RapidMiner Onomastics Extension, to extract origin and gender of names

RapidMiner version: RapidMiner 6.5-6.0 or 5.3

Extension version: Parse Name, Extract Gender, Extract Origin v6.5.10
API version: NamSor-API_v0.0.23

Document version: NamSor_RapidMiner_Extension v0.0.7
Document date: November 2015

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.**

Our values: we promote diversity, equal opportunity and support the @GenderGapGrader initiative.

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.

List of Operators

Extract Gender Operator

NamSor Gender API (formerly GendRE Genderizer) predicts the likely gender of a personal name. 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

NamSor Gender API goal is to hide this complexity, return an optimal result with a simple interface:

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

Can you guess the result of the following?

api/json/gender/בנימין/נתניהו/il
api/json/gender/声涛/周
api/json/gender/بي/مع ين/مرع الـ/lb

For the free version, we require input names to be properly parsed into a (firstName, lastName) format. The commercial version can handle unstructured names too (Andrea Rossini, Rossini Andrea, Andrea H. PARKER).

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

Extract Gender Editions:

Free Edition	Freemium API Key	Premium Customers
- no registration - unlimited number of calls - 2-digit precision - support via GitHub Ticketing	-1000/month free - parse unstructured names - higher performance & throughput (hundreds of names processed at a time) - full double precision - commercial support	- customized solutions - advanced caching (persistent caching between sessions)
Free	Get your API Key	Get a quote

Extract Origin Operator

NamSor Origin will guess the likely country of origin of a personal name, based on the sociolinguistics of the name (language, culture). This is a coarse grain classification, typically for marketing or social analytics. Finer-grain classification (regional level, ethnicity...) is available for more complex usage like Diversity Analytics or Migration Studies, but requires a specific paper contract.

The method for anthroponomical classification can be summarized as follow: judging from the name only and the publicly available list of all ~150k Olympic athletes since 1896 (and other similar lists of names), for which national team would the person most likely run? Here, the United-States, Australia, etc. are typically considered as a melting pot of other 'cultural origins': Ireland, Germany, etc. and not as a onomastic class on its own.

Extract Origin Editions:

Freemium API Key	Premium Customers
- 100/month free - parse unstructured names - high performance & throughput (hundreds of names processed at a time) - full double precision - commercial support	- customized solutions - advanced caching (persistent caching between sessions)
Get your API Key	Get a quote

