Create an NLP Dictionary for Spanish

The proposal period for 2022 internships is now closed
The proposal period for 2023 internships will open in November 2022

This is new project, more information coming soon. If you are interested in this project contact Lorraine Chapman.

Find out about the HPCC Systems Summer Internship Program.

Project Description

In order to eventually create digital human readers in Spanish, a dictionary must be established. This project will use the Spanish dictionary from Wiktionary. One interesting aspect of this project are the verbs in Spanish which have a rich morphology.

If you are interested in this project, please contact Add email link to mentor.

Completion of this project involves:

- Download the Spanish dictionary from wiktionary
- Write an NLP++ parser to extract the vocabulary from the wiktionary files into text files
- Write an NLP++ parser to transform the text files into knowledge base files
- Create Spanish test files for part-of-speech tagging
- Write an NLP++ part-of-speech tagger
- Run the tests using the NLP++ Plugin in ECL to show enhancements
- Create an NLP++ repository for the Spanish dictionary and analyzers

By the mid term review we would expect you to have:

- <What must be completed to pass the evaluation and continue on to complete the project>

Mentor	David de Hilster david.dehilster@lexisnexisrisk.com **Backup Mentor:** Add Backup Mentor Name Add link to Email Address
Skills needed	• Keen interest in natural language • Ability to learn and program in NLP++ • Ability to create test cases • Ability to write test code in ECL using the NLP++ plugin to test the enhanced dictionary
Deliverables	**Midterm** • Parts-of-speech text files **End of project** • A Spanish dictionary repository in the VisualText open source github including the dictionary files and NLP++ analyzers
Other resources	• HPCC Systems website • JIRA issue for this project • Wiktionary • Blog: Understanding Natural Language Processing • Github Repository • Video: Deploying Digital Human Readers Leveraging HPCC Systems Video: NLP++ ECL Plugin • Visual Text Open Source Website • NLP++ Language Extension • Formal language description • Learning ECL documentation and on-line training courses.