**Implementación de un analizador morfológico básico mediante expresiones regulares en Python. - Francisco José Morón Reyes**

### Aproximación

Nuestro analizador consta de dos partes fundamentales, un diccionario de términos comunes dónde hemos agrupado clases gramaticales fundamentalmente cerradas (preposiciones, determinantes, adjetivos…) y una lista con un conjunto de expresiones regulares. Ambos grupos determinan, en función de un patrón, una etiqueta para cada palabra que previamente hemos tokenizado.

Cabe destacar que el programa desarrollado se ha elaborado teniendo en mente una temática genérica. Todos los términos que hemos añadido, tanto en el diccionario como en la lista de expresiones regulares son de carácter general, pero teniendo también en cuenta que estamos en un dominio específico de colchones. Esto quiere decir que se ha buscado que, independientemente del texto que nos encontremos, el etiquetador arroje buenos resultados.

Debido a la gran generalidad que se ha cubierto con el diccionario de términos comunes, se ha decidido almacenarlo en un archivo externo (clases-cerradas.csv), desde el que importamos los términos y su correspondiente etiqueta. Se han añadido más de 700 términos para este archivo de configuración, incluyendo algunos verbos que por su irregularidad e importancia se han tenido en cuenta aquí.

Para las expresiones regulares, se ha decidido dejarlas alojadas a nivel de código, puesto que no son tan numerosas. Se ha intentado cubrir el mayor espectro posible de tiempos verbales y sustantivos, ordenándolos en función de la prioridad que menos problemas de ambigüedad nos diese. También hemos añadido en este apartado algunos grupos que podían cubrirse fácilmente con reglas, como por ejemplo la puntuación o los números.

### Análisis de los resultados

Como hemos hecho una fuerte apuesta por una amplia variedad de términos, los resultados que nos llegan tras analizar el corpus de test son bastante buenos. De 132 tokens que encontramos en el corpus de pruebas, el error oscila entre sobre los 15 tokens en función del umbral en el que queramos situar la ambigüedad.

Los grupos cerrados, como preposiciones o artículos funcionan muy bien. También los adverbios y adjetivos. Los pronombres y determinantes suelen presentar algunos errores de ambigüedad, pero han funcionado mejor de lo esperado. Por supuesto, ha ayudado mucho con los adjetivos hacer una revisión posterior dónde se tuviesen en cuenta adjetivos posiblemente relacionados con colchones, lo que ha mejorado aspectos no cubiertos en una primera versión.

En general nos dan problemas las terminaciones confusas que podrían ser verbos y sustantivos, como los finales en -as, que podrían ser plural o futuro. También algunos participios terminados en -o no contemplados, que se suelen confundir con sustantivos singulares y algunos sustantivos cuyas terminaciones no se han cubierto (bien por ser irregulares o por priorizarse antes los verbos) han dado fallos.

Como conclusión, diríamos que la efectividad en los textos es buena pero mejorable. A largo plazo es posible pulir el programa añadiendo sobre todo sustantivos y verbos irregulares y completando las listas de clases gramaticales cerradas.

La generalidad es el punto fuerte del programa. Como se ha realizado sin tener en cuenta dominio específico algunos, debería funcionar considerablemente bien con otros textos, independientemente del dominio con el que estén relacionados.

Por último, nuestra tasa de aciertos es de aproximadamente el 90% en el corpus de test. Es una buena cifra pero mejorable.

En general, con más tiempo y estudio sería posible pulir el programa y mejorar progresivamente los resultados.