

Por otro lado, al revisar trabajos relacionados, se encontró que ninguno de ellos tiene algún tipo de formalización. Pienso que hacer esto es muy importante, porque aunque la intuición es muy útil, al tener una base formal se elimina la subjetividad y los resultados se vuelven menos cuestionables, con mayor credibilidad y, por consiguiente, tienen mayor peso científicamente hablando. Además, al formalizar se definen parámetros que nos pueden servir como punto de referencia para medir el éxito de nuestros experimentos.

Sí se tiene conciencia que formalizar matemáticamente, describiendo propiedades de los datos, es difícil porque se trata de una teoría lingüística, que finalmente es un área de humanidades, pero no es imposible hacerlo, aunque no es trivial.

En general, se cumplieron los objetivos planteados; con respecto al corpus, se compil un corpus de textos para entrenamiento y otro para pruebas. Sin embargo, en los experimentos siempre hubo mucha incertidumbre, porque la recolección de la muestra no fue totalmente satisfactoria. Aunque al analizar los resultados se pudo ver que se logró construir la base inicial de fórmulas para cada dominio, y con ésta se pudieron

clasificar los textos de prueba aplicando el algoritmo diseñado.

Por otro lado, también de los errores se aprende, tal es el caso se SUBDUE que finalmente no fue utilizado, pero fue buena idea conocerlo porque finalmente influyó en el diseño del algoritmo de reconocimiento de patrones que se utilizó. Es decir, aunque el algoritmo en sí no fue aplicado directamente, nuestra implementación fue diseñada siguiendo la lógica de este método.

Se logró la implementación de la teoría formulaica a través de la elección de algoritmos adecuados que cumplieron con los requisitos de sencillez, independencia del dominio e idioma, establecidos desde el principio.

Al final, la mayor aportación es el intento de la formalización de la metodología, porque esta formalización facilitará la continuación del trabajo que quedó pendiente, sobre todo en lo que se refiere a metodología que permitirá que el tamaño de las fórmulas sea variable.

Se espera que este trabajo puede ser continuado para ser aplicado en las siguientes áreas potenciales: como herramienta para determinar el grado de formulaicidad de un dominio, con lo anterior se puede utilizar en el área de recuperación de información y así hacer *clustering* de textos; además se pueden construir diccionarios de expresiones comunes para cada dominio. También se pueden encontrar relaciones entre dominios, es decir, encontrar subdominios dentro de un dominio.