KLPT – Kurdish Language Processing Toolkit

Sina Ahmadi
Insight Centre for Data Analytics, National University of Ireland Galway, Ireland

**Objectives**
- Describe the Kurdish language as a less-resource language
- Carry out a survey on what has been done in Kurdish language processing
- Present a toolkit in Python for Kurdish language processing
- Address basic language processing tasks for Kurdish

**Introduction**
Despite the recent advances in applying language-independent approaches to various natural language processing tasks thanks to artificial intelligence, some language-specific tools are still essential to process a language in a viable manner.

Although there is a plethora of performant tools and specific frameworks for NLP, such as NLTK, Stanza and spaCy, the progress with respect to less-resourced languages is often hindered by not only the lack of basic tools and resources but also the accessibility of the previous studies under an open-source license.

**The Kurdish Language**

Kurdish is a less-resourced Indo-European language spoken by 20-30 million speakers in the Kurdish regions of Iran, Iraq, Turkey and Syria [1]

**Current state of KLP**
We reviewed the scientific publications that directly address an issue in these fields. We analyze previous publications from the following two perspectives:
- **Applicability**: Does the paper propose an approach or methodology that can be applied to solve the same problem in the other dialects of Kurdish? the majority
- **Open-source**: Does the paper provide the discussed resource or tool under an open-source license? only 18 out of 53

**KLPT modules**
KLPT is implemented in Python and is composed of four core modules with specific tasks without using any external NLP library:
- **Preprocess**: Handle diversities in scripts and orthographies in a formalized way
- **Transliterate**: transliterating the Arabic based and Latin based scripts of Kurdish
- **Stem**: spelling error detection and correction, stemming, lemmatization and morphological analysis
- **Tokenize**: tokenize words, sentences and multi-word expressions

**Lessons learned**
- Don’t reinvent the wheel: Release your project under an open source license
- Create community-driven initiatives
- Raise awareness by promoting good practices in content creation on the Web

**References**

**Use the tool**
This project is publicly available under a CC BY-SA 4.0 license. Find out more at:
- Toolkit: https://github.com/sinaahmadi/klpt
- Documentation: https://sinaahmadi.github.io/klpt/