Parse Name Operator

Parse Name Operator will try and guess the likely structure and order of personal name (firstName-lastName or lastName-firstName), based on the sociolinguistics of the name (language, culture). It is typically used as pre-processing before using the Extract Gender and Extract Origin operators.
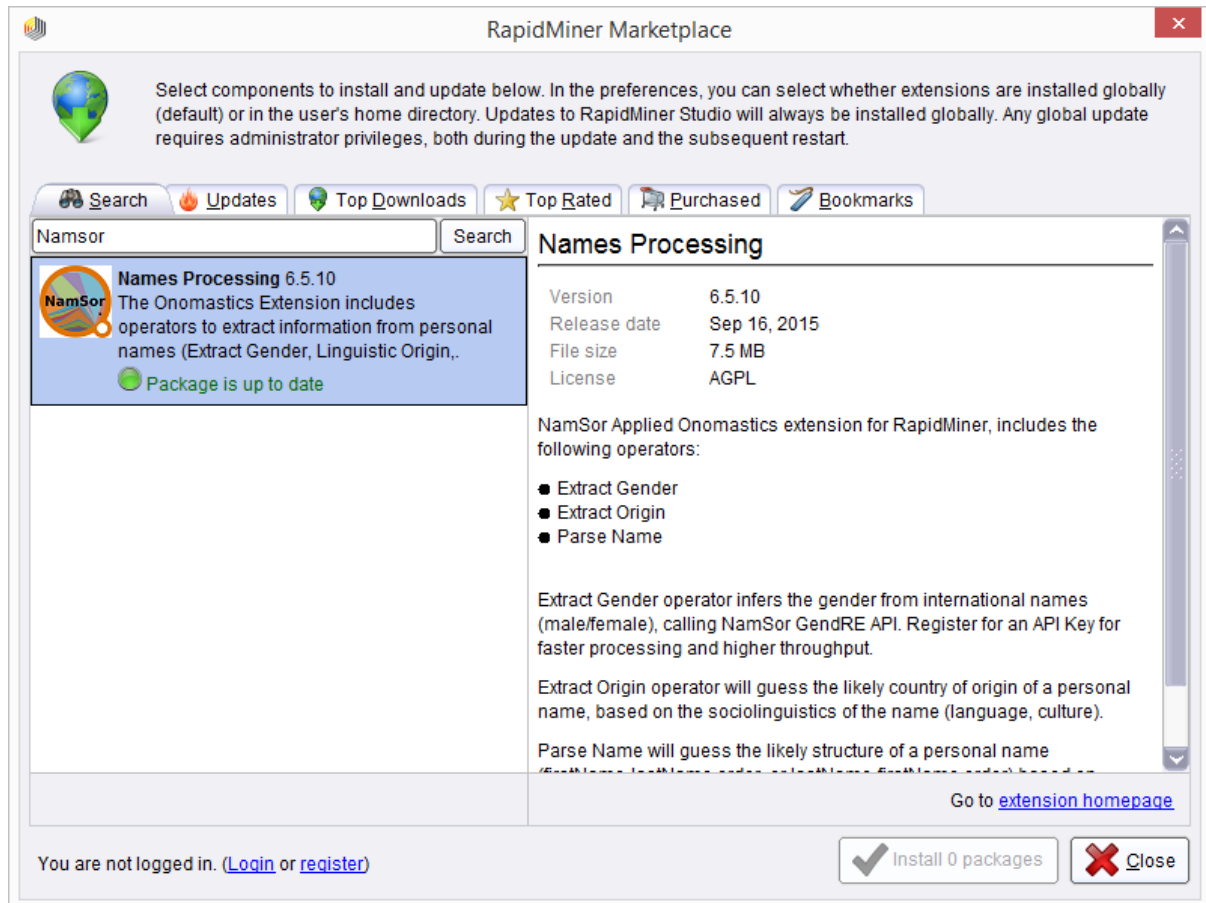
Getting Started

This section will get you started with NamSor Onomastics Extension. You can also view the online [tutorial video](#).

Installation

<u>Use RapidMiner MarketPlace:</u>

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

<u>Install manually in RapidMiner:</u>

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

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

<path>\RapidMiner\RapidMiner Studio\lib\plugins\

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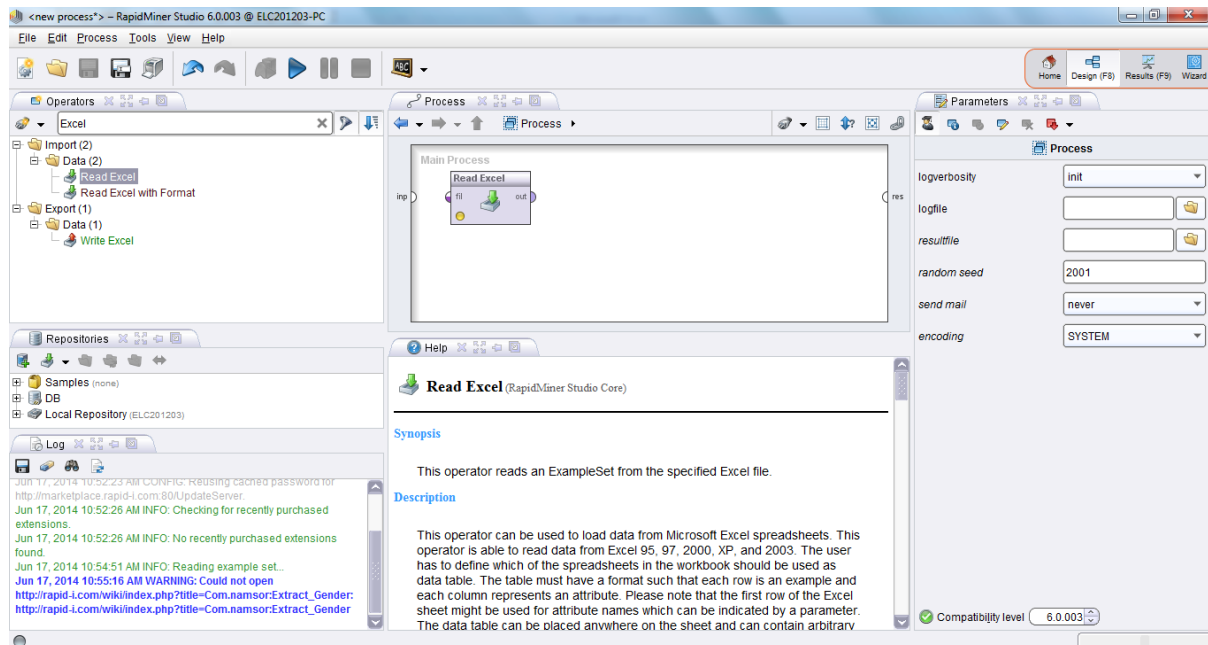
Default values should be OK through the wizard, except Attribute should be 'Text' and Encoding should be set to UTF-8 (Unicode, especially required if you would like to genderize Chinese, Russian or Arabic names).

