UNIVERSITÄT LEIPZIG

ScaDS DRESDEN LEIPZIG

ASV Automatische Sprachverarbeitung

European Summer School 2017

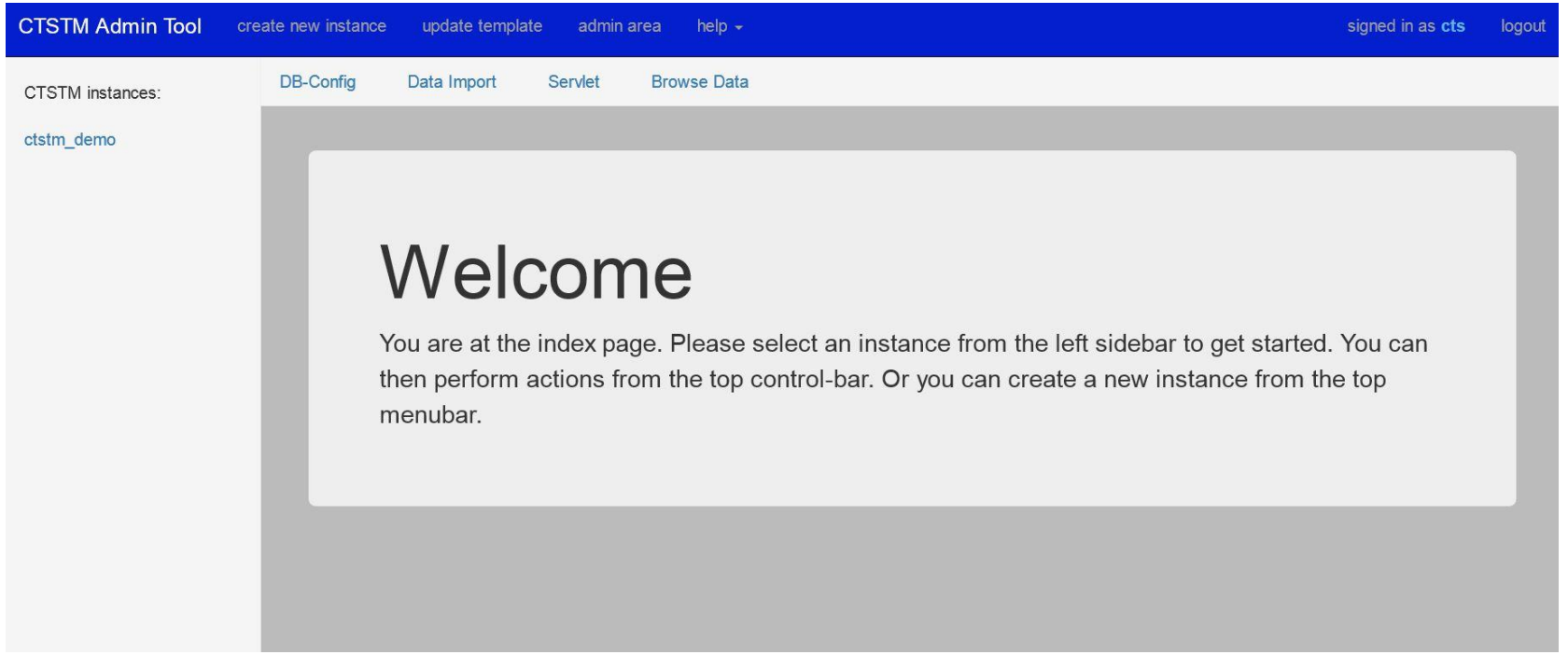
Text Mining with Canonical Text Services

