

# Module 3: GATE Cloud

Introduction to GATE Cloud

Ian Roberts ([i.roberts@sheffield.ac.uk](mailto:i.roberts@sheffield.ac.uk))

# This session will be recorded

Recorded video will be available after this session

# GATE Cloud Overview

- Online platform operated by the GATE team
- <https://cloud.gate.ac.uk>
- Two principal strands
  - Dedicated servers
  - Processing pipelines
    - Free rate limited REST API
    - PAYG batch processing
- Originally launched around 2010

## Dedicated servers - briefly

- Rent a virtual server pre-installed with our software
- Pay per hour of running time
- Two types of server available
  - Mimir indexing and search engine for GATE documents
  - Twitter data collector
- Covered in more detail this afternoon in Social Media module

# GATE Cloud Pipelines

- Recall: in GATE you build complex processing *pipelines* out of simple components
- GATE Cloud lets us publish “pipeline-as-a-service”
- 80+ pipelines available
- <https://cloud.gate.ac.uk/shopfront>

# GATE Cloud Pipelines

The screenshot shows a web browser window with the URL `cloud.gate.ac.uk/shopfront`. The page features a navigation bar with 'Home' and 'Services' buttons. Below the navigation, there is a filter section displaying 90 items with various category tags such as 'Abuse (3)', 'Archaeology (6)', 'Ariadne Infrastructure (6)', 'Basque (1)', 'Biomed (3)', 'Bulgarian (1)', 'Catalan (1)', 'Chunker (1)', 'Covid-19 (1)', 'Croatian (1)', 'Custom (3)', 'Czech (1)', 'Danish (1)', 'Dendrochronology (3)', 'Deprecated (1)', 'Dutch (4)', 'English (31)', 'Environment (2)', 'Estonian (1)', 'Finnish (1)', 'French (6)', 'German (7)', 'Greek (2)', 'Hyperpartisan (1)', 'Indonesian (1)', 'Italian (1)', 'Journalism (7)', 'Latvian (1)', 'Measurements (2)', 'Misinformation (1)', 'Morphology (1)', 'Named Entity (38)', 'News (1)', 'OCR (1)', 'OpenNLP (3)', 'Opinion Mining (4)', 'Part-of-Speech (23)', 'Polish (1)', 'Politics (3)', 'Portugese (1)', 'Romanian (2)', 'Russian (3)', 'Server (3)', 'Slovak (1)', 'Slovenian (1)', 'SoBigData (3)', 'spaCy (6)', 'Spanish (4)', 'Summarization (2)', 'Swedish (3)', 'Term Recognition (2)', 'Twitter (20)', 'Veracity (1)', 'Welsh (1)', and 'WeVerify (13)'. Three service cards are visible: 'English Named Entity Recognizer', 'TwitIE Named Entity Recognizer for Tweets', and 'Twitter Collector'. Each card includes a description, a price tag, and a request limit.

**English Named Entity Recognizer**  
Identify names of *persons, locations, organizations*, as well as *money amounts, time and date expressions* in English texts automatically.  
**1,200 free requests / day**  
Larger batches **£0.80 / CPU hour**

**TwitIE Named Entity Recognizer for Tweets**  
Named entity recognition service for Twitter data. Identifies *person, location, organization* etc. and also performs normalization of abbreviations and common shorthands ("brb", "gr8", "2day", etc.).  
**1,200 free requests / day**  
Larger batches **£0.80 / CPU hour**

**Twitter Collector**  
Collect tweets, view tweet statistics, and store results in your dashboard for further analysis.  
**£0.05 / CPU hour**

# Using a GATE Cloud Pipeline

- Two interfaces
  - “Online” API
    - Send one text, get back annotations immediately
    - Free but quota-controlled
  - Batch “annotation job”
    - Process large bundles of data in parallel
    - Unlimited, pay-as-you-go
- We will focus mainly on online API

### English Named Entity Recognizer



Identify names of *persons, locations, organizations*, as well as *money amounts, time and date expressions* in English texts automatically.

**1,200 free requests / day**  
Larger batches **£0.80 / CPU hour**

# Trying it out

- You can try out any pipeline direct from the user interface
- Enter text you want to process
- Results shown as colour highlights, much like the GATE Developer UI
- Try it yourself!