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.

Known RM Issue [#1805](): you may need to manually write the attribute mapping instead of selecting from the drop box.
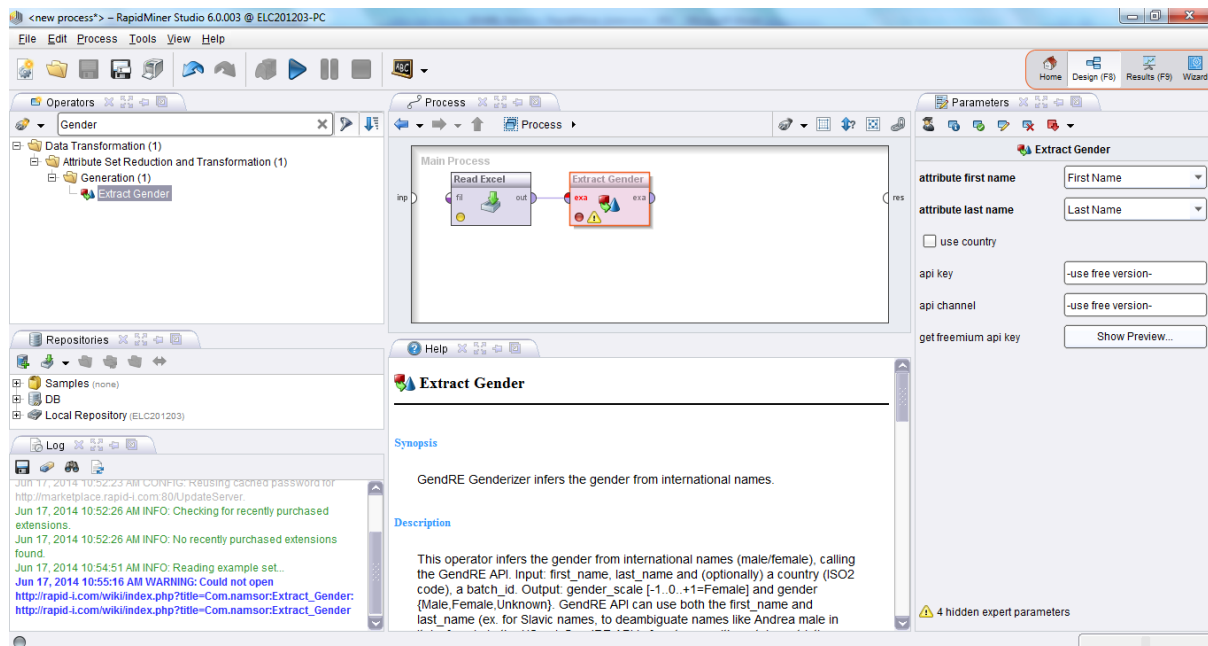
Leave the **api_key / api_channel** to use the free GendRE API.

Add a CSV exporter to view the results.
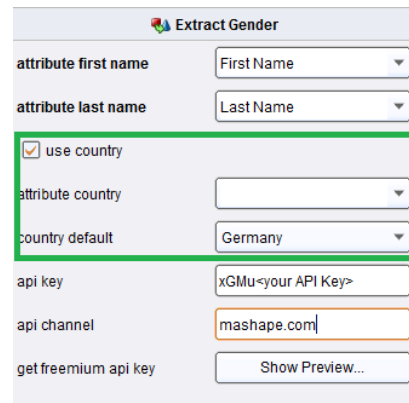
Additional attributes and parameters

Use Country

This information is inferred automatically in an upcoming API release, by recognizing the cultural origin of the (first_name, last_name) combination. Still, indicating the geography/locale improves the precision.

