NamSor™
Onomastics

RapidMiner Onomastics Extension, to recognize the gender of names

RapidMiner version: RapidMiner 6.0 or RapidMiner 5.3

Extension version: Extract Gender Operator v0.0.2
API version: GendRE API v0.0.14

Document version: NamSor_RapidMiner_Extension v0.0.2
Document date: June 2014

Contents

About NamSor

NamSor™ is a European vendor of Name Recognition Software. We offer specialized data mining to recognize the origin of personal names in any alphabet / language, with fine grain and high accuracy. NamSor's mission is **to help understand international flows of money, ideas and people.**

Please, reach us at contact@namsor.com or follow us on Twitter

Introduction

If you are reading this tutorial, you probably have already installed RapidMiner and gained some experience by playing around with the enormous set of operators.

At NamSor, we intend to deliver a set of operators for mining **proper names** in all geographies/alphabets/languages/cultures. We decided to start with a *seemingly* simple operator GendRE Gengerizer, to predict the likely gender of a personal name. Other useful operators will come in 2014.

Guessing the gender of name is not as simple as it seems:

- Andrea is a male name in Italy, a female name in the US. Laurence is a female name in France and a male name in the UK or in the US

- name demographics evolve, some names are genderless

- in Chinese or Korean, guessing the gender is almost impossible in Latin script, truly difficult even with the original script

- in most cultures, the gender is 'encoded' in the first name, in others it is encoded in the last name as well (for example, Slavic names, Lithuanian names ...) so you can guess the gender even if you have just the initials (for example, O. Sokolova is most likely a Slavic name and a female name)

- some names are very rare or just 'made up' and yet, because they *sound* like a male name or a female name, their gender is accurately perceived by the people in that same culture

GendRE API goal is to hide this complexity, offer a simple interface and return an optimal result:

api/json/gendre/John/Smith
{"scale":-0.99,"gender":"male"}

Can you guess the result of the following?

api/json/gendre/נתניהו/בנימין/il
api/json/gendre/声涛/周
api/json/gendre/الامرء بي/معـ ين/lb

Currently, we require input names to be properly parsed into a (firstName, lastName) format and our machine learning algorithm will progressively discover how names are parsed in different cultures. When this calibration is complete, we'll offer an even simpler interface.

In RapidMiner, simply connect the GendRE Genderizer operator in your process to infer the gender of a personal name and create new data/new segmentation.

Getting Started

This section will get you started with NamSor Onomastics Extension.

Installation

Pre-compiled extension binaries can be found in GitHub /dist/ directory.

Use RapidMiner MarketPlace:

Simply search for NamSor or 'Extract Gender' in the MarketPlace.

Install manually in RapidMiner 6:

Copy the Extension binary into RapidMiner extension directory. For example, on Windows:

C:\Program Files\RapidMiner\RapidMiner Studio\lib\plugins\rapidminer-NamSor-6.0.002.jar

Install manually in RapidMiner 5.3:

