IDP alpha tutorial

Max Talanov

May 7, 2008

1 Outline

Contents

1 Outline **1**

2 Overview **2**

3 Start up **2**

4 Client GUI **3**

4.1 GUI areas 3
4.2 Connect to the server 4
4.3 Upload files 4
4.4 Train 4
4.5 Apply with normalization 5
4.6 View annotated texts 5
4.6.1 Normalization 6
4.6.2 Annotated text colors 7
4.7 Change annotation and retrain 7
4.8 New element in Annotation tree 7

5 FAQ **8**

2 Overview

IDP stands for intellectual document processing. IDP is capable of creating XML structures on base of simple texts and machine learning technique, in other words IDP is a project of machine learned structurisation and normalization of the unstructured text of natural language.
Actually this means that IDP is capable to split ordinary text on logical spans based on learned information during training by human process. Also it is possible to setup normalization rules to transform contents of some spans to place it in DB or valid XML.
IDP uses: minorthird for natural language processing, JGAP for normalization of English and Russian dates, VISTAICO AERO PACK for icon set.

3 Start up

This is short description of setup and installation process.

1. You have to have Java version 6, please download from Sun microsystems site

2. Download client and server archives from IDP download page (You must have done this)

3. Unzip client and server archives to separate directories ex.: *server* and *client* respectively

4. Start server: in *server* directory: under Windoz: run *startServer.bat*, under Linux: run *chmod +x startServer.sh* then *./startServer.sh*. Though you can make some setup in properties file, I would recommend to stay on default settings at the moment.

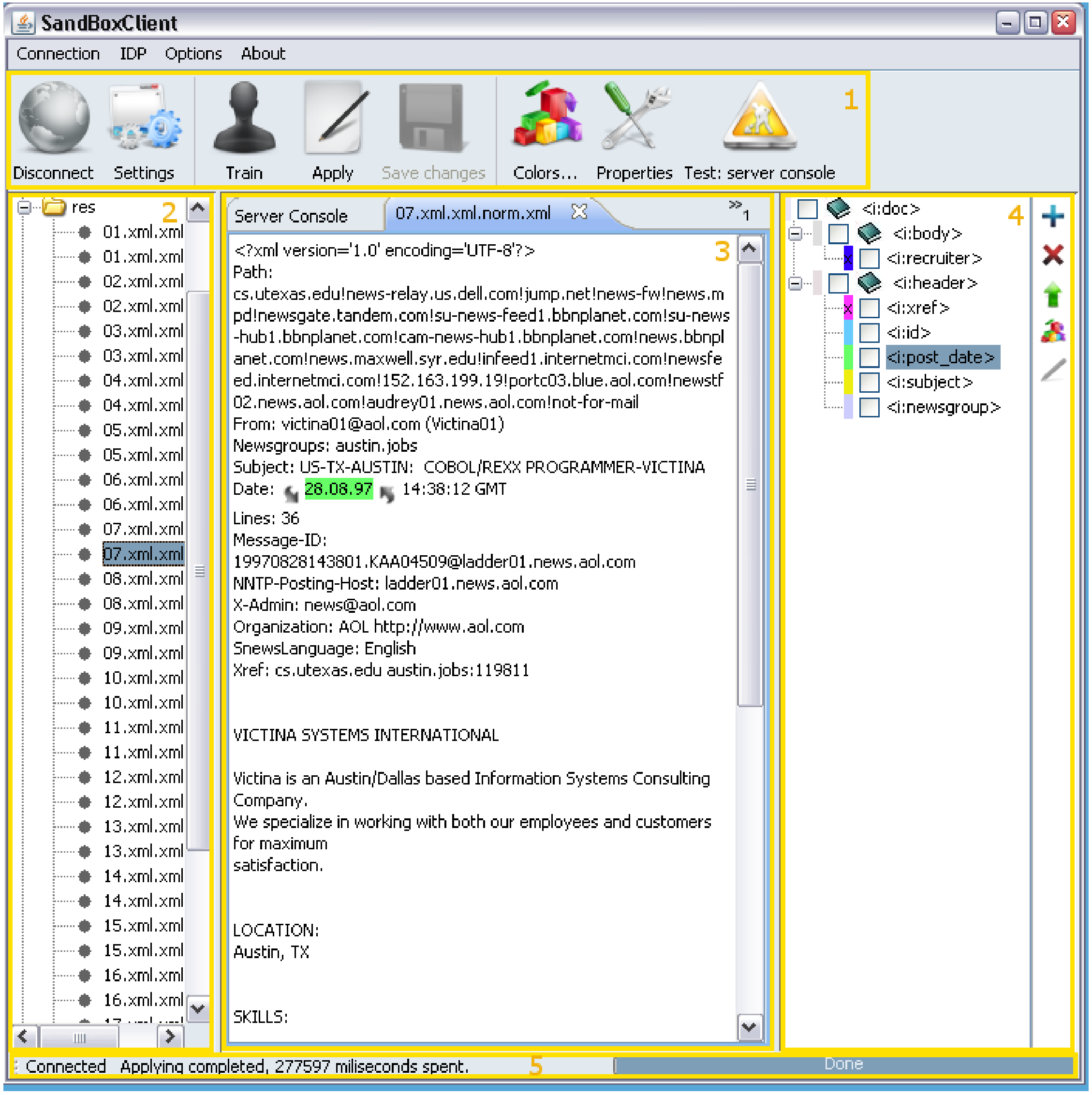
5. Start client: in *client* directory: under Windoz: run *startClient.bat*, under Linux: run *chmod +x startClient.sh* then *./startClient.sh*