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.
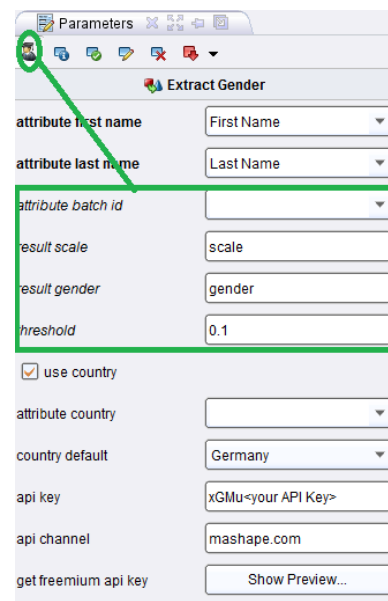
Expert Parameters

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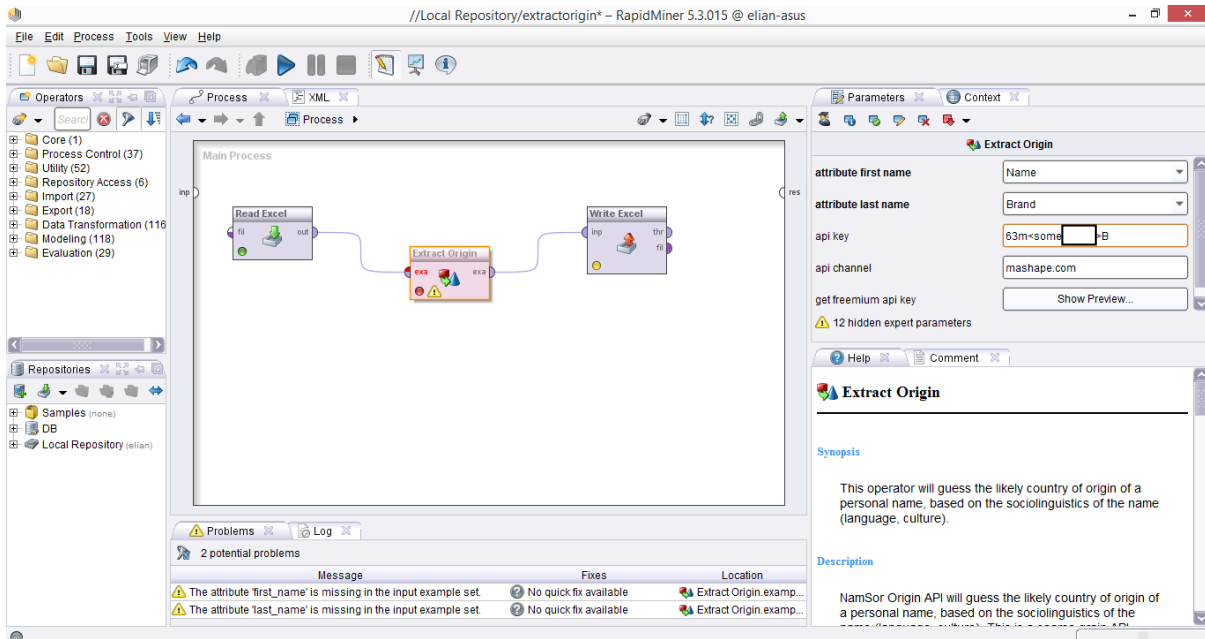
threshold: This parameter specifies threshold according to which Gender is considered 'Unknown' (evaluating *'Unknown' = abs(scale)<threshold*)

Create your first Extract Origin process

The steps are similar as for the Extract Gender process, only you need to obtain an API Key to activate the Freemium edition.

