NLC

A Watson Explorer Content Analytics UIMA analysis engine for Watson Natural Language Classifier.

Revision History

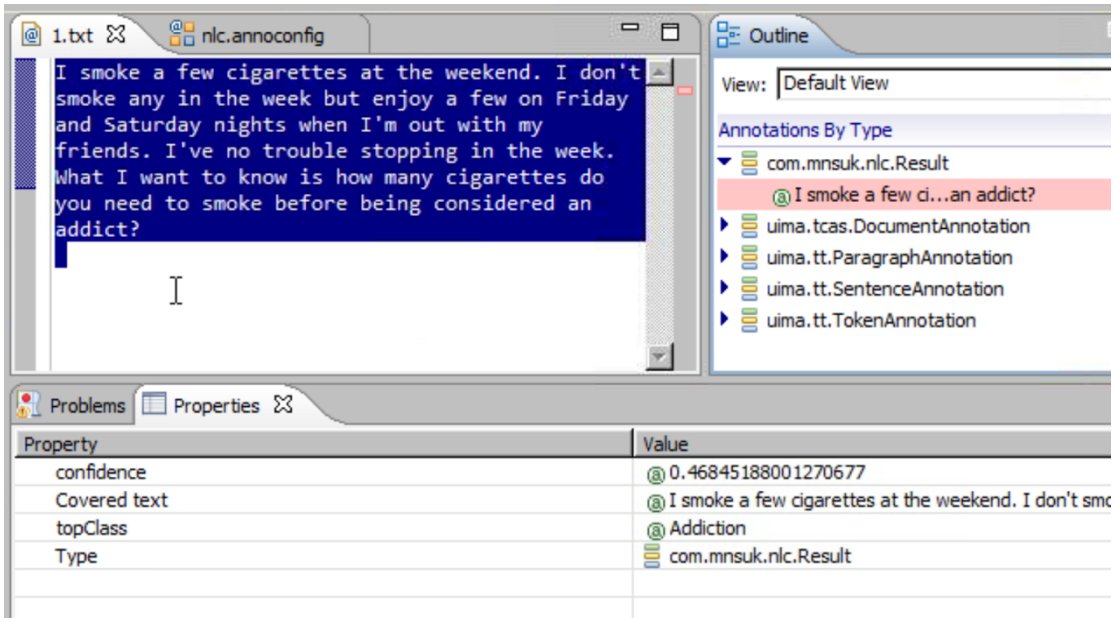
Date	Version	Status	Description	Author
07/04/2017	1.0	Release	Initial	Martin Saunders

Contents

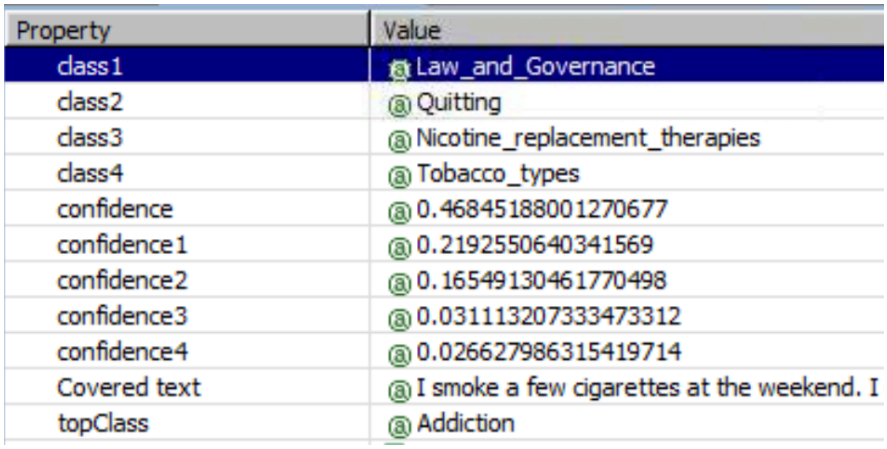
1. Introduction

The Watson Natural Language Classifier service uses uses machine learning algorithms to return the top matching predefined classes for short text input. You create and train a classifier to connect predefined classes to example texts so that the service can apply those classes to new inputs. This NLC annotator allows those identified classes to be annotated and utilised in any UIMA compliant pipeline such as Watson Explorer Content Analytics server or Content Analytics Studio.