This is actually all what you need to do to start using IDP.
This is possible that there could be some errors in server or in client parts, please be patient and submit bug reports, if you find some.

4 Client GUI

4.1 GUI areas

Client GUI has several sections or areas shown on the picture below.



```
<?xml version='1.0' encoding='UTF-8'?>
Path:
cs.utexas.edu!news-relay.us.dell.com!jump.net!news-fw!news.m
pd!newsgate.tandem.com!su-news-feed1.bbnplanet.com!su-news
-hub1.bbnplanet.com!cam-news-hub1.bbnplanet.com!news.bbnpl
anet.com!news.maxwell.syr.edu!infeed1.internetmci.com!newsfe
ed.internetmci.com!152.163.199.19!portc03.blue.aol.com!newstf
02.news.aol.com!audrey01.news.aol.com!not-for-mail
From: victina01@aol.com (Victina01)
Newsgroups: austin.jobs
Subject: US-TX-AUSTIN:  COBOL/REXX PROGRAMMER-VICTINA
Date:  28.08.97  14:38:12 GMT
Lines: 36
Message-ID:
19970828143801.KAA04509@ladder01.news.aol.com
NNTP-Posting-Host: ladder01.news.aol.com
X-Admin: news@aol.com
Organization: AOL http://www.aol.com
SnewsLanguage: English
Xref: cs.utexas.edu austin.jobs:119811


VICTINA SYSTEMS INTERNATIONAL

Victina is an Austin/Dallas based Information Systems Consulting
Company.
We specialize in working with both our employees and customers
for maximum
satisfaction.


LOCATION:
Austin, TX


SKILLS:
```

```
Connected  Applying completed, 277597 miliseconds spent.
```

1. Toolbar - mainly the server commands and settings.

2. Project tree - directory structure of the project.

3. Work panel - area for manual annotation and server console.
4. Annotation tree - XML tree of annotated text.
5. Status bar - Shows current state and progress of the commands.

4.2 Connect to the server

When you start client you are able to do no so much: Setup server settings, connect to the server. Note status bar is indicating *Not connected* state.

First of all, please connect to the server. Please press connect button Connect. Connect button should change to Disconnect, and status bar should indicate *Connected*.
In project tree, you should see *example_1* project name. Please expand *example_1*, double clacking on it. You should see *apply* and *train* directories.

4.3 Upload files

Please right click on *train* directory and choose *Upload...* item of menu, then choose three files from directory, where you have unpacked client archive, from *example_1/train* directory.
If files uploaded successfully, you should see *File has been uploaded successfully!* message. Please upload all *train* directory contents, this will more or less guarantee, correct training. Then upload 3-5 files to *apply* directory from *examle_1/apply* directory of client archive, for example `01.txt`, `02.txt`, `03.txt`. You should be able to see all uploaded files in directory tree, please do not forget to expand directories *apply* and *train*.