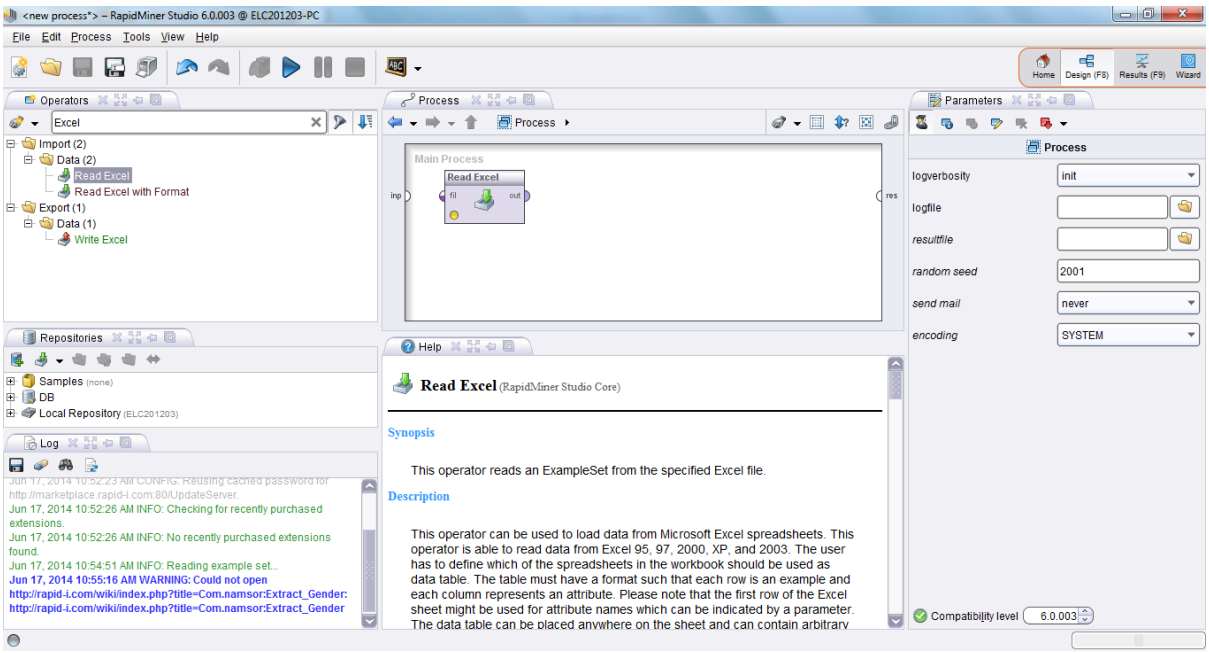
Copy the Extension binary into RapidMiner extension directory. For example, on Windows:

C:\Program Files (x86)\Rapid-I\RapidMiner5\lib\plugins\rapidminer-NamSor-5.3.002.jar

Create your first process

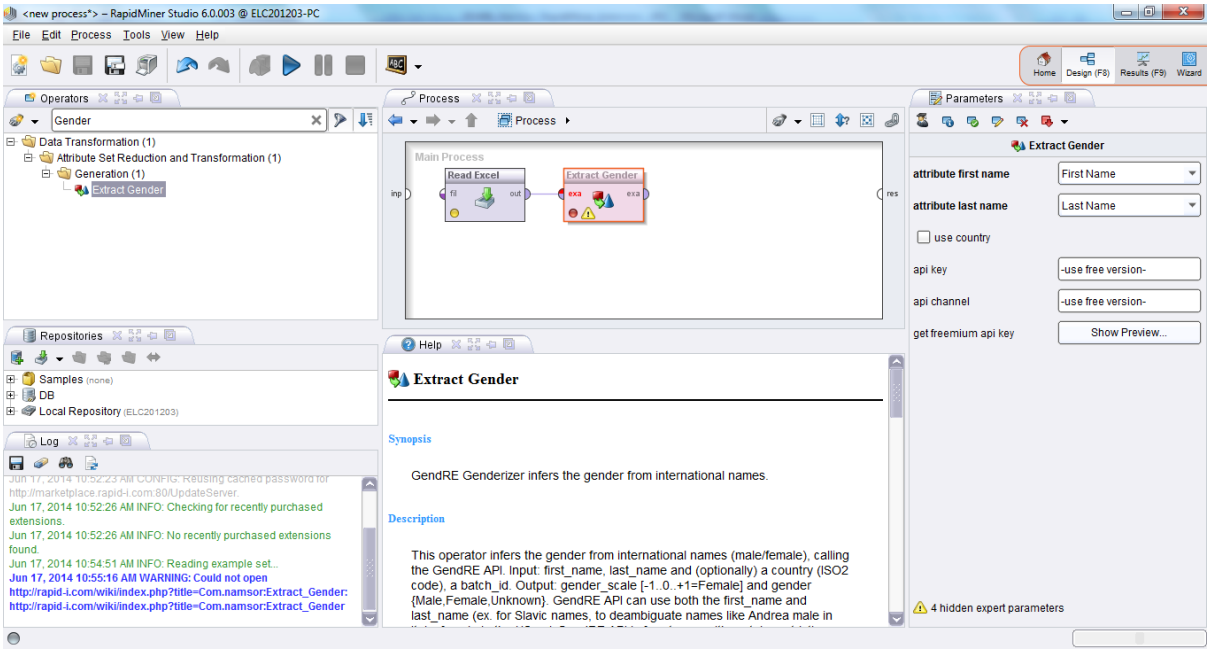
Create a simple Excel document with columns **first_name**, **last_name** and a few contacts.

