Resumen

El propósito de esta tesis es explorar el potencial de la teoría formulaica para agrupar textos no estructurados, basados en su contenido temático y estilo. La formulaicidad lingüística es una teoría lingüística reciente que propone que existen fórmulas o expresiones preformadas como parte de lo que se dice y es escrito. En este trabajo, textos de dominios que parecen contener formulas o secuencias formulaicas son recolectados, buscando las expresiones que son más típicas para cada dominio. El algoritmo de reconocimiento de patrones usado es un híbrido de SUBDUE. Esta es una herramienta nueva de minería de textos para encontrar subestructuras en grafos, este algoritmo es adaptado para resolver un problema de la lingüística computacional, proveyendo una alternativa al algoritmo de colocación clásico. Después, las expresiones encontradas son usadas por el algoritmo de agrupamiento formulaico que desarrollamos para identificar nuevos textos como pertenecientes o no a alguno de los dominios con los cuales el algoritmo está familiarizado. Los resultados son discutidos para probar el potencial de la teoría.