https://doi.org/10.2352/issn.2168-3204.2019.1.0.18

Preprocessing pipeline for Italian Cultural Heritage multimedia datasets

Maria Teresa Artese, Isabella Gagliardi – IMATI – CNR, Via Bassini 15, 20133 Milan, Italy email: {teresa, isabella}@mi.imati.cnr.it

Abstract

Preprocessing is an important task and a fundamental step in Information Retrieval, Text Mining, Natural Language Processing (NLP). While datasets in the English language can rely on well-established tools and methods for text preprocessing, the situation for the Italian language is more nuanced, due to a sum of factors, not least that fewer experiments and studies were made, and algorithms developed. Here we present an experimentation, a work in progress whose purpose is to define a pipeline able to preprocess texts. The different steps of the pipeline have been implemented and tested individually on Cultural Heritage datasets. The results obtained have been evaluated in the context of unsupervised automatic keyword extraction algorithms, such as RAKE or TextRank.

Introduction

Preprocessing is an important task and a fundamental step in Information Retrieval (IR), Text Mining, Natural Language Processing (NLP) and Automatic Unsupervised Keyword Extraction Algorithms [1][1][3][4]. Data mining, text analysis algorithms, such as automatic keyword extraction tools on text without any preprocessing obtain rather scarce results. Datasets written in English can rely on tools and methods for text preprocessing, using well-established methodologies to obtain a set of terms, in canonical form, eliminating, where appropriate, stoplist words and terms belonging to defined grammatical categories. Tools for tokenization, lemmatization/stemming, POS tagging tools, and almost standard stopword lists can be easily identified and applied to obtain acceptable results.

The situation for the Italian language is more nuanced.

First of all, Italian grammar, morphology, and syntax, is more complicated. As an example, some part of speech, as nouns and adjectives, are variable, that is, they are modified according to the number (singular and plural) and gender (feminine and masculine). The adjective beautiful (in Italian bello), for the positive grade, takes shapes 'bello' (masculine/singular), 'bella' (feminine/singular), 'belli' (masculine/plural), and 'belle' (feminine/plural).

Secondly, different writing styles [5] can lead to very different languages, even extremely popular ones. Datasets for cultural heritage can have different languages:.

1. the "ministerial" language of catalog cards on cultural heritage is formal, using rather rare words, mixed with technical terms, suitable for insiders;
2. description and transmission of traditions and fairs of intangible heritage is done by the communities and for the communities: so the language is common, with some dialectal terms;
3. sometimes datasets are written using a colloquial and informal language: for example, this is the case of food recipes, where directions and instruction are regulatory texts, with verbs in the imperative form.

Writing styles, in addition to a different vocabulary, follow the grammatical and syntactic rules in a different way: from strict observance of the formal register to a more relaxed way of writing of the informal one.

Finally, for the Italian language fewer experiments, studies and algorithms can be counted.

The purpose of the experimentation is to define a pipeline able to produce, for each text, together with the original version, a list of terms (also repeated), in canonical forms, such as the entries of the specific vocabulary of the dataset. The results will be the input for automatic unsupervised keyword extraction algorithms.

We present here a work in progress with the aim of defining a pipeline for the preprocessing of Italian datasets, in the context of tangible and intangible cultural heritage, to obtain the headwords. Different algorithms and options lead to different terms sets.

Related works

Preprocessing methods plays a very important role in text mining techniques and applications. It is the first step in the text mining process. [1] discusses the text mining preprocessing techniques, i.e. stop words elimination and stemming algorithms. [2] presents efficient preprocessing techniques and argues the need of Text Preprocessing in NLP System for two reasons: i) to reduce indexing (or data) file size of the text documents and ii) to improve the efficiency and effectiveness of the IR system. [3] investigates the impact of widely used preprocessing tasks including tokenization, stop-word removal, lowercase conversion, and stemming comparatively in two different languages, Turkish and English.

Approach

First, we have identified some Italian datasets, one of which are also available in English (a translation, made by experts from Italian to English), related to Cultural Heritage.

