UNIVERSIDAD NACIONAL DE LA PLATA

FACULTAD DE INFORMÁTICA

TESINA DE LICENCIATURA

Título: Incidencia de Idiomas Populares en la Lengua Española con Big Data: Análisis Masivo de Datos Mediante Amazon Elastic MapReduce y Google N-grams
Autores: Julián De Luca
Director: Waldo Hasperué
Codirector: Franco Chichizola
Asesor profesional: Ismael Pablo Rodriguez
Carrera: Licenciatura en Sistemas

Resumen

La presente tesina analiza, diseña e implementa una solución para el problema de la detección de neologismos y extranjerismos en la lengua Española. Para este fin, realizamos un análisis de investigaciones previas y problemáticas surgidas, y decidimos realizar un aporte mediante un sistema de cloud computing.

Nuestra solución se centra en el uso de tecnologías de análisis masivo de datos para el tratamiento de corpus grandes de una manera eficiente. Realizamos un repaso general de las tecnologías existentes, sin especificar herramientas puntuales, para luego introducir herramientas concisas. Logramos alcanzar una solución sencilla pero eficaz y escalable mediante el uso de herramientas provistas por Amazon Web Services, utilizando el corpus público llamado Google Books Ngrams.

Finalmente, abrimos la posibilidad de utilizar otras herramientas, y la misma metodología en otro tipo de estudios. Demostramos que logramos un aporte a la comunidad de lingüística computacional, y seguiremos trabajando en el tema en cuestión.

Palabras Claves

Google Books Ngrams, Amazon Web Services, AWS, Cloud Computing, Big Data, MapReduce, EMR, Hive, Neología, Extranjerismos.

Conclusiones

Se desarrolló un sistema de detección de neologismos novedoso que permite apreciar el uso de cloud computing como herramienta potente para su procesamiento.

Se otorga a futuros investigadores una herramienta de punto de partida para realizar sus estudios, con resultados claros y software extensible y modificable.

Trabajos Realizados

Se desarrolló un software para el lanzamiento de clústeres EMR en AWS, que mediante Hive permite tratar información de manera distribuida sencillamente.

Explicamos la instalación, funcionamiento, importancia y relación del sistema, y se propuso una metodología novedosa para la detección de neologismos, impulsada por teorías de otros autores.

Trabajos Futuros

Se propone:

- *Extensión para el procesamiento de información en distintos clústeres, que permita paralelizar tareas independientes.*
- *Extensión para el filtrado de extranjerismos sobre los neologismos detectados.*
- *Implementar la misma metodología en un corpus distinto, con unidades de tiempo menores a 1 año, posiblemente con información de internet como pueden ser foros.*

Fecha de la presentación: Marzo 2016