You can also cascade the different operators

Read CSV with Unstructured name -> Parse Name -> Extract Gender -> Extract Origin -> Write CSV

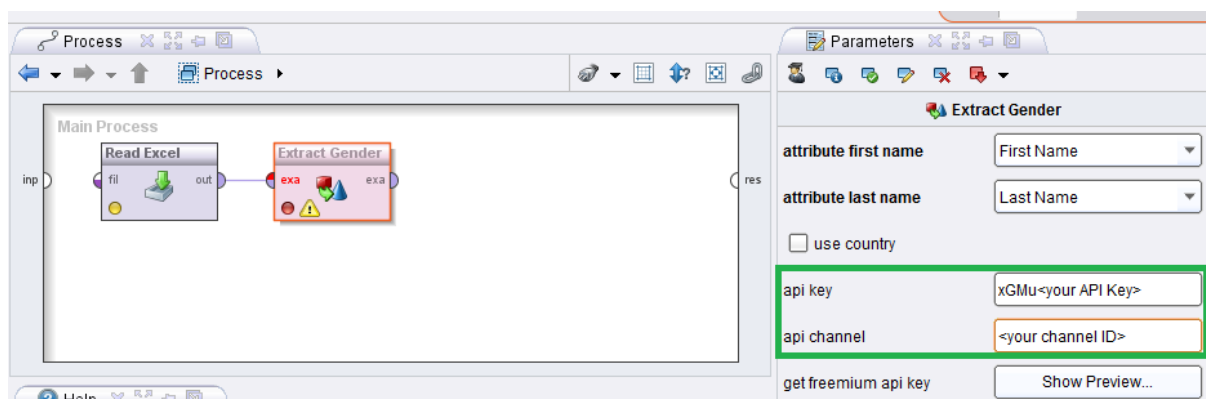# Ordering and getting support

Register on http://www.namsor.com/ to obtain an API Key.

NamSor API Key

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

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.namsor.com/onomastics/api/json/gender/John/Smith
```
{"scale":-0.99,"gender":"male"}
```

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

If NamSor extension for RapidMiner returns an error 500, it may be an issue with the proxy.
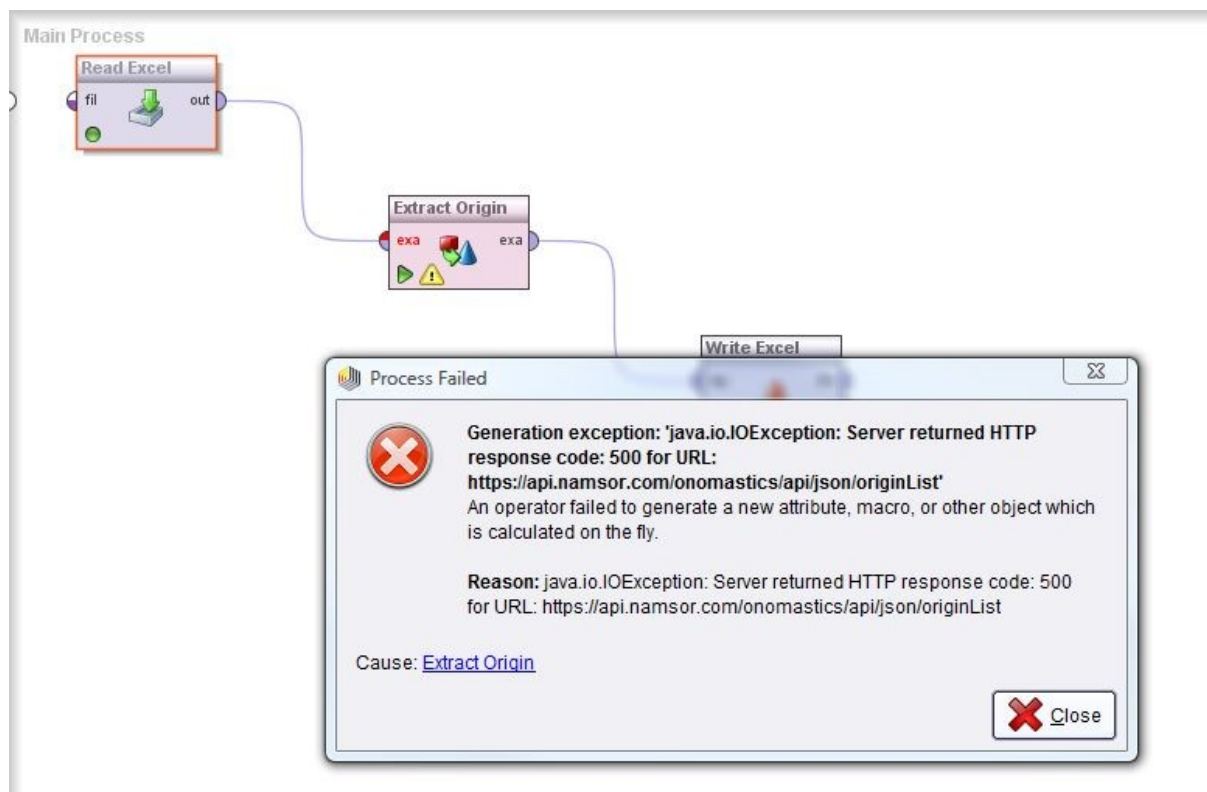
In the directory