Theory Session 6 – Canonical Text Miner

Federal Ministry of Education and Research

CLARIN-D

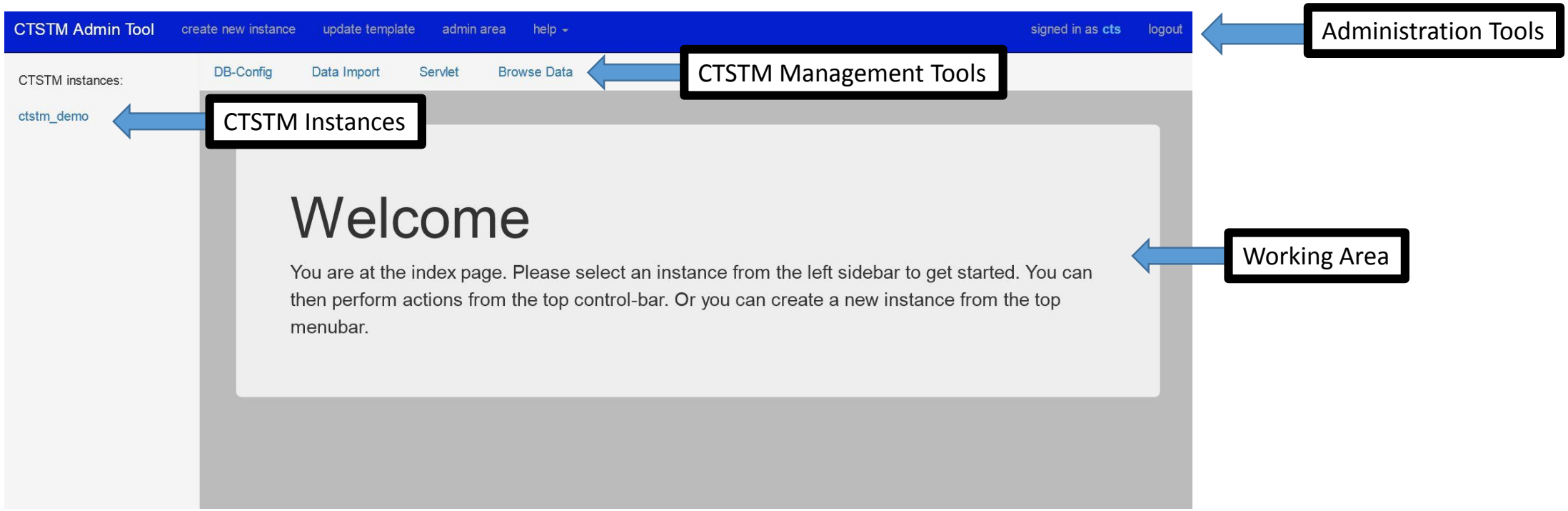
Overview



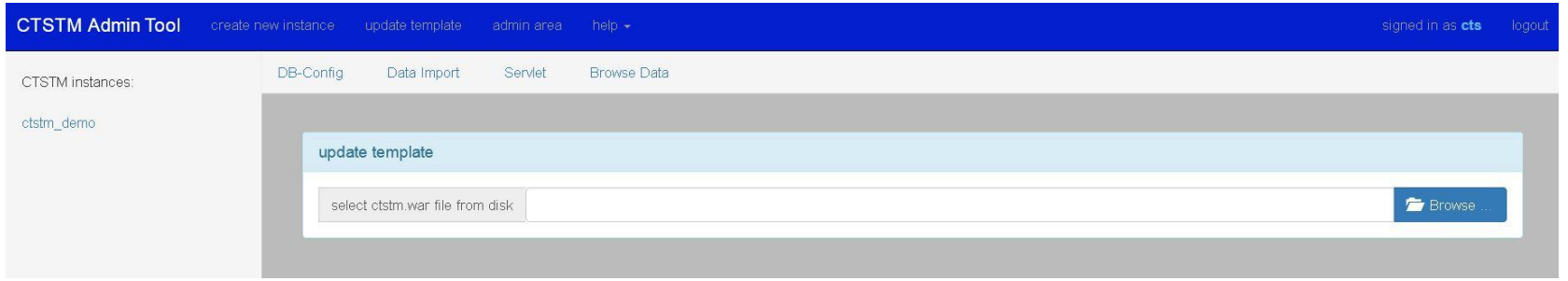
Can be deleted and redeployed without any impact on the CTS instances (In case the admin user is deleted or something else breaks).

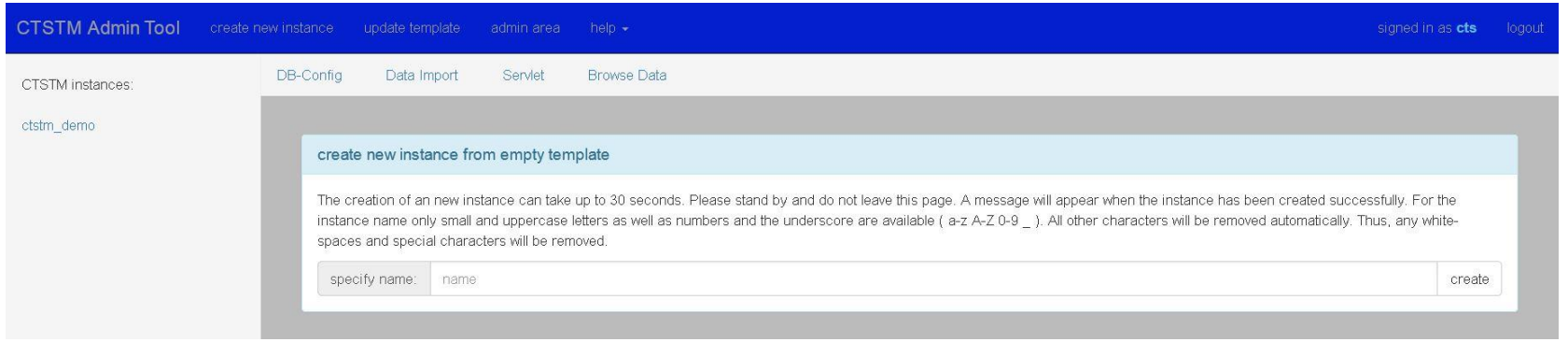
Improvised „Re-Skin“ of CTS Admin Tool

Overview


CTSTM Admin Tool
create new instance
update template
admin area
help
signed in as cts
logout
CTSTM instances:
ctstm_demo
DB-Config
Data Import
Servlet
Browse Data
Welcome
You are at the index page. Please select an instance from the left sidebar to get started. You can then perform actions from the top control-bar. Or you can create a new instance from the top menubar.
Administration Tools
CTSTM Management Tools
CTSTM Instances
Working Area

