# Natural Language Processing
# Italian Language Tokenizer with Emoji and Emoticons Support

Daniele Perrella

December 19, 2022

# Contents

# Chapter 1

# Introduction

## 1.1

# Chapter 2

# Language Tokenization

## 2.1 Literals

To support the Italian language, the default configuration of the tokenizer supports all the letters available for Italian, including accented ones: *à è é ì ò ù*

## 2.2 Punctuation

There are two rules for Italian punctuation, one including all the punctuations, and a second one (placed right before the generic one) to support exclusively the "..." punctuation.

# Chapter 3

# Domain Support

At the time, since almost every person has a email address, that's why there is a dedicated tokenization rule to tokenize emails. Email domain is not the only kind of domain supported, indeed there is also the support for classic URLs

## 3.1

# Chapter 4

# Tokenization for Numericals

## 4.1 Numbers, Floating Point and Scientific Notation

## 4.2 Operations

# Chapter 5

# Emoticons Support

An important featur included in this tokenizer, is support for emoticons, with an exaustive list of supported emoticons. To achieve such result, the emoticon regex has been created based on the wikipedia emoticons list, which cam be found at the following link

## 5.1

# Chapter 6

# Emoji Support

## 6.1

# Chapter 7

# The Output

## 7.1

# Chapter 8

# Custom Configurations

## 8.1 Adding custom configurations

METTERE LINKS