The screenshot shows a web browser window with the URL `cloud.gate.ac.uk/shopfront/displayItem/annie-named-entity-recognizer`. The page title is "English Named Entity Recognizer". In the top right corner, it displays "1,200 free requests / day" and "Larger batches £0.80 / CPU hour".

The main content area features a logo for ANNIE (a green circle with a white 'A') and the text: "ANNIE is a named entity recognition pipeline that identifies basic entity types, such as *Person, Location, Organization, Money amounts, Time and Date* expressions. It is the prototypical information extraction pipeline distributed with the GATE framework and forms the base of many more complex GATE-based IE applications." Below this is a blue button labeled "Annotation details".

To the right of the text is a preview of the GATE Developer UI, showing a document with various entities highlighted in different colors (green for Person, blue for Location, etc.).

Below the main text is a section titled "Test this pipeline". It contains a text input field with the placeholder "Type the content to annotate:". Below the input field are two options: "Or select a text file:" with a "Choose file" button and "No file chosen" text, and "Output type:" with a dropdown menu set to "JSON". Below that is "Document format:" with a dropdown menu set to "plain text". At the bottom of the section are two blue buttons: "Customize annotations" and "Test Pipeline".



English Named Entity Recognizer

cloud.gate.ac.uk/shopfront/displayItem/annie-named-entity-recognizer

## Test this pipeline

Type the content to annotate:

Ian works at The University of Sheffield.

Or select a text file:  No file chosen

Output type:

Document format:

[download](#)

Annotation types:  Organization  Person

### Annotations at this location

**Organization**

|           |                 |
|-----------|-----------------|
| orgType   | university      |
| rule      | GazOrganization |
| ruleFinal | OrgFinal        |

Use this pipeline

# Setting parameters

- One pipeline can produce several different annotations
- Each pipeline specifies its “default” and “additional” annotation types
- Click “annotation details”

## Annotation details

### Default annotations

|               |   |
|---------------|---|
| :Person       | Standard named entity types                                 |
| :Location     |   |
| :Organization |   |
| :Date         |   |
| :Address      | Includes email and IP addresses as well as street addresses |

### Additional annotations available if selected

|             |  |
|-------------|--|
| :Money      | Monetary amounts   |
| :Percent    | Expressions representing percentages                               |
| :Token      | The individual tokens of the text, with "category" feature for POS |
| :SpaceToken | The spaces between tokens  |
| :Sentence   | Sentences detected by the sentence splitter                        |

# Setting parameters

- “Customize annotations” to select the types you want

Address  Date  Location  Organization  Person  Money  Percent  Token  
 SpaceToken  Sentence

- GATE Cloud services can handle any document type that GATE can parse - select “document format”

Document format:

plain text ✓  
HTML  
XML  
Cochrane Library  
Pubmed  
MediaWiki  
Twitter JSON  
DataSift JSON

# Example - processing HTML

## Test this pipeline

Type the content to  
annotate:

```
<html>
<head>
<script language=JavaScript>
document.cookie='FTSection=hp/homepg;domain=.ft.com;path=/';document.cookie='FTPPage=0hparti;domain=.ft.co
m;path=/';document.cookie='FTReferrer='+escape(document.referrer)+';domain=.ft.com;path=/track/';document.coo
kie='FTURL='+escape(document.URL)+';domain=.ft.com;path=/track/';
</script>
```

Or select a text file:

airlines-27-jul-2001.html

Output type:

JSON

Document format:

HTML

Customize annotations

Test Pipeline

download

Annotation types:

Date

Location

Organization

Load an HTML file

Set the format

Text parsed out of  
the HTML

FT.com | TotalSearch | Global Archive | Print Return to Article | Print this Page Airlines take over running of air traffic control  
 FT.com site, Jul 27, 2001 BY KEVIN DONE, AEROSPACE CORRESPONDENT Seven UK airlines including British Airways,  
 Virgin Atlantic, BMI British Midland and EasyJet, on Friday took over control of the air traffic control system, completing  
one of the government's most controversial public-private partnership deals. Completion of the National Air Traffic Services deal  
comes at a critical time for the government as it tries to push through the PPP for the London Underground. The sale to a  
strategic investor of a 46 per cent stake in Nats is the first time in Europe that management control of en route air traffic