C:\Program Files\RapidMiner\RapidMiner Studio

Make a copy of RapidMiner-Studio.bat and add **-Djava.net.useSystemProxies=true** to the launch command:
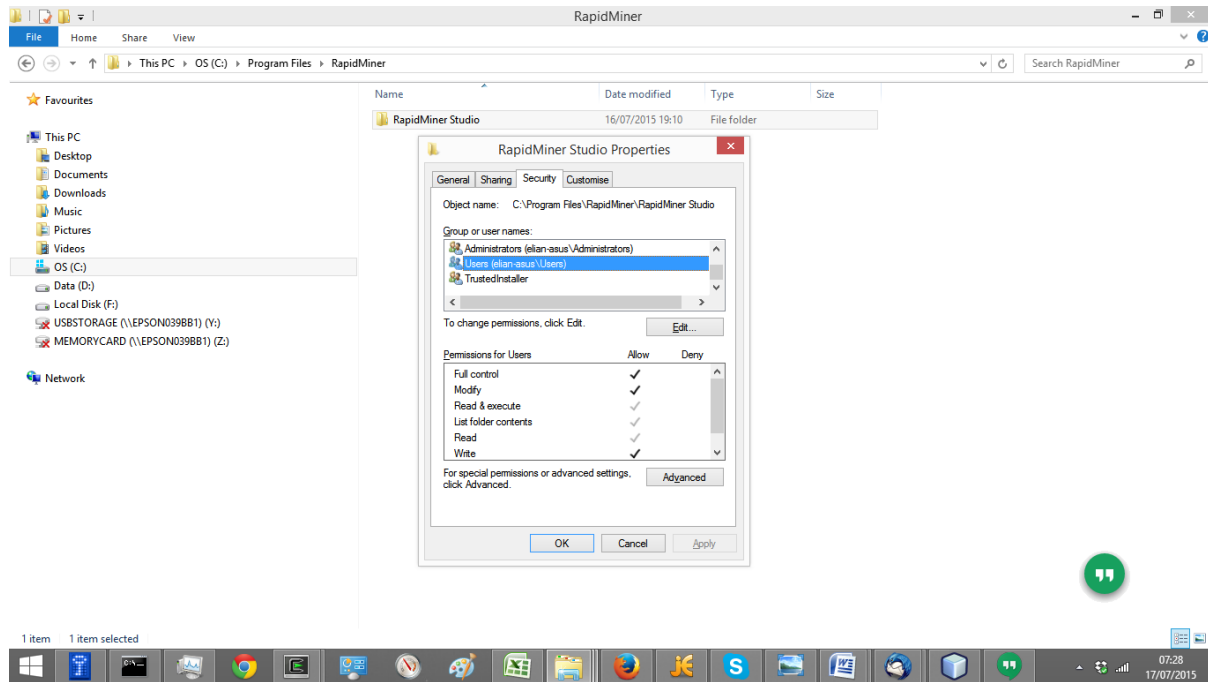
> echo Launching RapidMiner Studio GUI now...
>
> "%JAVA%" %JVM_OPTIONS% **-Djava.net.useSystemProxies=true** -cp *[...]*

Click on the modified file to launch RapidMiner, with the system proxy used as default.

NB/ You may need to change the permissions to write / edit

C:\Program Files\RapidMiner\RapidMiner Studio

(select and right click to edit permissions)

License & terms

Please review our licensing terms,

- the NamSor Onomastics Extension AGPL License,

https://raw.githubusercontent.com/namsor/rapidminer-onomastics-extension/master/LICENSE

- the NamSor API Terms & Privacy Policy,

https://namesorts.files.wordpress.com/2014/02/201507_namsor_api_terms_v006.pdf

- the RapidMiner MarketPlace Terms