Create New Instance / Update .war File



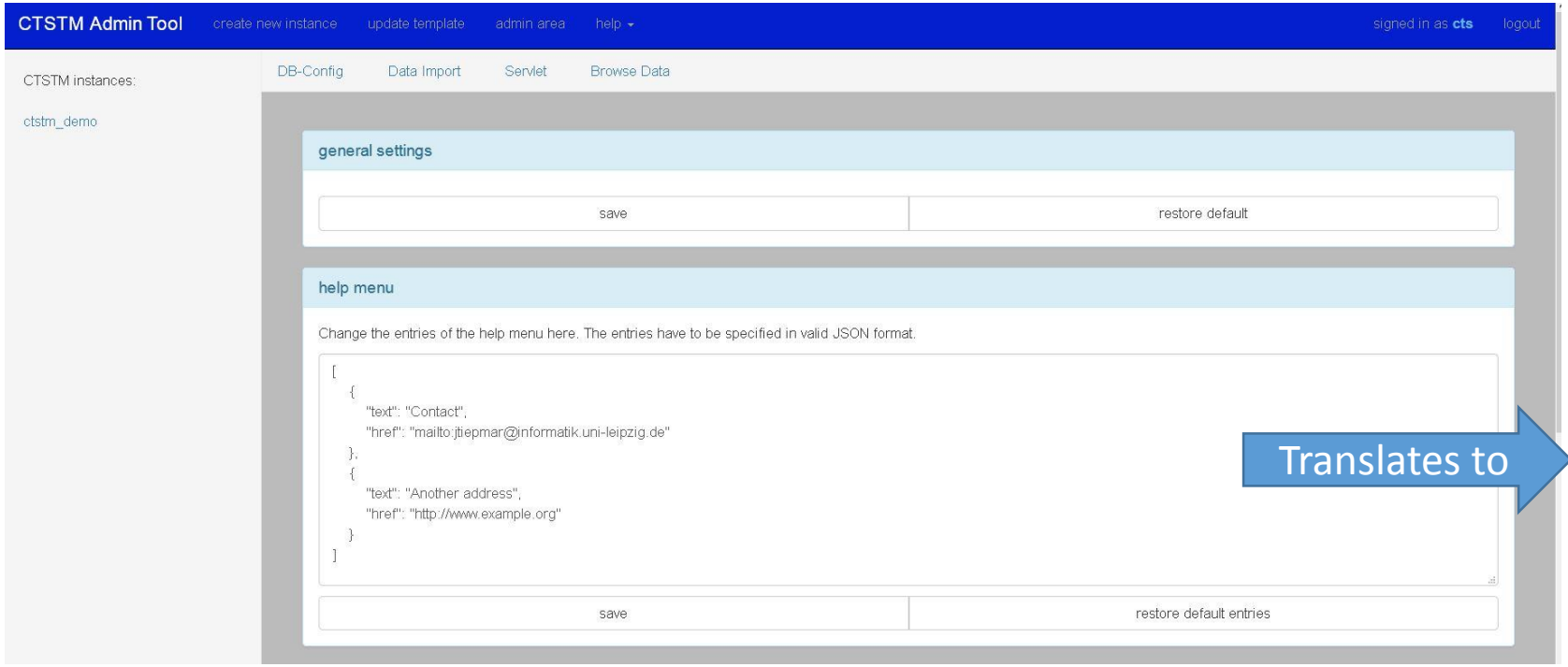


Old .war files are still available in admin area after update

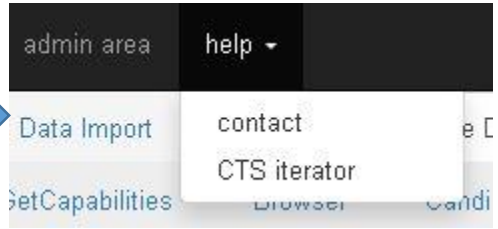
New .war file must be selected in admin area

Updating the template does not update the CTS instances (-> Servlet Menu)