# Output formats

- Choose JSON or XML output format
- Both produce the same clickable view but the “download” link gives access to the raw API response
- XML is the GATE standoff XML format - for examples see the documents you used in module 1
- JSON is based on Twitter data format

# JSON example

- “entities” has a list of annotations of each type
- “indices”: [start, end] positions within the “text”, exactly\* as in GATE Developer

```
1 {  
2   "text": "Ian works at The University of Sheffield",  
3   "entities": {  
4     "Organization": [  
5       {  
6         "indices": [17, 40],  
7         "orgType": "university",  
8         "rule": "GazOrganization",  
9         "ruleFinal": "OrgFinal"  
10      }  
11    ],  
12    "Person": [  
13      {  
14        "indices": [0, 3],  
15        "gender": "male",  
16        "kind": "firstName",  
17        "rule": "GazPersonFirst",  
18        "firstName": "Ian",  
19        "ruleFinal": "PersonFinal"  
20      }  
21    ]  
22  }  
23 }
```

\* not quite exactly, if text has emoji

# Using the API

- Try out interface is simply one front end on top of API
- You can call the same API directly from your own code, in any language with an HTTP client
- We have developed clients for
  - GATE Developer
  - Google Sheets

# Using the API

- Each pipeline has its own endpoint URL
- HTTP POST to that URL with the document as body (with appropriate value for `Content-Type` header)
- JSON response by default, `Accept: application/xml` to get XML instead
- [Full docs on cloud.gate.ac.uk](https://cloud.gate.ac.uk)

## Use this pipeline

### Single documents

---

You can process up to **1,200** documents per day free of charge using the **REST API**, at an average rate of **2 documents/sec**. Higher quotas are available for research users by arrangement, [contact us](#) for details.

The API endpoint for this pipeline is:

```
https://cloud-api.gate.ac.uk/process-document/annie-named-entity-recognizer
```

Create API Key



# Authentication and rate limits

- Public pipelines work without authentication, but at very low daily quota and rate limit
  - 1 call per 10 seconds on average
  - ~100 calls per day
- The advertised 1,200 per day, 2 calls per second is the baseline for *registered* users
- To get this quota you must authenticate using an *API Key*

# GATE Cloud API Keys

- If you haven't already, [register now](#) for an account on GATE Cloud
  - If you have already registered, [log in](#)
- Go to “<your name> account” in the top right
- Scroll to the bottom and “Manage your API keys”
- Generate a new key

Ian Roberts's account | Log out | Help

## API Keys

[Manage your API keys](#) for the GATE Cloud REST APIs.

# GATE Cloud API Keys

- API key is in two parts, the “key ID” and the “password”
- Copy and paste both into a text editor for future reference
- Note the warning:  
if you lose the  
password you must  
generate a new key  
(changing both ID  
and password)

## New API Key

These are the details of your newly generated API key. Make sure you have made a note of the password before leaving this page - **API key passwords cannot be recovered** and you will need to generate a replacement key if the password is lost.

Key ID **gcj7nx3h69lq**

Password **si6fegs5trxs8g0ui1bg**

Description

[Save changes](#)

[Return to the key management page](#)

# Python example

- Open [this notebook \(Google Colab\)](#)
- Save a copy in your own Google Drive
- Example of how to call GATE Cloud with the `requests` module

# Calling GATE Cloud in Developer

- We provide a GATE PR to call a GATE Cloud pipeline
- Start GATE Developer
- Load the “GATE Cloud Client” plugin
- Create a new “GATE Cloud API Client” processing resource

# Calling GATE Cloud in Developer

- Parameters
  - **apiKey** - the key ID you just created
  - **apiPassword** - the matching password
  - **endpointUrl** - the URL copied from “use this pipeline”