Drag and drop the Read Excel operator (Import->Data->Read Excel) and launch the Import Configuration Wizard.



Default values should be OK through the wizard, except Attribute should be 'Text' and Encoding should be set to UTF-8 (Unicode, especially required if you would like to genderize Chinese, Russian or Arabic names).

Drag and drop the **Extract Gender** operator (Operators>Data Transformation/Attribute Set Reduction and Transformation/Generation/Extract Gender). Connect the operator with the Excel file and map the attributes.

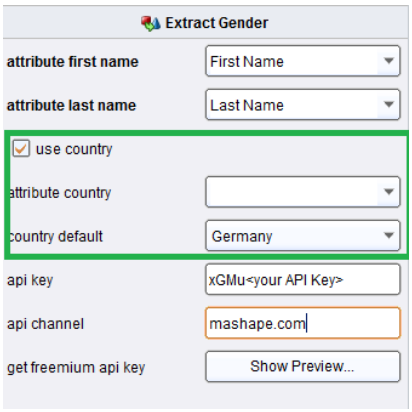


Leave the **api_key** / **api_channel** to use the free GendRE API.

Add a CSV exporter to view the results.

Additional attributes and parameters

Use Country

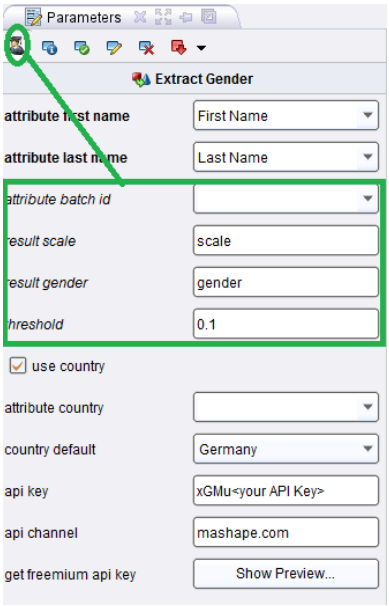


If your data has Country information, you can select the Country attribute on a row-by-row basis to specify which country statistics should be used when predicting gender (ex. Andrea is male in Italy, rather Female in the US).

You can also specify a default Country parameter which will be applied, unless there is a Country specified at the row level.

NB: this information will be inferred automatically in an upcoming API release, by recognizing the cultural origin of the (first_name, last_name) combination.

Expert Parameters



batch_id: If your data is logically grouped (ex. Twitter followers by Twitter user, etc.) you can set a Batch ID to maintain input/output data information.

result_scale, result_gender: Here you can change the default names for result/output attributes.

threshold: This parameter specifies threshold according to which Gender is considered 'Unknown' (evaluating *'Unknown' = abs(scale)<threshold*)

Network/API troubleshooting

In case of network error,

- check that you can access the API from behind your proxy, using your ordinary browser

http://api.onomatic.com/onomastics/api/json/gendre/John/Smith
{"scale":-0.99,"gender":"male"}

- check your RapidMiner proxy configuration in Tools>Preferences>System

Ordering and getting support

Please report any issue with the software or the GendRE API via GitHub

https://github.com/namsor/rapidminer-onomastics-extension/issues

You can upgrade the GendRE API (better precision) and subscribe for commercial support.