For example, the screen shot below taken from Content Analytics Studio shows a short piece of text taken from a health forum that's been classified with a classifier for healthy choices. The results show that this text has been classed as being about "Addiction" with a confidence level of 46.8%.



The annotator allows just the top classification to be returned (as shown above) or for the top five (see below). Additionally a threshold can be set to only return results where the confidence exceeds that level.



For larger documents only the first 1024 characters of the text are sent for analysis to stay within the limits of the Watson NLC service.

1.1 Languages Supported

The annotator supports the same languages as the Watson NLC service. At the time of writing these are: Arabic, English, French, German, Italian, Japanese, Portuguese, and Spanish.

2. Installation

2.1 Pre-Requisites

- Installation of Content Analytics Studio version 11 or above
- NLC files:
  - Core files (included in distribution):
    - NLC-ae.xml
    - NLC-n-n-n.jar
    - ICAUIMAUtils-n-n-n.jar
  - Watson Java JDK
    - core-n.n.n.jar
    - natural-language-classifier-n.n.n.jar
    - okhtpp-n.n.n.jar
    - okhttp-urlconnection-n.n.n.jar
    - okio-n.n.n.jar
    - gson-n.n.n.jar
    - logging-interceptor-n.n.n.jar
- Credentials for an instance of the Watson NLC service deployed in Bluemix.
- The unique identifier for a trained classifier deployed in the NLC instance

2.2 Content Analytics Studio

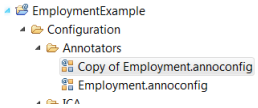
2.2.1 Create a project.

2.2.2 In this project create a folder to hold the NLU resources. This can be anywhere in the workspace but something like ***Resources/Custom/NLC*** would be a good choice.

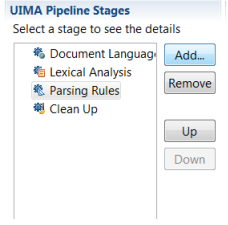
2.2.3 Copy the annotation engine configuration file (***NLC-ae.xml***) and all the jar files that make up NLU into this new folder.

2.2.4 In the relevant UIMA pipeline configuration file in your project add a custom stage as the penultimate stage of the pipeline. To do this:

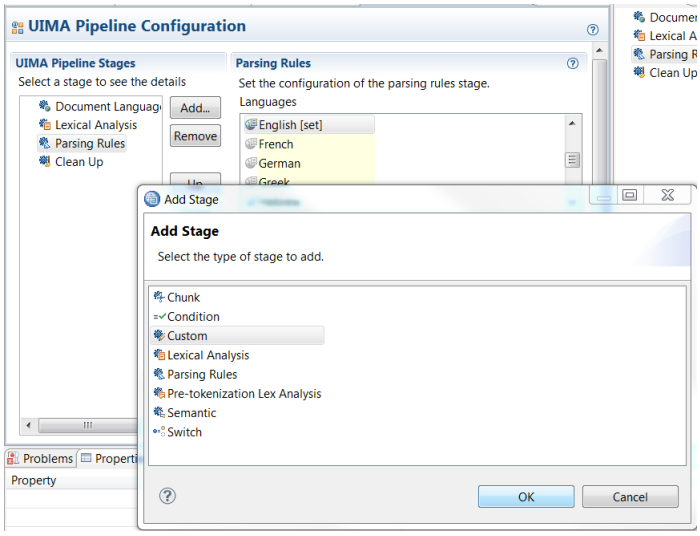
- Double click on the relevant pipeline configuration file under your ***Configuration/Annotators*** folder to open and edit it.