Parameters for the new GATE Cloud API Client

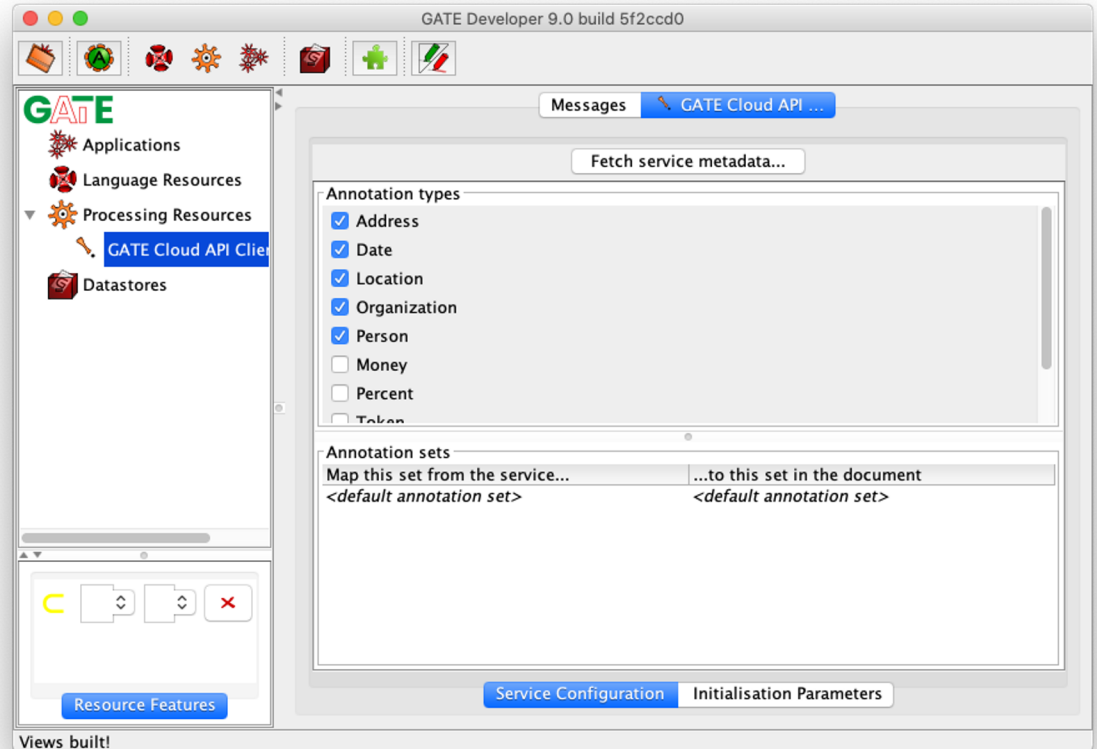
Name: GATE Cloud API Client 00008

| Name        | Type   | Required | Value                |
|-------------|--------|----------|----------------------|
| apiKey      | String |          | <input type="text"/> |
| apiPassword | String |          | <input type="text"/> |
| endpointUrl | URL    | ✓        | <input type="text"/> |

OK Cancel Help

# Configuring the PR

- Double-click the new PR in the tree, then “fetch service metadata”
- Annotation type checkboxes as on “test this pipeline”
- Annotation sets...



# Annotation sets and types

- Recall: in GATE each document has one or more annotation sets, each containing annotations of one or more types
- Think back to those “annotation details”
- See the colon - “:”
- Annotations are specified to GATE Cloud as *selector* expressions “SetName:AnnotationType”
- Most pipelines use the default set (which has no name)

## **Default annotations**

:Person  
:Location  
:Organization



# Mapping annotation sets

- By default, cloud client PR puts annotations that come back from service into the same set the service used
- But this is configurable
  - e.g. call ANNIE, but put its output into a set called “cloud”

| Annotation sets                  |                                |
|----------------------------------|--------------------------------|
| Map this set from the service... | ...to this set in the document |
| <default annotation set>         | cloud                          |

# Try it

- Configure your GATE Cloud Client PR to map ANNIE default set to local “cloud” set
- Load a document
- Create a corpus (shortcut - you can right-click on the document, then “new corpus with this document”)
- Load the default ANNIE application
- Add the GATE Cloud Client PR to the end of the ANNIE pipeline
- Run the app over your (one document) corpus

# Try it

Cloud ANNIE has  
found the same  
annotations as  
local ANNIE

The screenshot shows the GATE Developer 9.0 interface. The main window displays the text "Ian works at The University of Sheffield" with annotations. The left sidebar shows the GATE Applications tree, including ANNIE. The right sidebar shows the Annotation Sets list, with Organization and Person selected. The bottom status bar indicates "ANNIE run in 0.63 seconds".

| Type         | Set   | Start | End | Id  | Features                   |
|--------------|-------|-------|-----|-----|----------------------------|
| Person       |       | 0     | 3   | 114 | {firstName=Ian, gender=... |
| Person       | cloud | 0     | 3   | 143 | {firstName=Ian, gender=... |
| Organization |       | 17    | 40  | 115 | {orgType=university, ...}  |
| Organization | cloud | 17    | 40  | 144 | {orgType=university, ...}  |