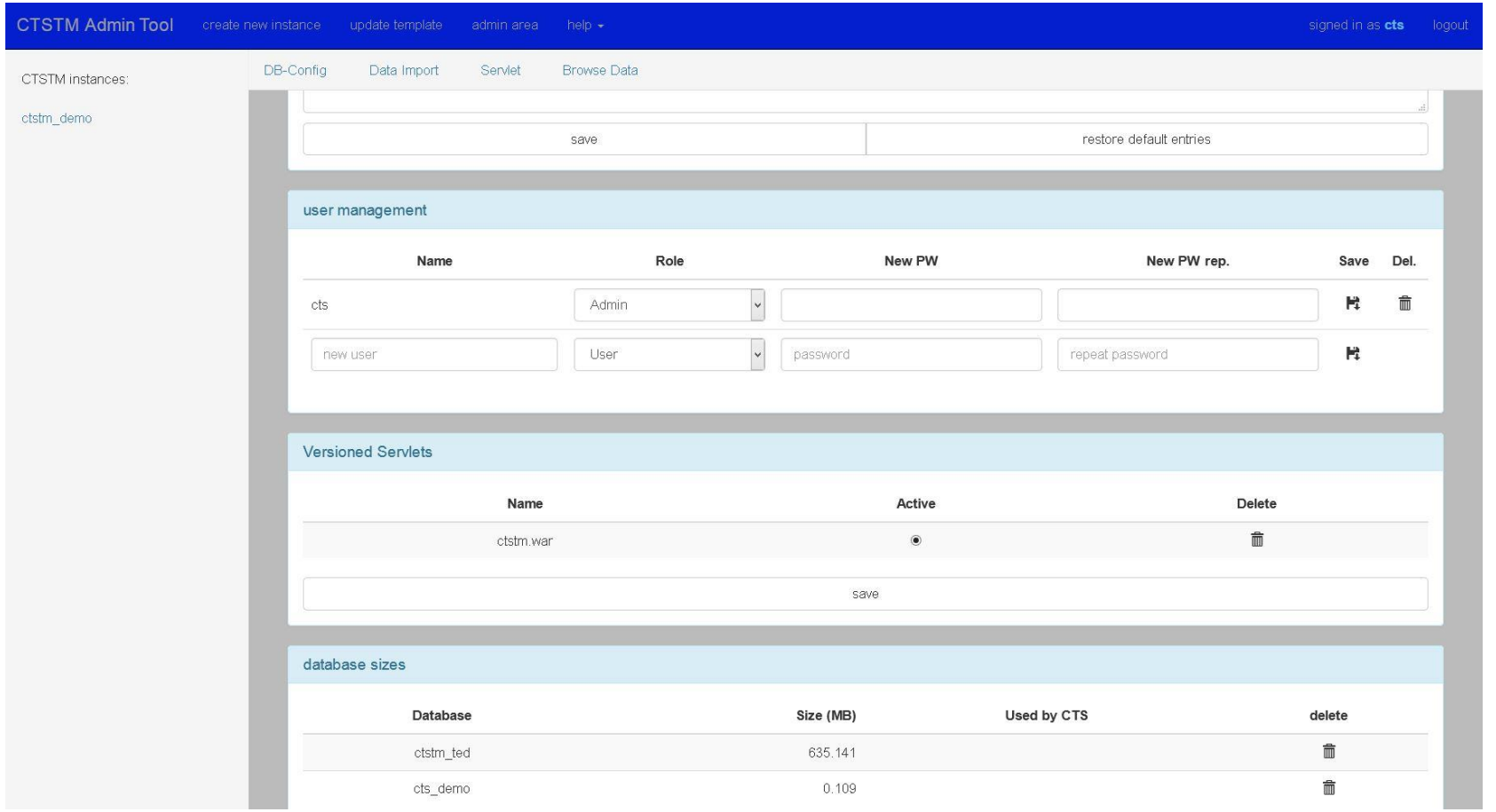
Admin Area / User Management



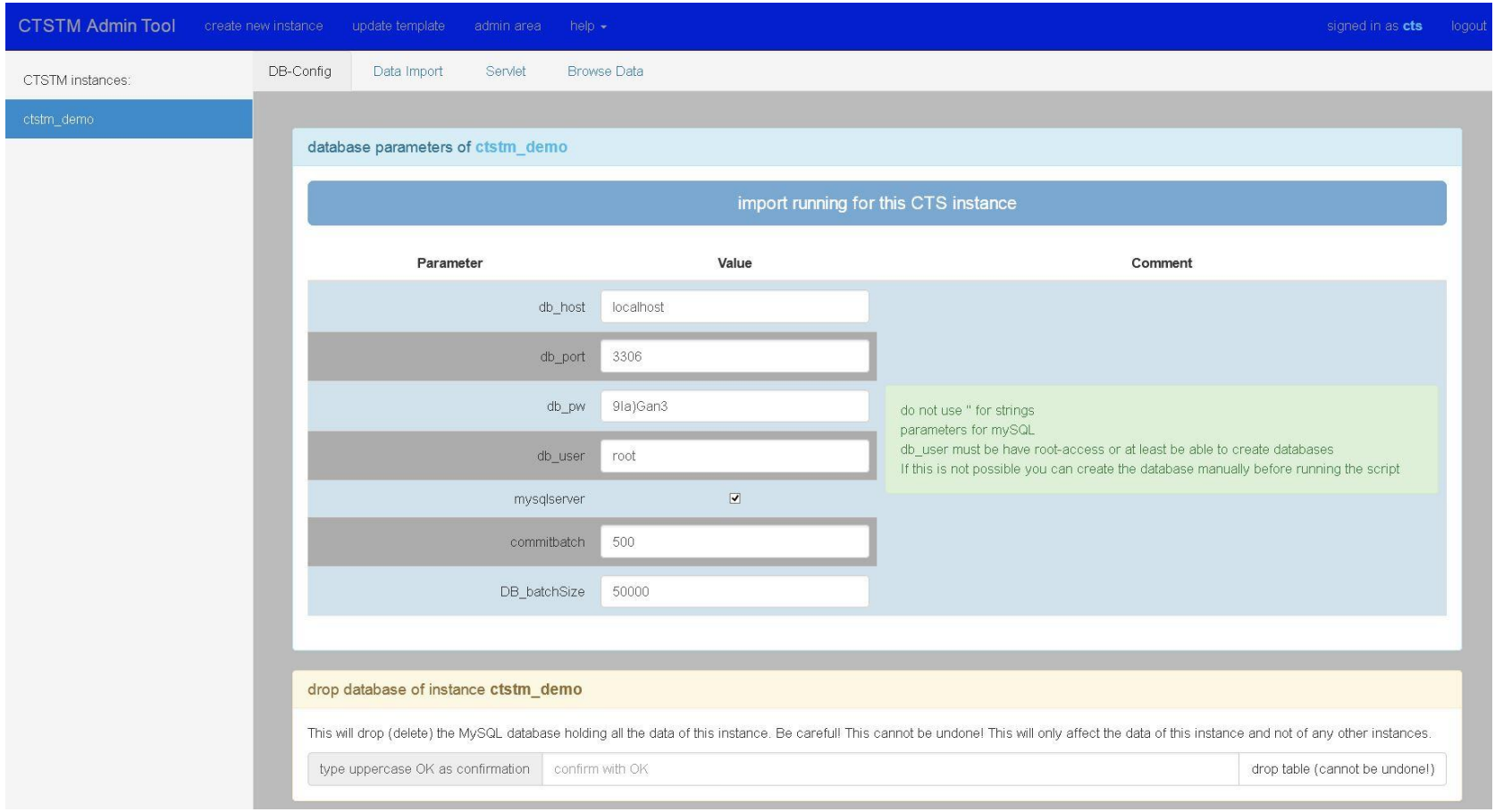
Translates to



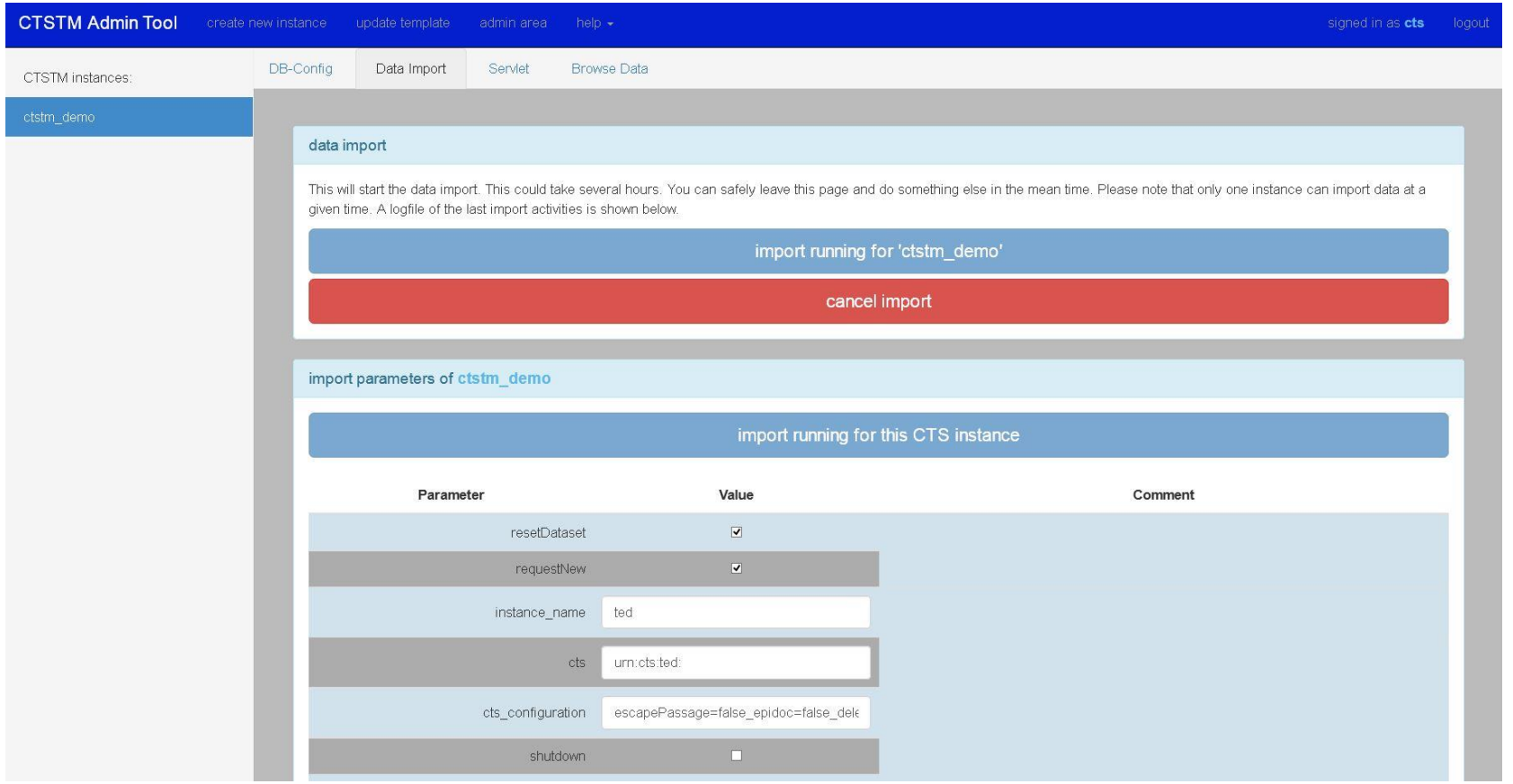
Admin Area / User Management



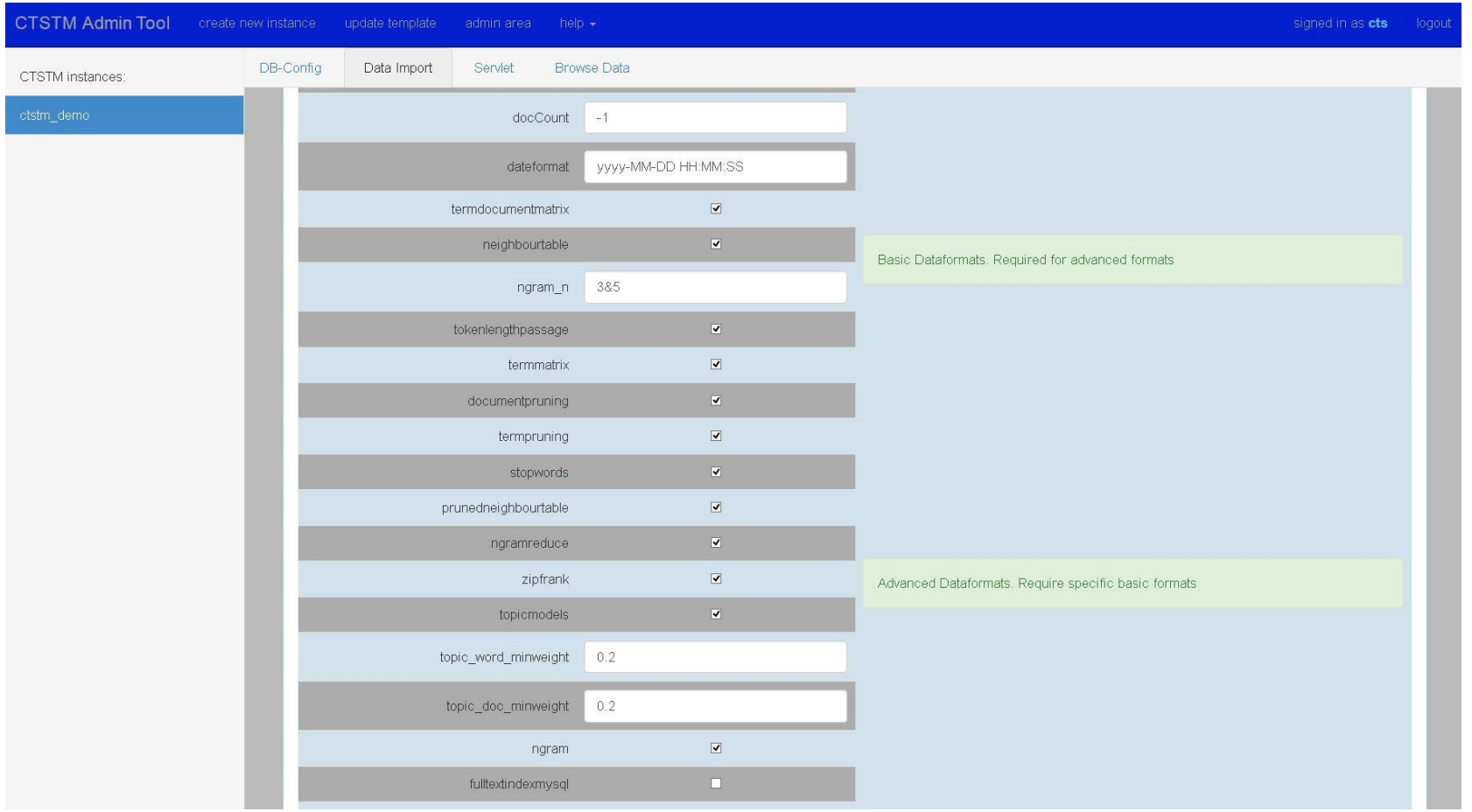
Database Configuration



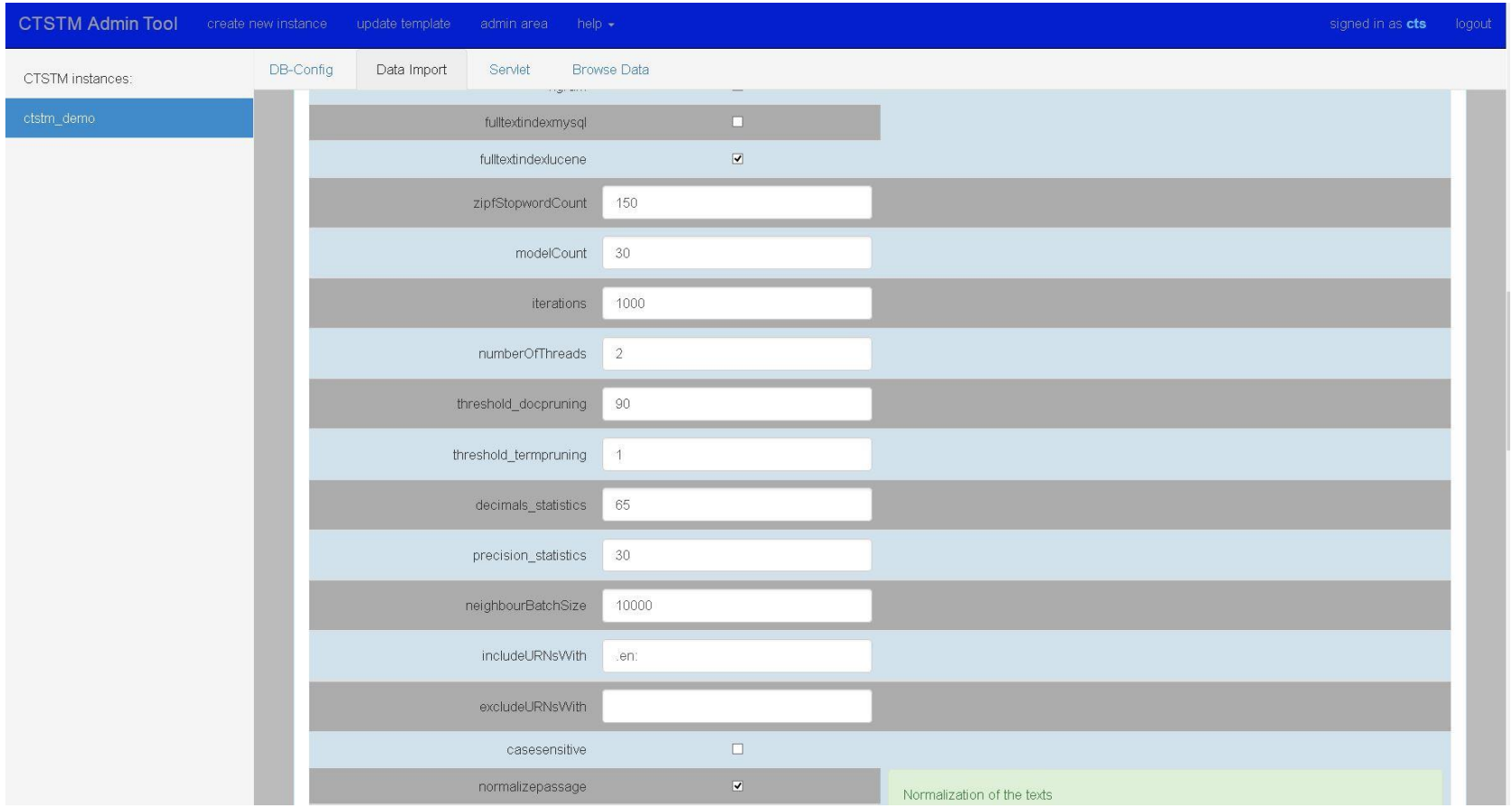
Data Import


CTSTM Admin Tool
create new instance
update template
admin area
help
signed in as cts
logout
CTSTM instances:
ctstm_demo
DB-Config
Data Import
Servlet
Browse Data
data import
This will start the data import. This could take several hours. You can safely leave this page and do something else in the mean time. Please note that only one instance can import data at a given time. A logfile of the last import activities is shown below.
import running for 'ctstm_demo'
cancel import
import parameters of ctstm_demo
import running for this CTS instance
Parameter
Value
Comment
resetDataset
requestNew
instance_name
ted
cts
urn:cts:ted:
cts_configuration
escapePassage=false_epidoc=false_del
shutdown

