CHEMDNER-patents CEMP subtask test text data (Version 14th July 2015)

For additional questions please send e-mail to: Martin Krallinger
(mkrallinger@cnio.es)

This directory contains the test set text for the CHEMDNER-patents CEMP, GPRO and CPD subtasks.

1. chemdner_patents_test_background_text_release.txt

This file contains plain-text, UTF8-encoded Patent abstracts in a
tab-separated format with the following three columns:

1- Patent identifier
2- Title of the patent
3- Abstract of the patent

It consists of 40,000 records that contain the subset of 7000 records that will be used as test set.

The other 33,000 records have been added as a background set to make sure that the systems are totally automatic and that no manual corrections of the test set predictions can be done.

2. General info related to test set prediction format

Before you submit your test set predictions make sure you use the right format for each of the subtasks. Check that you submit a tab-separated file with the requested number of columns and the correct information in each of the columns. Make sure you do not have duplicate predictions and that you specify correctly the ranks and confidence scores.

The most common prediction mistakes are:
1) incorrect column specification.
2) duplicate ranks (given a specific patent identifier, for the CEMP and GPRO task, it is not possible that two mentions have the same rank)
3) Incorrect range of ranks (ranks should start with 1)

Each column has to be separated by tabs, each row by newline characters.

In case you are not sure about the prediction format you can check out the baseline prediction example files at:

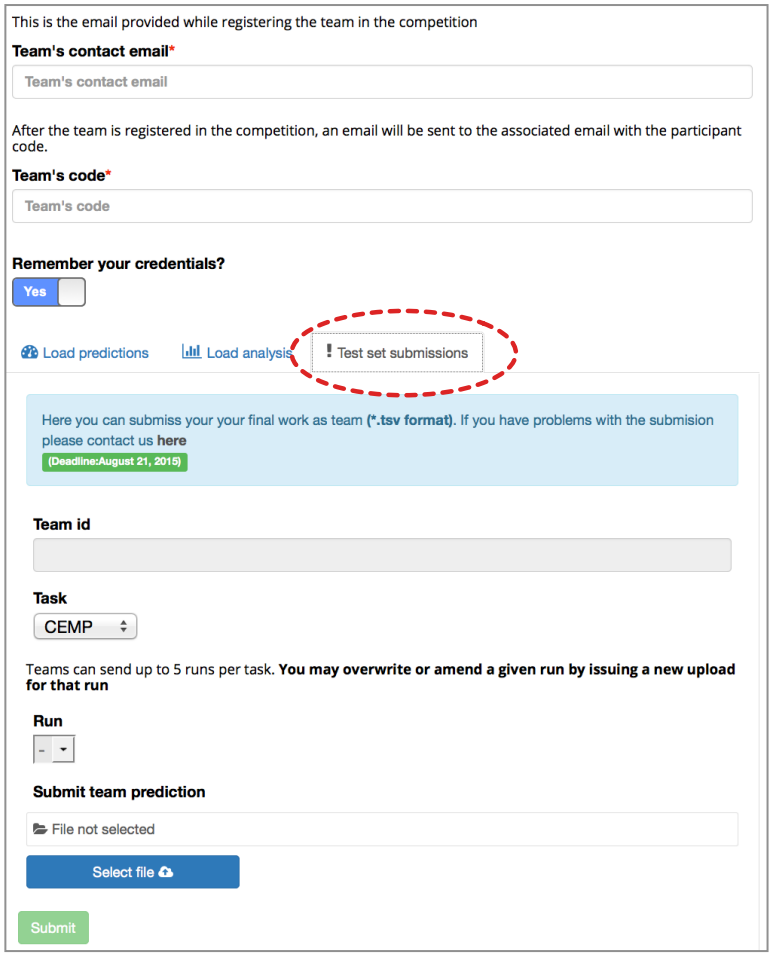
CEMP baseline:
http://www.biocreative.org/media/store/files/2015/CEMP_development_baseline_v02.tar.gz
GPRO baseline:
http://www.biocreative.org/media/store/files/2015/GPRO_development_baseline.tar.gz

3. Test set prediction submissions

Test set predictions have to be uploaded at the following webpage:

http://markyt.org/biocreative/analysis

At this page there is a submission tab called: 'Test set submissions'



You are allowed to submit up to 5 runs for each task. You have to fill in your team information and specify the Task and the Run.

4. CPD task prediction format

Column 1: patent identifier
Column 2: classification prediction. Has to be either 1 (does contain chemicals) or 0 (does not contain chemical mentions)
Columns 3: rank (a unique integer > 0 for that classification type - i.e., all true classifications and all false classifications each have to have unique ranks).
Column 4: confidence (a floating point number in the range (0,1], not more precise than the decimal precision available to Python on your OS)

5. CEMP task prediction format

For the CEMP task we will only request the prediction of the chemical mention offsets following a similar stetting as done for the BioCreative IV CHEMDNER task on PubMed abstracts. Given a set of patent abstracts, the participants have to return the start and end indices corresponding to all the chemical entities mentioned in this document.

It consists of tab-separated columns containing:

Column 1- Patent identifier
Column 2- Offset string consisting in a triplet joined by the ':' character. You have to provide the text type (T
: Title, A:Abstract), the start offset and the end offset.
Column 3- The rank of the chemical entity returned for this document
Column 4- A confidence score
Column 5- The string of the chemical entity mention

An example illustrating the prediction format is shown below:

WO2009026621A1	A:12:24	1	0.99	paliperidone
WO2011115938A1	T:0:17	1	0.99	Spiro-tetracyclic
WO2011115687A2	A:0:12	1	0.99	SP-B
WO2011115687A2	T:0:22	2	0.98989	Alkylated
WO2011115687A2	A:104:117	3	0.98978	SP-B
US20050101595	A:0:13	1	0.99	Aminothiazole
US20050101595	A:60:67	2	0.98989	2-amino
US20050101595	T:0:50	3	0.98978	N-containing
US20050101595	A:29:52	4	0.98967	N-containing
WO2010147138A1	A:252:262	1	0.99	nucleotide
WO2010147138A1	A:363:373	2	0.98989	amino
WO2010147138A1	A:92:102	3	0.98978	fatty
CN103087254A	A:196:218	1	0.99	stearyl

6. GPRO task prediction format

For the GPRO task we will only request the prediction of the GPRO mention offsets following a similar stetting as done for CEMP task. Given a set of patent abstracts, the participants have to return the start and end indices corresponding to all the GPRO type 1 entities mentioned in this document.

It consists of tab-separated columns containing:
Column 1- Patent identifier
Column 2- Offset string consisting in a triplet joined by the ':' character. You have to provide the text type (T: Title, A: Abstract), the start offset and the end offset.
Column 3- The rank of the chemical entity returned for this document
Column 4- A confidence score

Column 5- The string of the GPRO entity mention (type I)
An example illustrating the prediction format is shown below:

CA2274314C	T:0:7	1	0.99	Insulin
CA2275686C	A:377:379	1	0.99	R2
CA2282254C	A:227:239	1	0.99	HIV protease
CA2282254C	T:49:61	2	0.99	hiv protease
CA2291391C	A:365:368	1	0.99	BET
CA2295678C	A:150:156	1	0.99	mGluR5
CA2308315C	A:74:76	1	0.99	R2
CA2318184C	A:266:268	1	0.99	R2