Click here to be redirected to the virtual room of the Project Expo.

Project abbreviation: Microservices

Project name: Microservices at your Service - Bridging the gap between NLP Research and Industry

Project coordinator: Lingsoft

Project consortium: Lingsoft, Gradiant, Reykjavik University, University of Tartu

Funding: Connecting Europe Facility (CEF)

Project duration:1.3.2021-28.2.2023

Main key words: NLP tools, Microservices, Open source tools

Background of the research topic: Ever since the META-NET white paper series on NLP coverage in the languages of the European Union/EEA, it has been recognized in Europe that the existing large American corporations cannot be expected to provide language technology solutions for many of the smaller languages in Europe. National initiatives in e.g. Finland , Sweden , Estonia , Iceland and Spain, have therefore been developed to address this situation by supporting research and development of NLP tools and services in the local languages.

Goal of the project: Our mission is to simplify finding and using the many superb open source speech and language processing tools that the European Research Community has to offer. The European Union's Connecting Europe Facility has given us support to fill the ELRC-SHARE and European Language Grid with such tools! We are taking yet another step towards fulfilling the vision of a European digital single market.

Project abstract: The Microservices project strives to make existing tools available through the ELRC-SHARE and ELG repositories. The tools we target are the speech and language technology (NLP) tools stemming from the research institutions in languages in our targeted areas: Spanish, Portuguese, Icelandic, Faroese, Danish, Finnish, Swedish, Norwegian, Northern Sami, Estonian, Latvian and Lithuanian.

Universities and other research institutions in Europe, and sometimes companies, are nowadays often publishing open source software on various platforms, primarily GitHub. In the best case, these software packages are associated with research papers, usage instructions, and linked also to other sharing platforms, such as the CLARIN network or META-SHARE. But, often GitHub is the only place, and even if the license is commercially permissive it's difficult to determine if the tool even works as intended. The threshold to using these tools can also be high due to the reliance on various Unix dependencies that may not be compatible between any two tools developers/researchers would like to combine.

Our project is assisting in making these, sometimes very good, NLP tools more accessible to a larger audience of software developers, as well as researchers. We both investigate and search for potential tools, test them when test data is available and then dockerize a selection of tools adding an industry standard API to these services. We then share the docker images via the CEF ELRC-SHARE and ELG platforms.

As a result, we expect that these platforms are enhanced with a large set of good or great NLP research tools. The docker images and API access facilitate the utilisation of these tools by researchers and software/service developers. These docker images with exposed APIs are a form of microservices that facilitate combining many independent tools in a modular microservice architecture. This type of architecture has certain advantages when it comes to integrating open

Click here to be redirected to the virtual room of the Project Expo.

source tools, in particular, it can facilitate quickly adding support for new languages, or quickly replace an old module with a state-of-the-art module.

Project website: https://www.lingsoft.fi/en/microservices-at-your-service-bridging-gap-between-nlp-research-and-industry