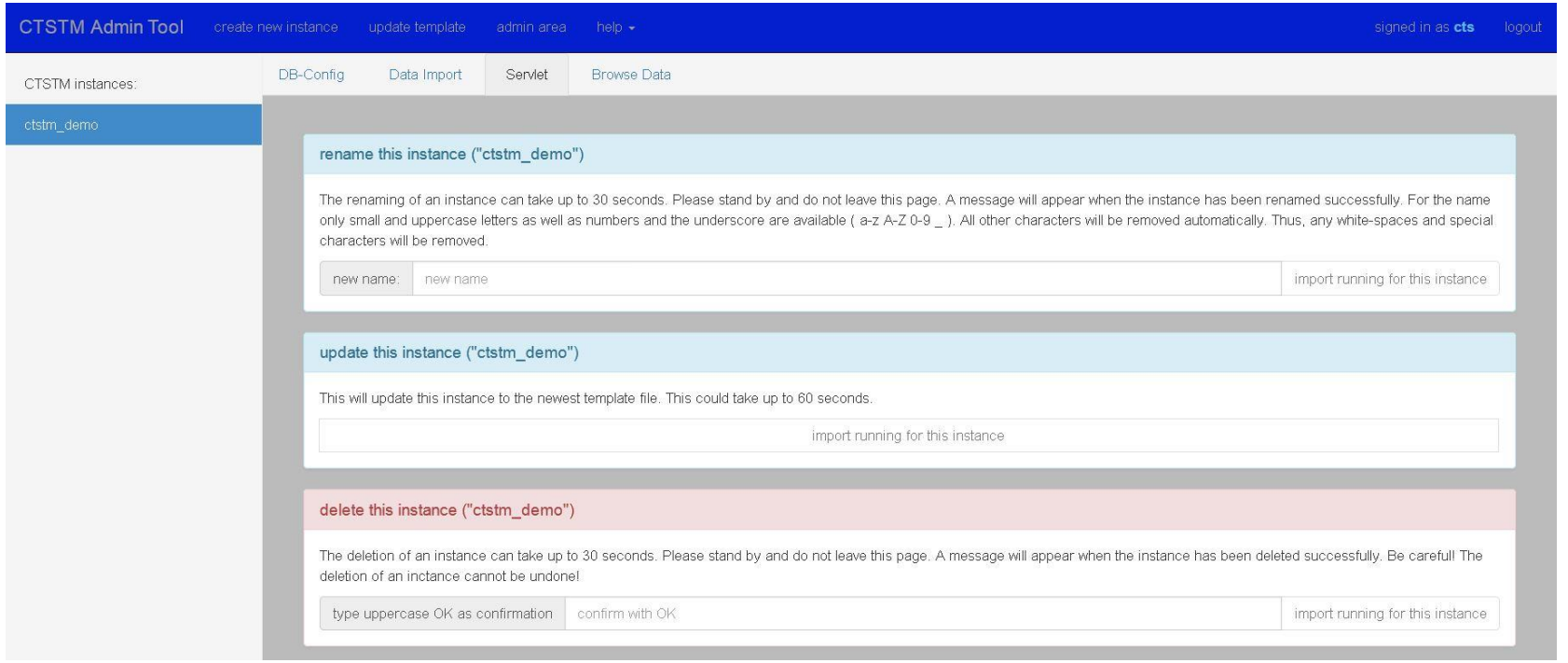
Data Import


CTSTM Admin Tool
create new instance
update template
admin area
help
signed in as cts
logout
CTSTM instances:
ctstm_demo
DB-Config
Data Import
Servlet
Browse Data
docCount
-1
dateformat
yyyy-MM-DD HH:MM:SS
termdocumentmatrix
neighbourtable
Basic Dataformats. Required for advanced formats
ngram_n
3&5
tokenlengthpassage
termmatrix
documentpruning
termpruning
stopwords
prunedneighbourtable
ngramreduce
zipfrank
Advanced Dataformats. Require specific basic formats
topicmodels
topic_word_minweight
0.2
topic_doc_minweight
0.2
ngram
fulltextindexmysql

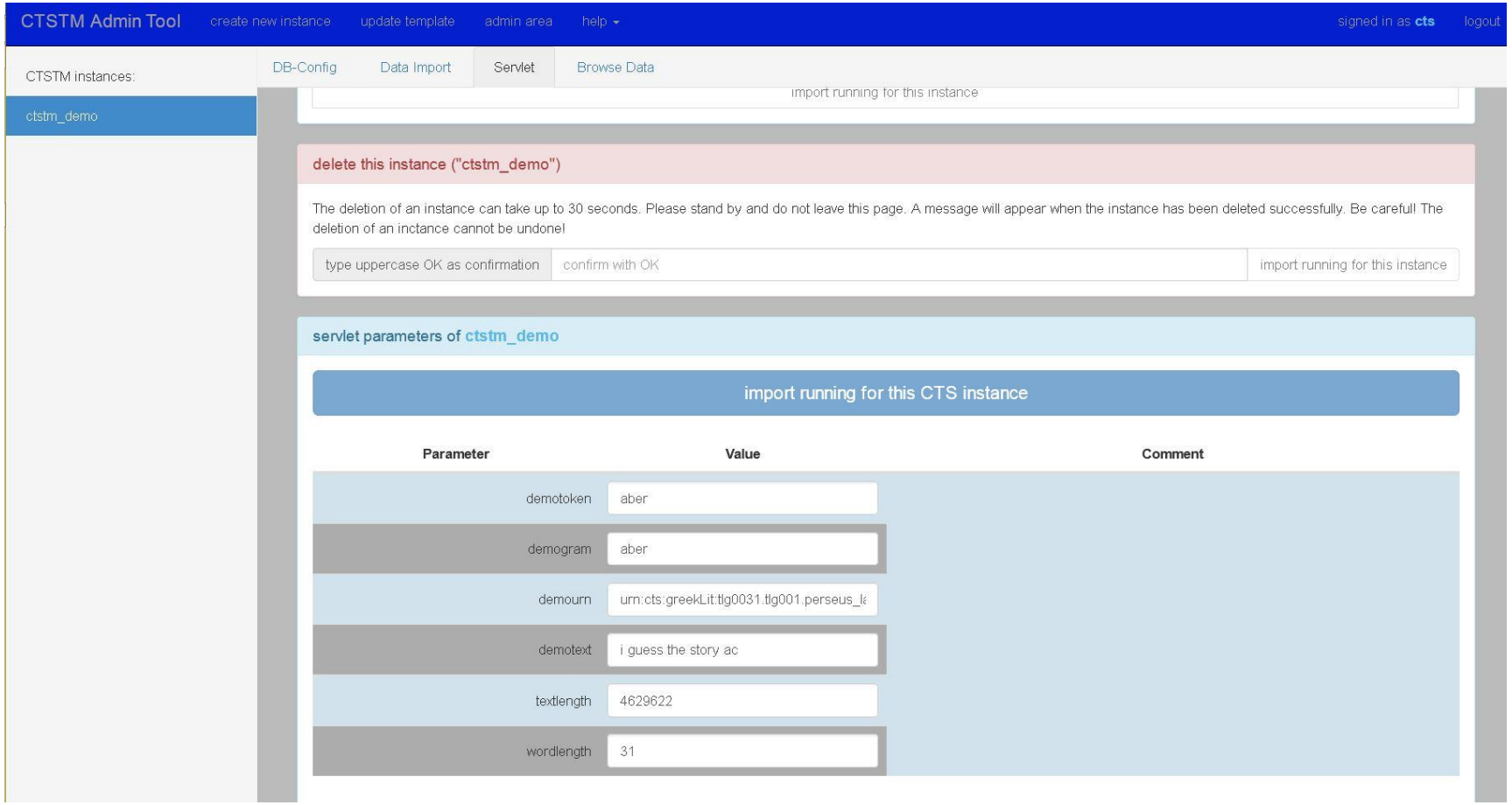
Data Import


CTSTM Admin Tool
create new instance
update template
admin area
help
signed in as cts
logout
CTSTM instances:
ctstm_demo
DB-Config
Data Import
Servlet
Browse Data
fulltextindexmysql
fulltextindexlucene
zipfStopwordCount
150
modelCount
30
iterations
1000
numberOfThreads
2
threshold_docpruning
90
threshold_termpruning
1
decimals_statistics
65
precision_statistics
30
neighbourBatchSize
10000
includeURNsWith
.en:
excludeURNsWith
casesensitive
normalizepassage
Normalization of the texts

