BigNLP: Aproximando a Computación de Altas Prestacións ás Tecnoloxías Big Data: Aplicación ao Procesamento da Linguaxe Natural

Descrición

O procesamiento lingüístico de grandes cantidades de texto é unha tarefa complexa que require do uso de varias subtarefas organizadas en módulos interrelacionados. Un dos maiores problemas das técnicas de procesamiento lingüístico é o seu alto custo computacional e os seus problemas de escalabilidade, o que as fan inviables para a análise de grandes volumes (Gigabytes e mesmo Terabytes) de documentos. Doutra banda, cabe apuntar que a filosofía dos enfoques máis recentes da lingüística de corpus baséanse na "Web As Corpus", liña de investigación onde se postula que con máis datos e máis texto obtéñense mellores resultados.

Por esta razón, consideramos que a computación de altas prestacións e o uso de estratexias orientadas a Big Data encaixan de forma natural como solución á limitada eficiencia computacional dos módulos para o procesamento lingüistico. No entanto, a relativa simplicidade modular dos procesos, así como a clara independencia das unidades lingüísticas de entrada (frases, parágrafos, textos...), son factores a ter en conta que poden facilitar a integración dos módulos de PLN no contexto dos sistemas computacionais de altas prestacións mediante o uso de tecnoloxías Big Data.

Obxectivos

O obxectivo principal do proxecto será o de desenvolver un conxunto de novas ferramentas e solucións para procesamento Big Data, o que vai permitir integrar nunha suite paralela e escalable un conxunto de módulos multilingües para o procesamiento da linguaxe natural. Esta suite debe procesar grandes cantidades de texto en tempos de execución reducidos e, ao mesmo tempo, facer un uso eficiente das plataformas hardware de altas prestacións que se consideren, prestando especial atención ás arquitecturas heteroxéneas. En concreto, vanse a considerar módulos para a Extracción de Termos Multipalabra, Análises Sintáctico, Extracción de tripletas, Análise de Correferencia e Análise de sentimentos. Debemos destacar que os novos módulos PLN que se van a desenvolver neste proxecto poderán utilizarse en aplicacións lingüísticas máis complexas e de alto nivel como a tradución automática, a recuperación de información, sistemas de vixilancia tecnolóxica, etc. Así mesmo, as ferramentas xeradas como froito das investigacións do proxecto serán de propósito xeral e, por tanto, poderían aplicarse a códigos ou aplicacións provenientes de áreas diferentes á do procesamiento da linguaxe natural.

INVESTIGADORES

Investigador principal
Juan Carlos Pichel Campos

Investigadores do CiTIUS
Tomás Fernández Pena

DETALLES

Data de execución:
07/09/2015 - 06/09/2018

Financiado por
Proyectos de I+D+i para Jóvenes Investigadores del Programa Estatal de I+D+i Orientada a los Retos de la Sociedad, Ministerio de Economía y Competitividad, TIN2014-54565-JIN

PO FEDER Galicia 2014-2020 "*Unha maneira de facer Europa*"

PUBLICACIÓNS

LinguaKit: a Big Data-based multilingual tool for linguistic analysis and information extraction
International Conference on Social Networks Analysis, Management and Security, 2018

Towards a Big Data Multi-language Framework using Docker Containers
Jornadas Sarteco, 2018

A New Approach for Sparse Matrix Classification Based on Deep Learning Techniques
IEEE Cluster, 2018

Ver todas

PROGRAMAS CIENTÍFICOS

Computación avanzada