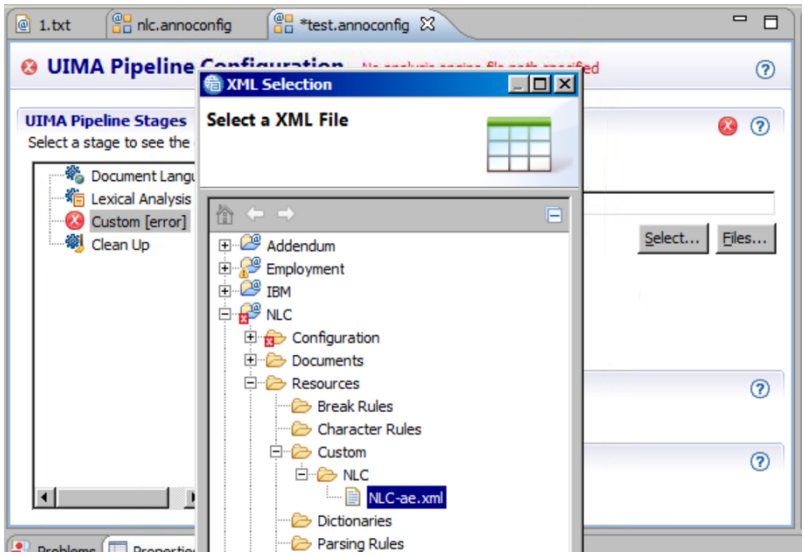
- select the Lexical Analysis stage
  click the “Add” button



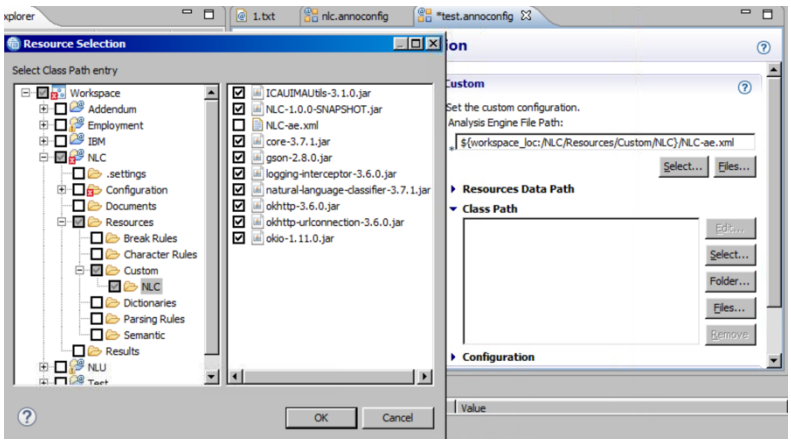
- select “Custom” in the popup window
  click “OK”



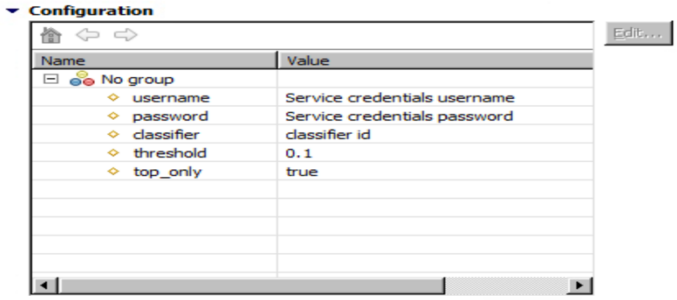
- Click in the text box of the Custom panel under where it says ***Analysis Engine File Path:*** and click the ***Select*** button. Navigate to the folder created in step 2.2.2 and select the ***NLC-ae.xml*** annotation engine configuration file.



- Back in the ***Custom*** panel click the small right arrow to open the ***Class Path*** panel and click ***Select***. Again navigate to the new folder and this time click the check boxes to select all the jar files.
  Click ***OK***



- Back in the ***Custom*** panel click the small right arrow to open the ***Configuration*** panel.



Set the username and password with the credentials from your Watson NLU service deployed in Bluemix.

Set the additional parameters according to the guidance in the table below.

Parameter	Default value	Note
classifier		The unique identifier of the trained classifier to use.
threshold	0.1	0.0 <= threshold < 1.0 Results for classes that have a confidence value below this threshold will not be returned.
top_only	true	true | false By default only the class with the highest confidence will be returned. Set to false to have the top 5 classes and their confidence levels returned (where they exceed the set threshold).

- Save the UIMA pipeline configuration file
- You're ready to go!