Apache UIMA AlchemyAPI Annotator Documentation

Authors: The Apache UIMA Development Community

Version 2.3.0

Copyright © 2008, 2010 The Apache Software Foundation

Incubation Notice and Disclaimer. Apache UIMA is an effort undergoing incubation at the Apache Software Foundation (ASF). Incubation is required of all newly accepted projects until a further review indicates that the infrastructure, communications, and decision making process have stabilized in a manner consistent with other successful ASF projects. While incubation status is not necessarily a reflection of the completeness or stability of the code, it does indicate that the project has yet to be fully endorsed by the ASF.

License and Disclaimer. The ASF licenses this documentation to you under the Apache License, Version 2.0 (the "License"); you may not use this documentation except in compliance with the License. You may obtain a copy of the License at

http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, this documentation and its contents are distributed under the License on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

Trademarks. All terms mentioned in the text that are known to be trademarks or service marks have been appropriately capitalized. Use of such terms in this book should not be regarded as affecting the validity of the the trademark or service mark.

Table of Contents

Introduction

The AlchemyAPI Annotator is a set of annotators that wrap the AlchemyAPI (http://www.alchemyapi.com) services provided by Orchestr8 (http://www.orchestr8.net).

To use AlchemyAPI Annotator, choose which service you want to put inside your UIMA pipeline, then find the corresponding AE descriptor, put your API Key as value of the 'apikey' parameter, eventually modify the parameters' defaults and you're done.

Chapter 1. Wrapped services

This chapter describes the AlchemyAPI services wrapped inside AlchemyAPI Annotator.

1.1. Categorization

AlchemyAPI's Categorization service can be used to categorize text, HTML, or web-based content, assigning the most likely topic category (news, sports, business, etc.)

Table 1.1. Category Types

Arts and Entertainment	Arts (Drawing, Sculpting, etc.) and Entertainment (Movies, Music, etc.) News and Discussion.
Business	Business and Finance News, SEC filings, etc.
Computers and Internet	Information Technology (Computers, Internet, Mobile Phones, etc.) News and Discussion.
Culture and Politics	Political News and Discussion, and Culture / Society-related issues.
Gaming	Gaming (Computer Games, Video Games, Board Games) News and Discussion.
Health	Health (Medications, Treatments, Diseases, etc.) News.
Law and Crime	Legal and Crime-related News and Discussion.
Religion	Religious News and Discussion.
Recreation	Recreational Activities (Boating, etc.)
Science and Technology	Science (Physics, Astronomy, etc.) News and Discussion.
Sports	Sports News and Commentary.
Weather	Weather News and Discussion.

Supported languages are: *English*, *French*, *German*, *Italian*, *Portuguese*, *Russian*, *Spanish*, *Swedish* .

The information extracted on the category of (for example) the passed text document is stored inside a FeatureStructure of type org.apache.uima.alchemy.ts.categorization.Category, with text and score.

1.2. Keyword Extraction

AlchemyAPI's Keyword Extraction service can be used to extract topic keywords from HTML, text, or web-based contents. Supported languages are: *English*, *French*, *German*, *Italian*, *Portuguese*, *Russian*, *Spanish*, *Swedish*

Each keyword is put inside a org.apache.uima.alchemy.ts.keywords.KeywordFS Feature Structure with the extracted keyword as text feature.

1.3. Language Detection

AlchemyAPI's Language Detection service can be used to extract language from a text, HTML, or web-based content. AlchemyAPI identifies more of 95 languages. Supported languages:

- Afrikaans ISO-639-3: afr

  Albanian ISO-639-3: sqi

  Amharic ISO-639-3: amh

  Amuzgo Guerrero ISO-639-3: amu

  Arabic ISO-639-3: ara

  Armenian ISO-639-3: hye

  Azerbaijani ISO-639-3: aze

  Basque ISO-639-3: eus

  Breton ISO-639-3: bre

  Bulgarian ISO-639-3: bul

  Catalan ISO-639-3: cat

  Cebuano ISO-639-3: ceb

  Central K'iche' ISO-639-3: qut

  Central Mam ISO-639-3: mvc

  Chamorro ISO-639-3: cha

  Cherokee ISO-639-3: chr

  Chinese ISO-639-3: zho

  Comaltepec Chinantec ISO-639-3: cco

Comaltepec Chinantec ISO-639-3: cco

Croatian ISO-639-3: hrv

Cubulco Achi' ISO-639-3: acc

Czech ISO-639-3: ces

Dakota ISO-639-3: dak

Danish ISO-639-3: dan

Dutch ISO-639-3: nld

English ISO-639-3: eng

Esperanto ISO-639-3: epo

Estonian ISO-639-3: est

Faroese ISO-639-3: fao

Fijian ISO-639-3: fij

Finnish ISO-639-3: fin

French ISO-639-3: fra

Fulfulde Adamawa ISO-639-3: fub

Georgian ISO-639-3: kat

German ISO-639-3: deu

Greek ISO-639-3: ell

Guerrero Nahuatl ISO-639-3: ngu

Gujarti ISO-639-3: guj

Haitian Creole ISO-639-3: hat

Hausa ISO-639-3: hau

Hawaiian ISO-639-3: haw

Hebrew ISO-639-3: heb

Hiligaynon ISO-639-3: hil

Hindi ISO-639-3: hin

Hungarian ISO-639-3: hun

Icelandic ISO-639-3: isl

Indonesian ISO-639-3: ind

Irish ISO-639-3: gle

Italian ISO-639-3: ita

Jacalteco ISO-639-3: jac

Japanese ISO-639-3: jpn

Kabyle ISO-639-3: kab

Kaqchikel ISO-639-3: cak

Kirghiz ISO-639-3: kir

Kisongye ISO-639-3: sop

Korean ISO-639-3: kor

Latin ISO-639-3: lat

Latvian ISO-639-3: lav

Lithuanian ISO-639-3: lit

Low Saxon ISO-639-3: nds

Macedonian ISO-639-3: mkd

Malay ISO-639-3: msa

Maltese ISO-639-3: mlt

Maori ISO-639-3: mri

Micmac ISO-639-3: mic

Mòoré ISO-639-3: mos

Ndebele ISO-639-3: nde

Nepali ISO-639-3: nep

Norwegian ISO-639-3: nor

Ojibwa ISO-639-3: oji

Pashto ISO-639-3: pus

Persian ISO-639-3: fas

Polish ISO-639-3: pol

Portuguese ISO-639-3: por

Q'eqchi' ISO-639-3: kek

Romanian ISO-639-3: ron

Romani ISO-639-3: rom

Russian ISO-639-3: rus

Serbian ISO-639-3: srp

Shona ISO-639-3: sna

Shuar ISO-639-3: jiv

Slovak ISO-639-3: slk

Slovenian ISO-639-3: slv

Spanish ISO-639-3: spa

Swahili ISO-639-3: swa

Swedish ISO-639-3: swe

Tagalog ISO-639-3: tgl

Thai ISO-639-3: tha

Todos Santos Cuchumatan Mám ISO-639-3: mvj

Turkish ISO-639-3: tur

Ukrainian ISO-639-3: ukr

Urdu ISO-639-3: urd

Uspanteco ISO-639-3: usp

Vietnamese ISO-639-3: vie

Welsh ISO-639-3: cym

Wolof ISO-639-3: wol

Xhosa ISO-639-3: xho

Zarma ISO-639-3: ssa

1.4. Ranked Entities Extraction

AlchemyAPI's Ranked Entities Extraction service can be exploited for identifying people, companies, organizations, cities, geographic features, and other typed entities within your HTML, text, or web-based content. To see the complete list of supported entity types see here: http://www.alchemyapi.com/api/entity/types.html[1] This service also expose 'entity disambiguation' mechanism to resolve detected companies, locations, and people to a unique "instance". Such thing is stored inside the 'disambiguation' feature of all the FeatureStructures inside the package org.apache.uima.alchemy.ts.entity, except for AlchemyAnnotation. This service also provides comprehensive support for RDF and Linked Data. This is done filling some features like 'dbpedia', 'geonames' and other linked data nodes' information. In the end also advanced quotations extraction is integrated into all named entity extraction API calls, enabling the identification of utterances ("quotations") within unstructured text, this will be stored inside 'quotations' feature.

1.5. Microformats Extraction

This AlchemyAPI's service handles Microformats data standards and is capable of extracting hCard, adr, geo, and rel-* formatted content from web pages. The information is stored inside a Feature Structure called MicroformatFS under org.apache.uima.alchemy.ts.microformats package.

1.6. Annotated Entities Extraction

This service is somehow an enhancement of the Ranked Entities Extraction service, with annotation of piece of text relating to extracted entities.

> **Note:** Annotator and descriptor still live but the service underneath is not available for free API keys.

[1] http://www.alchemyapi.com/api/entity/types.html

Chapter 2. Configuring parameters

This chapter describes how to configure AlchemyAPI Annotator analysis engines' parameters.

2.1. Common Parameters

- *apikey* : this parameter contains the (free or not) API key string needed to be able to call AlchemyAPI services.
- *output* : this parameter contains one of xml,json,rdf,rel-tag respectively to specify the desired output in XML, JSON, RDF or MicroFormat formats. Beware that each AE wrapping an AlchemyAPI service has a set of supported output formats that is a subset of the previous list.

2.2. Service Specific Parameters

Entity Extraction Text : see here[1]

Text Categorization : see here[2]

Text Keyword/Term Extraction : see here[3]

HTML Micrformats Extraction : see here[4]

URL Micrformats Extraction : see here[5]

[1] http://www.alchemyapi.com/api/entity/textc.html
[2] http://www.alchemyapi.com/api/categ/textc.html
[3] http://www.alchemyapi.com/api/keyword/textc.html
[4] http://www.alchemyapi.com/api/mformat/htmlc.html
[5] http://www.alchemyapi.com/api/mformat/urls.html