T5: Natural Language Processing for MIR

An increasing amount of musical information is being published daily in media like Social Networks, Digital Libraries or Web Pages. All this data has the potential to impact in musicological studies, as well as tasks within MIR such as music recommendation. Making sense of it is a very challenging task, and so this tutorial aims to provide the audience with potential applications of Natural Language Processing (NLP) to MIR and Computational Musicology.

In this tutorial, we will focus on linguistic, semantic and statisticalbased approaches to extract and formalize knowledge about music from naturally occurring text. We propose to provide the audience with a preliminary introduction to NLP, covering its main tasks along with the stateof-theart and most recent developments. In addition, we will showcase the main challenges that the music domain poses to the different NLP tasks, and the already developed methodologies for leveraging them in MIR and musicological applications. We will cover the following NLP tasks:

Basic text preprocessing and normalization

Linguistic enrichment in the form of partofspeech tagging, as well as shallow and ↵dependency parsing.

Information Extraction, with special focus on Entity Linking and Relation Extraction.

Text Mining

Topic Modeling

Sentiment Analysis

Word Vector Embeddings

We will also introduce some of the most popular python libraries for NLP (e.g. Gensim, Spacy) and useful lexical resources (e.g. WordNet, BabelNet). At the same time, the tutorial analyzes the challenges and opportunities that the application of these techniques to large amounts of texts presents to MIR researchers and musicologists, presents some research contributions and provides a forum to discuss about how address those challenges in future research. We envisage this tutorial as a highly interactive session, with a sizable amount of hands-on activities and live demos of actual systems.

Sergio Oramas received a degree in Computer Engineering by the Technical University of Madrid in 2004, and a B.A. in Musicology by the University of La Rioja in 2011. He is a PhD candidate at the Music Technology Group (Pompeu Fabra University) since 2013, holding a “La Caixa” PhD Fellowship. His research interests are focused on the extraction of structured knowledge from text and its application in Music Information Retrieval and Computational Musicology.

Luis Espinosa-Anke is a PhD candidate at the Natural Language Processing group in at Pompeu Fabra University. His research focuses in learning knowledge representations of language, including automatic construction of glossaries; knowledge base generation, population and unification; and automatic taxonomy learning. He is Fulbright alumni, “laCaixa” scholar, and member of the Erasmus Mundus Association as well as the European Network of eLexicography.

Shuo Zhang is a PhD candidate in Computational Linguistics at Georgetown University, USA, and a collaborator/researcher at the Music Technology Group, Universitat Pompeu Fabra. He has worked in both text (NLPinformation extraction) and sound (speech processingtimeseries data mining in speech prosody) aspects of computational linguistics and their applications in MIR. His past and current projects include areas such as coreference resolution, search and visualization of multilayered linguistic corpora, text mining & topic modeling in MIR, temporal semantics, timeseries mining in speech and music, etc. Shuo holds B.Sci. from the Peking University, M.A. from the Department of Music, University of Pittsburgh, and M.Sci. in Computational Linguistics from Georgetown University.

Horacio Saggion is Profesor Agregado at the Department of Technologies, Universitat Pompeu Fabra. He holds a PhD in Computer Science from Université de Montréal (Canada). He is associated to the Natural Language Processing group where he works on automatic text summarization, text simplification, information extraction, text processing in social media, sentiment analysis and related topics. His research is empirical combining symbolic, patternbased approaches and statistical and machine learning techniques. Before joining Universitat Pompeu Fabra, he worked at the University of Sheffield for a number of UK and European research projects developing competitive human language technology. He was also an invited researcher at Johns Hopkins University in 2011. Horacio has published over 100 works in leading scientific journals, conferences, and books in the field of human language technology.