We made a first 'manual', exploratory data analysis to understand the steps to be undertaken to extract the terms in canonical form, along with the POS (part-of-speech) tagging [6], that is their grammatical category (noun, verb, adjective, etc.). Depending on the datasets, the grammatical styles can vary greatly, and consequently the results.

The pipeline identified has to be general enough to be applied to all datasets, in any language and style, to obtain a "clean" texts or set of headwords, and is composed of tokenization, POS (part-of-speech) tagging, lemmatization/ stemming and removal of stopwords. In the following each step will be described in details.

Tokenization is the process of decomposing a text, considered as a continuous set of words – or a string-, into a set of terms, composed of a single or compound words. In Italian, for example, the apostrophe is a character of division of words. It is obligatory in case of elision, such as a *bell'amico* (good friend) or *quest'alunna* (this (girl) student). In these cases, two distinct words should be identified. Moreover, regardless of language, different methods can be applied to identify compound words. For example, you can use statistical methods that identify groups of words frequently together and consider them as n-gram or use headlines of Wikipedia page as a dictionary or identify their presence in WordNet.

POS (part-of-speech) tagging is the process of associating the corresponding grammatical category to every single word.

Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential *there*
FW	Foreign word
…	…
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
…	…
PRP	Personal pronoun
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	*to*
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
…	…

Table 1: Alphabetical list of part-of-speech tags used in the Penn Treebank Project: from https://www.ling.upenn.edu

This process requires understanding the structure of the sentence because, in Italian as in English, the same word can belong to different grammatical categories (e.g. be an adjective, a noun or a verb) depending on the composition of the sentence.

Normalization: the process of lemmatization/stemming reduces the words to the lemma or stem. The Porter algorithm is the most widespread stemmer for the English language. For example, according to the Porter stemmer [7] or Snowball algorithm [8], ceremony, as ceremonies, become ceremoni. The lemmatizes, on the other hand, should bring the various inflected forms to a single lemma, that is, the voice present in the vocabulary. For example ceremony, ceremonies should lead to ceremony; or listening, (he) listens, (they) listened, ... to "to listen". For Italian, there are no similar "universal" algorithms.

Stopword: traditionally, in information retrieval systems, texts to be indexed are provided without stopwords, those words that are so frequent that they have no meaning: for example and, in, is/are, …. Obviously, stoplists are different for Italian and English. Also, depending on the context of use, they should be customized.

The results of the pre-processing are lists of headwords or vocabularies, to be used - together with the original data - for the next steps of text analysis.

Pre-processing pipeline

For the datasets in English and Italian, the pre-processing phases are the same, while the specific tools and algorithms to be applied differ. In general, in this experimentation, we have always privileged the standard tools, using the version developed for the Italian language. For the English language, data sets were processed using well known and consolidated tools, such as Stanford's core NLP suite Natural Language Toolkit of Python[9], or the NLTK package [10] with PENN Treebank as a POS tagger and tokenizer [11]. Table 1 reports a subset of Treebank POS tag for the English language.

http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
(Copyright Prof. Achim Stein, University of Stuttgart)

ABR	abbreviation
ADJ	adjective
ADV	adverb
CON	conjunction
DET:def	definite article
DET:indef	indefinite article
FW	foreign word
INT	interjection
LS	list symbol
NOM	noun
NPR	name
NUM	numeral
PON	punctuation
PRE	preposition
PRE:det	preposition+article
PRO	pronoun
PRO:demo	demonstrative pronoun
PRO:indef	indefinite pronoun
…	
VER:cimp	verb conjunctive imperfect
VER:cond	verb conditional
VER:cpre	verb conjunctive present

Table 2: Italian tagset used in the TreeTagger parameter file

For **tokenization,** the Italian language makes great use of apostrophes and accented letters:specific treatments to identify and replace accented letters, apostrophes, quotation marks and indicators of "direct speech" are essential to obtain a clean text for subsequent processing. For example, in order to provide the POS tagging and lemmatization/stemming algorithms with all the necessary information, punctuation marks must be maintained. Tests have shown that NLTK tools were adequate in dividing texts into sentences and sentences into single or compound words.