4.4 Train

Train is done mainly automatically by the server. To train please click on Train button Train. Work panel should display a lot of messages and progress bar in lower left corner should display name of training stage and percent of completeness of training. When training is complete you should be able to see *Training has been completed successfully!* message, progress bar should display *Done* message.

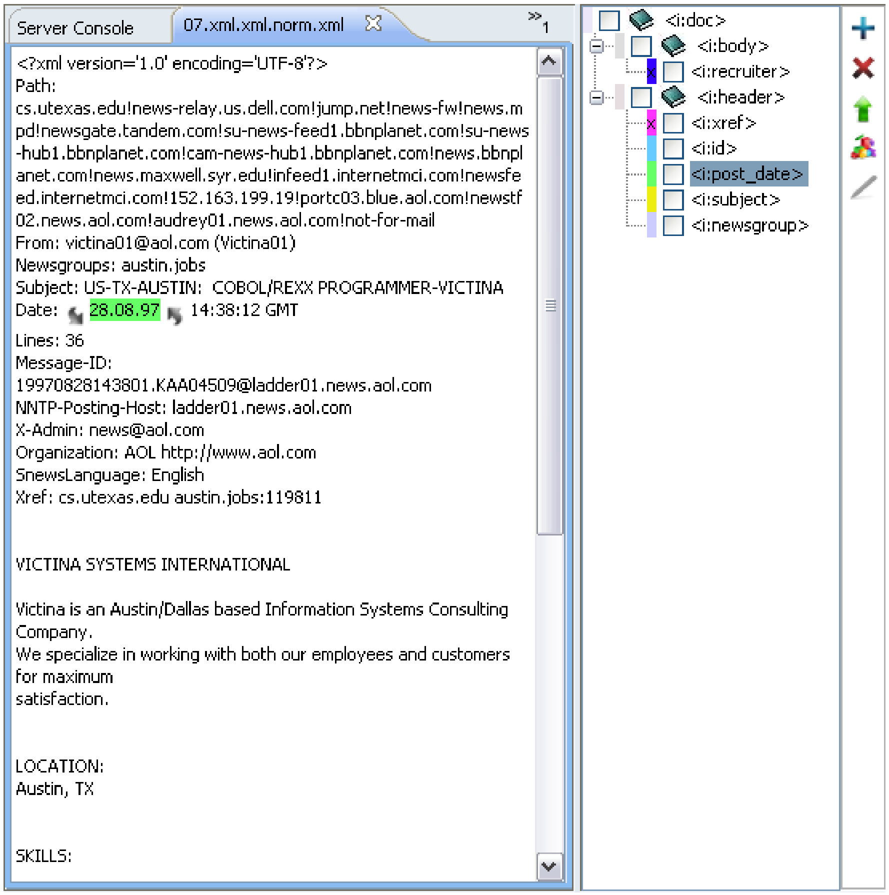
4.5 Apply with normalization

This is mainly most complex operation for the server, and requires some significant time.

Please click on Apply.
First you should see a lot of messages in the server console and progress bar should indicate *Loading annotators*, *Annotating texts*, *Normalising result xml files* stages. Please note that *Normalization* should take time, because of Genetic Algorithms invoked in the process.
After all you should see *Annotation has been completed successfully!* message and *res* directory should emerge in project tree.

4.6 View annotated texts

Expand *res* directory, double click on some file in *res* directory, that ends with *txt.xml*, but without *norm.xml* at the end, ex.: `01.txt.xml`. Work panel should display contents of xml file without xml tags. Annotation tree panel should display root element of xml tree result of annotation *i:doc*.