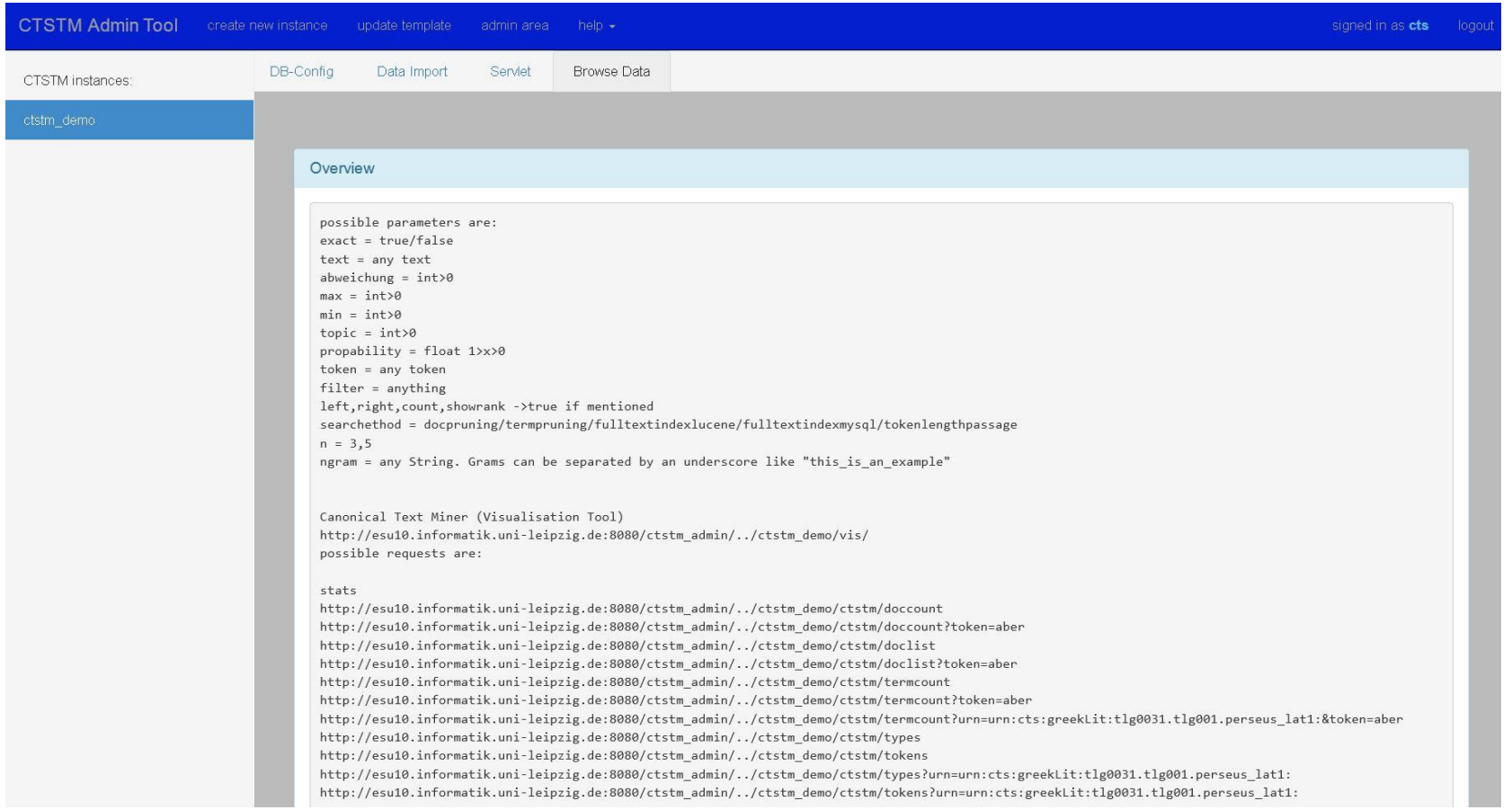
Servlet Management



Servlet Management



Browse – “API“


CTSTM Admin Tool
create new instance
update template
admin area
help
signed in as cts
logout
CTSTM instances:
ctstm_demo
DB-Config
Data Import
Servlet
Browse Data
Overview
possible parameters are:
exact = true/false
text = any text
abweichung = int>0
max = int>0
min = int>0
topic = int>0
propability = float 1>x>0
token = any token
filter = anything
left,right,count,showrank ->true if mentioned
searchethod = docpruning/termpruning/fulltextindexlucene/fulltextindexmysql/tokenlengthpassage
n = 3,5
ngram = any String. Grams can be separated by an underscore like "this_is_an_example"
Canonical Text Miner (Visualisation Tool)
http://esu10.informatik.uni-leipzig.de:8080/ctstm_admin/../ctstm_demo/vis/
possible requests are:
stats
http://esu10.informatik.uni-leipzig.de:8080/ctstm_admin/../ctstm_demo/ctstm/doccount
http://esu10.informatik.uni-leipzig.de:8080/ctstm_admin/../ctstm_demo/ctstm/doccount?token=aber
http://esu10.informatik.uni-leipzig.de:8080/ctstm_admin/../ctstm_demo/ctstm/doclist
http://esu10.informatik.uni-leipzig.de:8080/ctstm_admin/../ctstm_demo/ctstm/doclist?token=aber
http://esu10.informatik.uni-leipzig.de:8080/ctstm_admin/../ctstm_demo/ctstm/termcount
http://esu10.informatik.uni-leipzig.de:8080/ctstm_admin/../ctstm_demo/ctstm/termcount?token=aber
http://esu10.informatik.uni-leipzig.de:8080/ctstm_admin/../ctstm_demo/ctstm/termcount?urn=urn:cts:greekLit:tlg0031.tlg001.perseus_lat1:&token=aber
http://esu10.informatik.uni-leipzig.de:8080/ctstm_admin/../ctstm_demo/ctstm/types
http://esu10.informatik.uni-leipzig.de:8080/ctstm_admin/../ctstm_demo/ctstm/tokens
http://esu10.informatik.uni-leipzig.de:8080/ctstm_admin/../ctstm_demo/ctstm/types?urn=urn:cts:greekLit:tlg0031.tlg001.perseus_lat1:
http://esu10.informatik.uni-leipzig.de:8080/ctstm_admin/../ctstm_demo/ctstm/tokens?urn=urn:cts:greekLit:tlg0031.tlg001.perseus_lat1:

Contact

Jochen Tiepmar
E-Mail: jtiepmar@informatik.uni-leipzig.de

Scalable Data Solutions (ScaDS) Leipzig
Universität Leipzig
Ritterstraße 9-13
04109 Leipzig