Initially, for **lemmatization and POS tagging**, the same tools were also used for Italian - where they were present - but the results obtained were not useful due to the presence of too many errors and inaccuracies, both for the POS tagging and for the lemmatization.

Subsequently, tools developed specifically for the Italian language were tested, for example, the Italian version of Snowball [8], Pattern python package specific for Italian (Pattern clips 2.6) or LinguA [12][13].

In the end, we used and integrated two different tools:

1. TreeTagger [14] [15] is a free tool developed by Helmut Schmid in the TC project at the Institute for Computational Linguistics of the University of Stuttgart, using the Italian standard tagset. Table 2 shows its grammatical categories, taken from “Italian tagset used in the TreeTagger”. A second tagset was also tested, but the results were not as satisfactory.
2. Spacy [16] is a tool designed for Natural Language Processing, and offers statistical neural network models for many languages, including Italian, for POS tagging, lemmatization and NER (Named Entity Recognition).

Each of the tools makes errors in the POS tagger and lemmatization, but by integrating them you can minimize the errors and get acceptable results, as will be briefly discussed below.

Stopword lists: In order to eliminate or limit noisy words, appearing too frequently to be of some use, stopword lists have to be used. There are various lists of generic stopwords, both for Italian and English. For example, in Ranks NL website[1] there are several stoplists for English, while only one for Italian[2]. NLTK, Spacy, Pattern package, … have stopword lists, general purpose.

Preliminary Results and evaluation

We run the system on different datasets of intangible cultural heritage data, with texts in Italian and one in English, and one of cooking recipes [17][18][19].

AESS (Archivio di Etnografia e Storia Sociale) database stores information related to the oral history of the Lombardy region. **IntangibleSearch** is a dataset of Intangible Cultural Heritage of Lombardy Region and Alp territories. Languages in which data are available are Italian and English.

Cookit is a dataset of traditional Italian recipes, written in Italian language, taken from well-known culinary websites, as Giallozafferano[3] or cucinaitaliana[4]. Writing styles of recipes is regulatory texts, with verbs in the imperative form. Only the results per Cookit are reported here.

Treetagger and Spacy have been tested for lemmatization and POS tagging. Treetagger performs better than Spacy for lemmatization and POS tagging, being able to correctly identify and interpret verbs, nouns, and adjectives. In table 3 and 4 some results for Italian sentences taken from CookIt dataset. ‘Dividetela’ (in English ‘separate it’) is a word composed of a verb in imperative form and a personal pronoun. Treetagger correctly identifies the verb, but miss the personal pronoun; Spacy fails to recognize the flexed form and bring the verb back to the to + infinitive form.

Original	**POS**	**Lemma**	**Correct?**
dividetela	VER: impe	dividere	**Yes**
acqua bolle	NOM NOM	acqua bolla	**No**: bolle is a verb
torta sbrisolona	ADJ NOM	torto sbrisolona'	**No** Torta is a noun

Table 3: TreeTagger results

Original	**POS**	**Lemma**	**Correct?**
prendete	VERB Tense=Pres|VerbForm=Fin	prendere	Yes
dividetela	VERB Tense=Pres|VerbForm=Fin	dividetela	**No**:Lemma: dividere
tanto	ADV	tangere	**No** Tanto: invariant

Table 4: Spacy results

‘Acqua bolle’, in English ‘water boils’, is composed of a noun and a verb. TreeTagger, mistakenly, locates 2 neighboring nouns, instead of a name and a verb. In this case, Spacy, with its ability to identify the grammatical and syntactic structure of the sentence, correctly identifies the noun (subject) + verb. Table 3 and 4 tables show other errors or inaccuracies using TreeTagger or spacy.

Spacy package has a function that displays every single sentence as a tree, with syntactic and grammatical dependencies. Figure 5 shows the tree for the sentence: “Quindi prendete la zucca e dividetela in parti per eliminare più facilmente i semi e i filamenti interni” (“then take the pumpkin and divide it into parts to more easily eliminate seeds and internal filaments”).