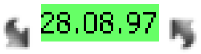


```
<?xml version='1.0' encoding='UTF-8'?>
Path:
cs.utexas.edu!news-relay.us.dell.com!jump.net!news-fw!news.m
pd!newsgate.tandem.com!su-news-feed1.bbnplanet.com!su-news
-hub1.bbnplanet.com!cam-news-hub1.bbnplanet.com!news.bbnpl
anet.com!news.maxwell.syr.edu!infeed1.internetmci.com!newsfe
ed.internetmci.com!152.163.199.19!portc03.blue.aol.com!newstf
02.news.aol.com!audrey01.news.aol.com!not-for-mail
From: victina01@aol.com (Victina01)
Newsgroups: austin.jobs
Subject: US-TX-AUSTIN:  COBOL/REXX PROGRAMMER-VICTINA
Date:  28.08.97  14:38:12 GMT
Lines: 36
Message-ID:
19970828143801.KAA04509@ladder01.news.aol.com
NNTP-Posting-Host: ladder01.news.aol.com
X-Admin: news@aol.com
Organization: AOL http://www.aol.com
SnewsLanguage: English
Xref: cs.utexas.edu austin.jobs:119811


VICTINA SYSTEMS INTERNATIONAL

Victina is an Austin/Dallas based Information Systems Consulting
Company.
We specialize in working with both our employees and customers
for maximum
satisfaction.


LOCATION:
Austin, TX


SKILLS:
```

4.6.1 Normalization

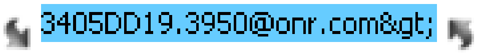
Please double click on *i:doc*, you should see tree drop down *i:body* and *i:header*, double click on *i:header*, then select *i:post_date*. Please double click on another file that ends with *.norm.xml* but with same name to previously opened, ex.: `01.txt.xml.norm.xml`. Select *i:post_date*, you should be able to notice difference of representation of the date, for example in file *01.txt.xml i:post_date* is `23 Aug 97` and in *01.txt.xml.norm.xml* `23.08.97`. Actually this is result of normalization.

4.6.2 Annotated text colors

At the moment you should see curly arrows around *post_date* and value should be selected with color  28.08.97 .

Please use toolbar button *Change tag color ...* to setup color of selected tag.

4.7 Change annotation and retrain

Please open three *txt.xml* files from *res* directory. Select *i:id* tag in *Annotation tree* and expand annotation to include `&gt;` or `&lt;`. Please don't forget to press *Save changes* button every time you change annotation. *corrected* directory with edited files should emerge in project tree. Please run *Train* again, as described above. Note that at the end of *Train* stage, you should be able to see corrected files in *train* directory. Add not processed txt files from *examle_1/apply* directory of client archive, for example `18.txt`, `19.txt`, `20.txt`. Run *Apply*. After all you should be able to see new files in *res* directory. Please click on new files, you should see annotation of *i:id* includes `&gt;`  3405DD19.3950@onr.com&gt; .

4.8 New element in Annotation tree

Please close all opened tabs in work panel, delete all files from *apply* directory, upload *17.txt* file from client archive to *apply* directory. Expand *res* directory, open `18.txt.xml`, `19.txt.xml`, `20.txt.xml`, expand *i:doc*, *i:header* elements in *Annotation tree*. Select *i:header*, press ➕ button, set *org* name of the tag. Do not forget to save changes, click ⬆ in *Annotation tree* panel. Expand tree back.

Select *i:org* element of the tree. Press ✏ tool of *Annotation tree* toolbar. Answer *Ok* in pop-up dialog. Select text to the right of *Organization:*,

for example in *20.txt.xml* it is *New Resources Corporation*. Press *Yes* in choose dialog to confirm your selection. Click on Save changes button to save changes of each file. Run *Train*, delete all files from *res* directory, run Apply. Expand *res* directory, open result file *17.txt.xml*, note *Annotation tree* contains *i:org* element and it is annotated.
Close client.
To stop server please run in *server* directory: under Windoz: *stopServer.bat*, under Linux:*chmod +x stopServer.sh* then *stopServer.sh*.

5 FAQ

Why is it client-server - Because Server really needs fast machine with 1Gb of free memory, especially in case of other than Enlish languages.

Can I use it for not English - Yes. IDP has been tested with Russian CV documents. We can not guarantee that it will really work on your documents (well it is still alpha).

What about integration with different systems - Not yet. IDP produces XML files, and that's it at the moment.

Can I report a bug - Yes please, use IDP issues list.

Why is it alpha - Because it has not really been tested on different datasets. We are sure, there are a lot of bugs. Please be patient.

Does your server really support multiple clients - Not at the moment.

Can I use different to text and xml input files - Not at the moment.

How many documents are ok to train - Depends, train set should describe context of text to be annotated as completely as possible.