## **Lematización y *stemming***.

En los lenguajes donde es frecuente el uso de formas flexionadas (es decir, en plural, en femenino, conjugada, etc) la *lematización* consiste en hallar el lema correspondiente a una palabra, es decir, la forma que por convenio se acepta como representante de todas las formas flexionadas de una misma palabra. Hallar el ***lema*** correspondiente es equivalente a una ***entrada de diccionario***.

De esta manera múltiples representaciones y tiempos de una palabra ('*ataco*', '*atacante*', '*atacado*'…) pueden reducirse a un solo token - ***atac[[1]](#footnote-1)***- y, por tanto, a un solo atributo.

Una buena lematización requiere un conocimiento del contexto en que se usa la palabra, para lo cual es necesario realizar el *etiquetado gramatical* (*POS Tagging*) del texto. No obstante, es más frecuente, sencillo y rápido la utilización de una forma tosca de lematización llamada *stemming* que prescinde del etiquetado gramatical y que, entre otras aplicaciones, se ha utilizado en Google Search (1) desde 2003.

De los algoritmos de *stemming* de uso libre, el más conocido es el de Porter (2) que se encuentra incluido dentro de *Snowball* (3)que es un software para procesamiento de cadenas de caracteres orientado a la creación de algoritmos de *stemming*.

Se han llevado a cabo diferentes ensayos de evaluación de la eficacia de este tipo de algoritmos en distintos idiomas. Hollink et al. (4) manifiestan que las técnicas básicas pero ciertos métodos auxiliares (palabras vacías, *stemmers*,..) deben adaptarse al idioma que se esté utilizando. Básicamente comparan la eficacia de la recuperación de información sobre textos sin *stemmizar* o *stemmizados* para siete idiomas europeos (alemán, español, finlandés, francés, holandés, inglés y sueco).

Las conclusiones fueron las siguientes:



Fig. 12: Resultados de la prueba de V. Hollink et al.

## Bibliografía.

1. **MIT Libraries. [En línea] http://libguides.mit.edu/c.php?g=176061&p=1159432.**

**2. *An algorithm for suffix stripping.* Porter, M. F. 1980, Program , Vol. 14 Issue: 3, págs. 130-137.**

**3. Snowball. [En línea] http://snowball.tartarus.org/.**

**4. *Monolingual Document Retrieval for European Languages.* Hollink, V., Kamps, J., Monz, C. et al. 4, 2004, Information Retrieval, Vol. 7.**

1. Con *SnowballStemmer.* [↑](#footnote-ref-1)