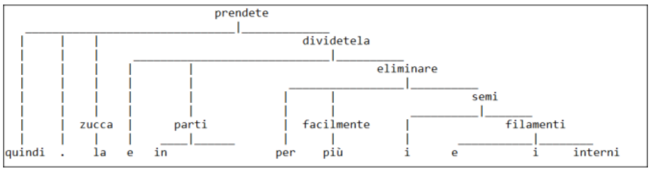

Table 5: Sentence Tree of Spacy

The aim of our work is to prepare ‘clean’ texts to maximize and improve results by applying unsupervised automatic extraction algorithms for keywords, such as tf-idf, RAKE or TextRank [22][23][24]. We have therefore applied different methods of pre-processing to the datasets to test how the different sets of terms produce results (keywords or keyphrases). First, we considered the documents and we tokenized and lemmatized (or not), producing three elaborations of the original data:

Original data: no tokenization steps applied (orig)

Tokens: only tokenization of texts (tokens)

Tokens more than 1: after tokenization + lemmatization, only tokens appearing more than once are kept (tokens_moreonce).

POS tagging and lemmatization were performed by integrating TreeTagger and Spacy and minimizing their errors. Also, since some inaccuracies remain, we compared the results with the synsets of Italian MultiWordNet, in the same grammatical category as the starting term [20][21].

After POS tagging processing, tokens that are names (_nouns), and tokens that are names and adjectives (_nouns_adj)

[1] https://www.ranks.nl/stopwords

[2] https://www.ranks.nl/stopwords/italian

[3] Giallozafferano http://giallozafferano.it

[4] Cucina Italiana http://www.cucinaitaliana.it

are kept on the three processed datasets (orig, tokens, tokens, tokens_moreonce).

In our test, we used the list of predefined NLTK stopwords, easily configurable, adding, if useful, custom words, specific to the context of use.

Different preprocessing methods select a set of different terms, leading to different keywords/keyphrases extracted from the algorithms. In Table 6, it can be observed that the most frequent terms in datasets are different depending on the pre-treatment methods used. Each step has been executed several times, testing different algorithms and/or options and/or options and/or stopwords/grammatical categories.

Evaluation is difficult and depends on the purpose of the system. Once lemmatization and POS tagging errors have been eliminated or minimized, the effectiveness of the results can only be assessed by the adequacy of the extracted keywords. Some preliminary evaluation demonstrates that applying TextRank and RAKE to tokens or tokens_moreonce result in a very poor set of keywords, compared to those obtained on unprocessed texts. This depends on how the TextRank and RAKE algorithms work. Tf idf works well with tokens, and tokens_moreonce, as long as you keep the cardinality of the terms in the text.

Dataset\pre-proc. method	orig_nouns_adj	tokens_nouns
Cookit	minuto acqua olio impasto sala|sale fuoco pasta forno	acqua olio fuoco impasto pasta sala farina forno
Intangible English	people procession year small time village group traditional	use people year procession place time day man

Table 6: Top Most Frequent Terms

The prototype was developed in Python, using standard packages like Numpy, Matplotlib, Pandas and other more specific ones for processing of textual data such as NLTK [10], Gensim [25], Skit-learn [26], TreeTagger [14][15] and Spacy [16].

Conclusions

In the paper, we present a possible pipeline for the preprocessing of texts, for their subsequent use for the unsupervised automatic extraction of keywords. We have shown the steps, the tools to be used for datasets in the Italian language. The work is still in progress and the different phases of the pipeline have been implemented and tested individually, with different options. For example, headwords can be extracted based on n-grams, words with specific parts of voice tags (nouns, verbs, adjectives), or different stopword lists.

Future developments will include:

- test additional tools for the Italian language
- a broader evaluation of the results.

References

[1] Vijayarani, S., Ms J. Ilamathi, and Ms Nithya. "Preprocessing techniques for text mining-an overview." International Journal of Computer Science & Communication Networks 5.1 (2015): 7-16.

[2] Kannan, S., and Vairaprakash Gurusamy. "Preprocessing Techniques for Text Mining." (2014).