4 Annotations (0 selected) Select: [ ] [New]

# Try a different service

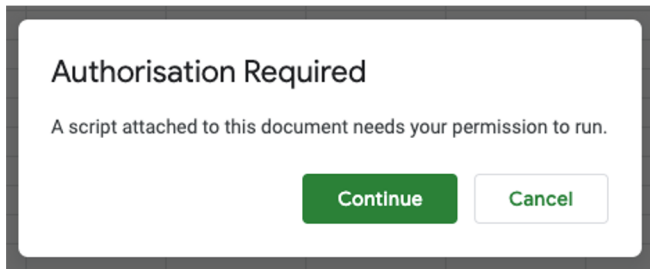
- ANNIE is a good test, but not one you'd really use in this way
- There are other pipelines on GATE Cloud that are not available for you to run locally
- Delete your Cloud Client PR
- Create another one calling a different service
  - e.g. [YODIE](#)
- YODIE uses the “Shef” annotation set by default

# Google Sheets

- We have recently developed an add-on for Google Sheets to process text in a spreadsheet
- Particularly useful for people who are comfortable with spreadsheets but not programming
- Currently in beta - not yet a published add-on you can install to your own sheets
- For testing purposes take a copy of [this Google Sheet](#)
  - You can only *view* the original, not edit it
  - go to File → “Make a copy”

# Authorisation

- In your copy of the sheet, “Add-ons” → “GATE Cloud Text Analysis” → “Open sidebar”
- First time you do this it will request authorisation
  - Your email will appear as the developer, this is normal!



## Google hasn't verified this app

The app is requesting access to sensitive info in your Google Account. Until the developer ([you@gmail.com](mailto:you@gmail.com)) verifies this app with Google, you shouldn't use it.



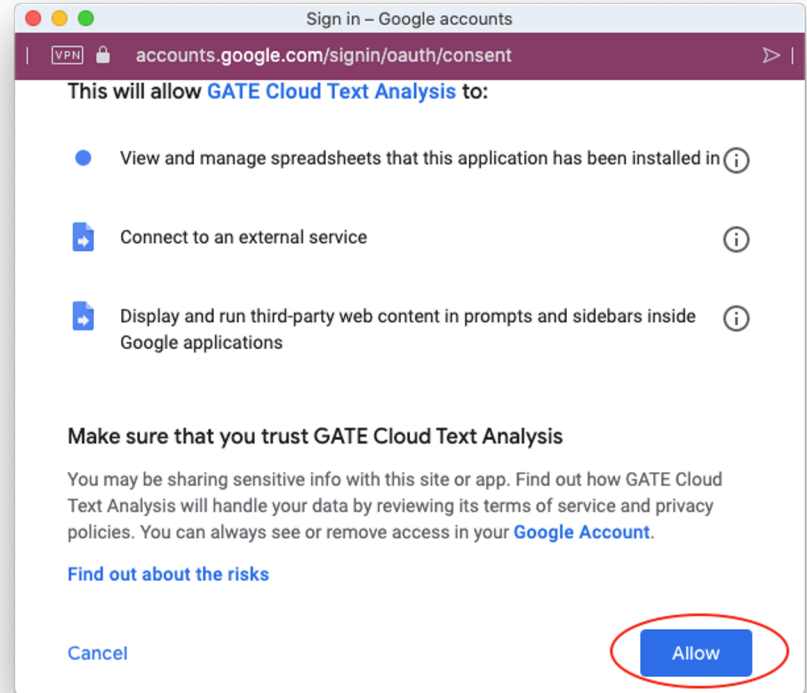
BACK TO SAFETY

# Authorisation

Continue only if you understand the risks and trust the developer  
([you@gmail.com](#)).

[Go to GATE Cloud Text Analysis \(unsafe\)](#)

- When the add-on is published properly then we will be the developer and it won't be considered "unsafe"



# Using the add-on

- Once authorised, “open sidebar” again
- “Configure credentials” lets you supply your API key ID and password
- Intended workflow
  - Choose a column of texts to process
  - Typically, set destination immediately to the right of the top-most input cell
    - For input B2:B9, destination would be C2
  - Choose the service to use
  - Configure annotations you want to extract using column headings

## Configuration

Configure credentials

## Analyse Text

Range to analyse

Use selected range

Specify a rectangular range of cells containing the text to analyse

Destination

Use current cell

Leading cell of the range where results should be placed

Process with service

ANNIE Named Entity Recognizer (English) ⇅

### Description

Finds named entities including person names, locations and organizations in English text.