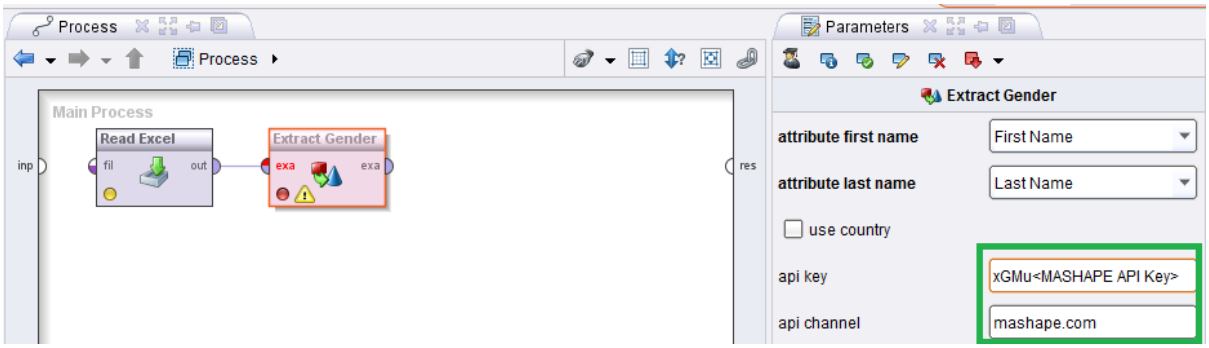
GendRE API Freemium on Mashape

- Register on Mashape.com to obtain an API Key
https://www.mashape.com/namsor/gendre-infer-gender-from-world-names

In RapidMiner, set Extract Gender parameters

- **api_key** : enter your Mashape.com API Key

- **api_channel** : enter **mashape.com/*<your project_name (optionally)>***



- to report an issue or ask a question to our support team, please use

https://www.mashape.com/namsor/gendre-infer-gender-from-world-names#!issues

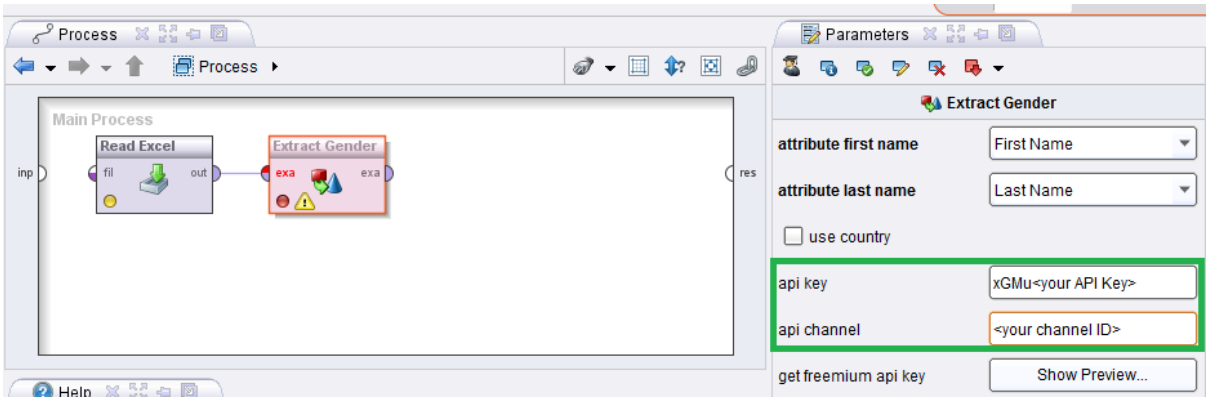
Premium Corporate Customers and RapidMiner MarketPlace Customers

Your will have been provided with the following information:

- **api_key**: enter your API Key

- **api_channel** : enter your Channel and Contract ID, for example

namsor.com/client_id/project_id



Cloud processing & on-premise software

If you need to process very large datasets in very a short time span, start first with a Freemium/Premium offer of GendRE API to prepare and test your process, then contact our professional services via the support tool.

Licensing

Please review our licensing terms,

- the NamSor Onomastics Extension AGPL License,
https://raw.githubusercontent.com/namsor/rapidminer-onomastics-extension/master/LICENSE

- the GendRE API Terms & Privacy Policy,
http://namesorts.files.wordpress.com/2014/02/20140223_gendre_api_v004_terms.pdf

- the RapidMiner MarketPlace Terms