[3] Uysal, Alper Kursat, and Serkan Gunal. "The impact of preprocessing on text classification." Information Processing & Management 50.1 (2014): 104-112.

[4] Salton, Gerard, and Christopher Buckley. "Term-weighting approaches in automatic text retrieval." Information processing & management 24.5 (1988): 513-523.

[5] Strunk, William. The elements of style. Penguin, 2007.

[6] Part-of-speech tagging (POS tagging): https://en.wikipedia.org/wiki/Part-of-speech_tagging (last consulted 10/10/2018)

[7] Porter, Martin F. "An algorithm for suffix stripping." *Program* 14.3 (1980): 130-137.

[8] Porter, Martin F. "Snowball: A language for stemming algorithms." (2001).

[9] CoreNLP, Stanford. "a suite of core NLP tools." URL http://nlp. stanford. edu/software/corenlp. shtml (Last accessed: 2018-09-06)

[10] Bird, Steven, and Edward Loper. "NLTK: the natural language toolkit." Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. Association for Computational Linguistics, 2004.

[11] Marcus, Mitchell, Beatrice Santorini, and Mary Ann Marcinkiewicz. "Building a large annotated corpus of English: The Penn Treebank." (1993).

[12] Dell'Orletta, Felice. "Ensemble system for Part-of-Speech tagging." Proceedings of EVALITA 9 (2009): 1-8.

[13] Attardi, Giuseppe, et al. "Accurate dependency parsing with a stacked multilayer perceptron." Proceedings of EVALITA 9 (2009): 1-8.

[14] Schmid, Helmut. "Treetagger| a language independent part-of-speech tagger." Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart 43 (1995): 28.

[15] Schmid, H., et al. "The enriched TreeTagger system." proceedings of the EVALITA 2007 workshop. 2007.

[16] Industrial-Strength Natural Language Processing http://spacy.io (last consulted 13/03/2019)

[17] Intangible Search: https://intangiblesearch.eu (last consulted 10/10/2018)

[18] Archivio di Etnografia e Storia Sociale (AESS) https://aess.regione.lombardia.it (last consulted 10/10/2018)

[19] CookIT https://arm.mi.imati.cnr.it/cookIT (last consulted 10/10/2018)

[20] Miller, George A. "WordNet: a lexical database for English." *Communications of the ACM* 38.11 (1995): 39-41.

[21] Fellbaum, Christiane. "WordNet." *The Encyclopedia of Applied Linguistics* (2012).

[22] Hasan, Kazi Saidul, and Vincent Ng. "Automatic keyphrase extraction: A survey of the state of the art." Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vol. 1. 2014.

[23] Mihalcea, Rada, and Paul Tarau. "Textrank: Bringing order into text." Proceedings of the 2004 conference on empirical methods in natural language processing. 2004.

[24] Rose, Stuart, et al. "Automatic keyword extraction from individual documents." Text Mining: Applications and Theory (2010): 1-20.

[25] Rehurek, R., and P. Sojka. "Gensim–python framework for vector space modelling." NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic 3.2 (2011).

[26] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." Journal of machine learning research 12.Oct (2011): 2825-2830.

Author Biography

Maria Teresa Artese *took her degree in Computer Science at the University of Studies of Milan in 1990. She has been working at the CNR since 2000. Now she works at IMATI – CNR Unit of Milan. Dr. Artese major areas of work are functional analysis and software development, technical support, database structuring and development, dynamic web database sites, design, and implementation. Recently she has focused his research on multimedia information systems development and integration of information, also available as open linked data, from different sources. On these topics, she has several national and international publications, and she has been working in national and international research projects.*

Isabella Gagliardi *took her degree in Physics at the University of Studies of Milan in 1985. She has been working at the CNR since 1986. Now she works at IMATI – CNR Unit of Milan. Dr. Gagliardi major areas of research include Hypermedia Information Retrieval models and methodologies, automatic generation of hypertextual links between text-text, text-image, and audio-audio, dynamic web-based database design and implementation, and clustering algorithms. She has worked on the development of multimedia information systems available on the web and development of participative online platform. Recently she has focused her work on data mining, information retrieval, and text summarization.*