API - Autocomplete

Introduction

The autocomplete API is a *Backend* service that returns a list of the suggestions most likely to complete an initial *input*.

The API is based on a principle of REST access, but since it only deals with searches, only the GET method is available.

Access

API root URL

All API calls start with the url: https://apife.ine.pt

Positioning URL

From the point of view of the consumer of the service, the segment "**dic**" (which indicates that you want to access a dictionary) follows. Finally, the identifier segment of the dictionary that you intend to use in the autocomplete.

At this stage, the available dictionaries are:

- Occupation: https://apife.ine.pt/dic/CPP2010
- Economic Activity: https://apife.ine.pt/dic/CAEREV3
- Higher Education: https://apife.ine.pt/dic/CURSOSUPC2021

Use

There are two use cases available for consumption:

Prefetch
/preftech (https://apife.ine.pt/dic/{dictionary id}/prefetch)

For the identified dictionary, returns a list of the most frequent entries. It can be invoked and cached in the autocomplete client.

Search

?q=XXXX (https:// apife.ine.pt/dic/{dictionary id}/?q={query_text})

Example: https:// apife.ine.pt/dic/CPP2010/?q=baila

Returns the list of occupations that contain the particle "baila". As "Bailarino" is not a frequent occupation, it will not be included in the list obtained in the "prefetch".

Structure

Prefetch and lookup return *arrays* in *JSON* with objects that have the structure:

[{ c : "AAA", d : "BBBB", t : "CCCCC"}, ...]

In each element:

- "**c**" contains the code;
- "**d**" the designation to be presented as a suggestion;
- "**t**" is a *string* of words separated by spaces that we will call *tokens*.

The order of the elements in the *array* reflects their ordering by relevance (the most relevant ones come at the beginning).

Dictionaries

The basis for the construction of the Dictionaries (only available in Portuguese), beside the official coding lists (CAE Rev3, CPP 2010, CNAEF), is sourced from all the manual coding history of more than 30 statistical operations carried out over about 8 years within the scope of the Household Surveys. At the time, the total number of interviews conducted exceeded 600,000. All expressions (1) with a frequency equal to or greater than 10 and a coding consistency of 90% and (2) with a frequency equal to or greater than 5 and a coding consistency of 100% were considered eligible to enrich the classifiers. Then a metric distance was calculated between the expressions already existing in the classifier and the rest of the history. The Optical String Alingment - an extension of the Levenshtein measure - was used to calculate the distance at an interval of 1 to 3. After validation, the expressions that were equivalent in meaning, but distinct in spelling, were integrated into the dictionaries.
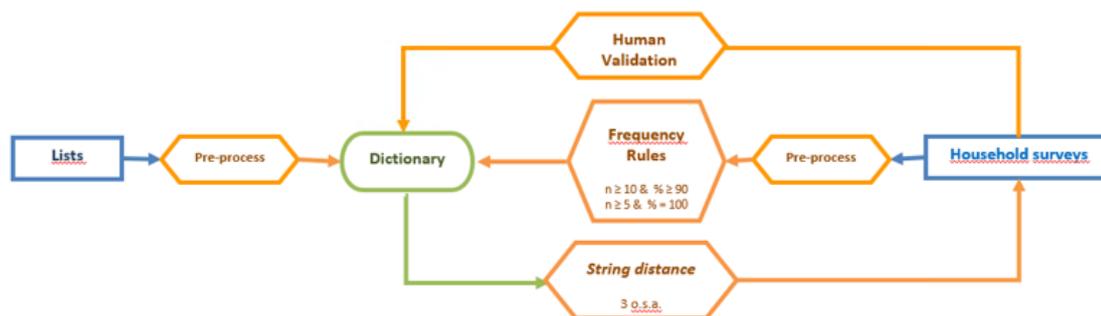
Figure 1- Dictionary Creation Schema

Nomenclatures

As mentioned, the API classifies expressions based on three nomenclatures

- Occupation: https://apife.ine.pt/dic/CPP2010
- Economic Activity: https://apife.ine.pt/dic/CAEREV3
- Higher Education: https://apife.ine.pt/dic/CURSOSUPC2021

The SMI Version used for the classification of Occupation is: V02014- Portuguese classification of professions, CPP 2010 which can be consulted at: https://smi.ine.pt/Versao/Detalhes/2014?modal=1

The SMI Version used for the classification of Economic Activity is: V00554 - Portuguese classification of economic activities, revision 3 that can be consulted at: https://smi.ine.pt/Versao/Detalhes/554?modal=1

The SMI Version used for the classification of Higher Education Courses is: V04477 - Higher Education Qualifications, 2020 (Courses - IINQE) which can be consulted at: https://smi.ine.pt/Versao/Detalhes/4477?modal=1