[Full details \(opens in new window\)...](#)



# Demo

- The “BBC example” sheet has a few BBC news articles
- Select the cells with input text (B2 to B9)
- Under “Range to analyse” click “use selected range”
- We will put the output alongside - click into cell C2, then under destination click “use current cell”
  - Results from processing B2 will go in C2, D2, ...
  - B3 will go in C3, D3, ...
  - Etc.
- Select ANNIE as the service to use

# Configuring annotations

- One service can produce many different annotations
- Approach taken here is to map output into several columns
- “Configuration range” (typically column headers) defines what to extract
- Scroll sidebar down, click “Open configuration helper”

Configuration range

Use selected range

Open configuration helper

Specify a rectangular range of cells (typically a set of row or column headers) specifying how to handle the results from the service.

# Configuring annotations

- Click into cell C1, fill a horizontal range of 6 cells

To help you get started, the buttons below will let you generate an initial configuration range for this service, starting from the currently selected cell. Choose the cell you want to start from, then:

Fill a  range of  cells

- You will see C1 to H1 filled with headings, and the range selected - click “Use selected range”
  - Column headings are annotation types

# Explanation

- To summarise: you've now configured the tool to
  - process the text in B2:B9
  - using ANNIE
  - extract the annotations specified by C1:H1
  - and put the results in cells starting at C2
- “Submit job”
  - The tool will post each input cell in turn to GATE Cloud
  - If the input set were bigger, you'd see a percentage progress bar
  - Tool is careful to respect rate limit & quota, and will slow down processing or wait if necessary for quota to reset

# More advanced configuration

- For a column heading like “Person”, tool outputs the text under each Person annotation in the API response
- But we know GATE annotations have other features
- Configuration is actually a pattern language that can extract any combination of features for each annotation

# More advanced configuration

AnnotationType pattern

- `pattern` can be any sequence of feature names separated by spaces or punctuation

Person gender → male

Location text (`locType`) → Sheffield (city)

- `text` is a “magic” feature name representing the text under the annotation (like `@string` in JAPE)

# More advanced configuration

- Configuration helper creates drop-down options on the headers but these are not exhaustive
- You can edit the headings to fit your requirements
- Use “test this pipeline” to see what features are available
  - “Full details” link under service description will take you there

## Another example - YODIE

- Let's run a different pipeline on the same texts
- Range to analyse is still B2 : B9
- Service: "YODIE named entity disambiguation (English)"
- Destination: use cell I2
- Configuration helper: click in I1 and generate a one-cell range, then "use selected range"
- Drop down I1 to see example configs
  - All the same `Mention` annotation type, but different combinations of features - `inst` is the DBpedia instance URI



## Another example - YODIE

- Submit the job and let it process
- Note “Myanmar” and “Burma” both map to the same concept in DBPedia
  - It’s not perfect - “the WHO” is treated as the band, not the World Health Organization

# Batch processing overview

- If you want to process large numbers of documents in a hurry
  - Upload documents as ZIP files
  - or collect data from Twitter (see module 4)
  - Reserve an “annotation job”, either for a standard pipeline, or upload your own
  - Processing done on Amazon EC2 with many VMs in parallel
  - Output as JSON or (ZIP files of) XML like the online API
  - Pay for the compute time you use - buy credit vouchers from the University of Sheffield online shop

# Summary

- GATE Cloud offers a REST API to process text with any of our published pipelines
- API can be called by any client that speaks HTTP(S)
  - “Test this pipeline”
  - PR from GATE Developer
  - Your own code in Python/Java/golang/etc.
  - Google Sheets
- Everyone gets a basic free quota, enhanced quotas available for research users on request
- PAYG batch mode for larger jobs

# Links

- Home page: <https://cloud.gate.ac.uk>
- API spec: <https://cloud.gate.ac.uk/info/help/online-api.html>
- Java library: <https://github.com/GateNLP/cloud-client>

**Thank you!**