Usage as GATE plugin

Alexander Schutz

November 11, 2008

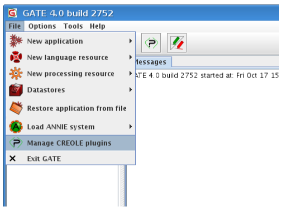
The keyphrase extraction has been developed as a number of GATE plugins and it makes use of some of the nice features the GATE framework has to offer. Therefore, it is possible to run the keyphrase extraction as a pipeline inside the GATE gui, for testing or evaluation purposes, or for prototyping. The following provides a complete walk through of what has to be done in order to run the keyphrase extraction as GATE plugin, and how results can be inspected.

1. Download the GATE plugins at `http://resources.smile.deri.ie/nlp/keyphrase-extraction/current/keyphrase-extraction-1.0.0-gate.zip`.

2. Unzip in a folder of your choice, in the remainder denoted as keyphrase-extraction-folder.

3. Start GATE.

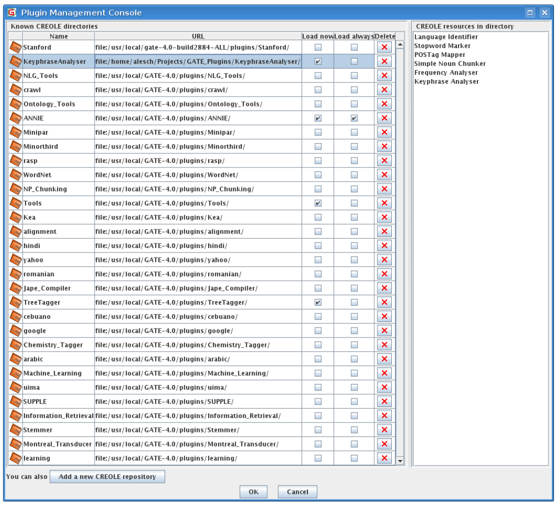
4. When the user interface comes up, select *File > Manage CREOLE plugins*.



ATE 4.0 build 2752 started at: Fri Oct 17 15

5. In the following dialogue, click on *Add a new CREOLE repository*, then select the directory *keyphrase-extraction-folder* (this is where you unzipped the plugins), confirm with OK in the file chooser. Now, an additional entry should appear in the Plugin Management Console, named *KeyphraseAnalyser*. By clicking on it, a number of CREOLE resources should be listed on the right hand side.

6. Click on *Load Now* (or *Load Always*).



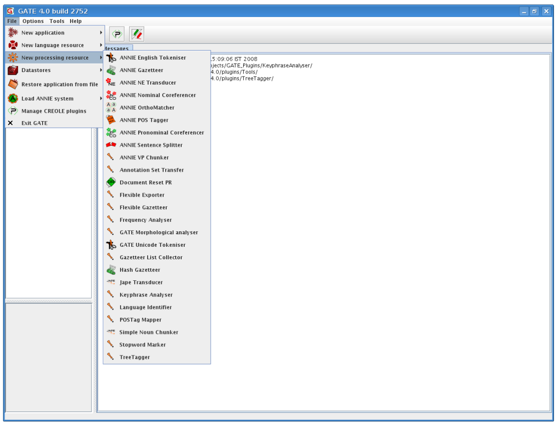
Name	URL	Load now	Load always	Delete
Stanford	file:/usr/local/gate-4.0-build2884-ALL/plugins/Stanford/	☐	☐	✕
KeyphraseAnalyser	file:/home/alesch/Projects/GATE_Plugins/KeyphraseAnalyser/	☑	☐	✕
NLG_Tools	file:/usr/local/GATE-4.0/plugins/NLG_Tools/	☐	☐	✕
crawl	file:/usr/local/GATE-4.0/plugins/crawl/	☐	☐	✕
Ontology_Tools	file:/usr/local/GATE-4.0/plugins/Ontology_Tools/	☐	☐	✕
ANNIE	file:/usr/local/GATE-4.0/plugins/ANNIE/	☑	☑	✕
Minipar	file:/usr/local/GATE-4.0/plugins/Minipar/	☐	☐	✕
Minorthird	file:/usr/local/GATE-4.0/plugins/Minorthird/	☐	☐	✕
rasp	file:/usr/local/GATE-4.0/plugins/rasp/	☐	☐	✕
WordNet	file:/usr/local/GATE-4.0/plugins/WordNet/	☐	☐	✕
NP_Chunking	file:/usr/local/GATE-4.0/plugins/NP_Chunking/	☐	☐	✕
Tools	file:/usr/local/GATE-4.0/plugins/Tools/	☑	☐	✕
Kea	file:/usr/local/GATE-4.0/plugins/Kea/	☐	☐	✕
alignment	file:/usr/local/GATE-4.0/plugins/alignment/	☐	☐	✕
hindi	file:/usr/local/GATE-4.0/plugins/hindi/	☐	☐	✕
yahoo	file:/usr/local/GATE-4.0/plugins/yahoo/	☐	☐	✕
romanian	file:/usr/local/GATE-4.0/plugins/romanian/	☐	☐	✕
Jape_Compiler	file:/usr/local/GATE-4.0/plugins/Jape_Compiler/	☐	☐	✕
TreeTagger	file:/usr/local/GATE-4.0/plugins/TreeTagger/	☑	☐	✕
cebuano	file:/usr/local/GATE-4.0/plugins/cebuano/	☐	☐	✕
google	file:/usr/local/GATE-4.0/plugins/google/	☐	☐	✕
Chemistry_Tagger	file:/usr/local/GATE-4.0/plugins/Chemistry_Tagger/	☐	☐	✕
arabic	file:/usr/local/GATE-4.0/plugins/arabic/	☐	☐	✕
Machine_Learning	file:/usr/local/GATE-4.0/plugins/Machine_Learning/	☐	☐	✕
uima	file:/usr/local/GATE-4.0/plugins/uima/	☐	☐	✕
SUPPLE	file:/usr/local/GATE-4.0/plugins/SUPPLE/	☐	☐	✕
Information_Retrieval	file:/usr/local/GATE-4.0/plugins/Information_Retrieval/	☐	☐	✕
Stemmer	file:/usr/local/GATE-4.0/plugins/Stemmer/	☐	☐	✕
Montreal_Transducer	file:/usr/local/GATE-4.0/plugins/Montreal_Transducer/	☐	☐	✕
learning	file:/usr/local/GATE-4.0/plugins/learning/	☐	☐	✕

CREOLE resources in directory

Language Identifier
Stopword Marker
POSTag Mapper
Simple Noun Chunker
Frequency Analyser
Keyphrase Analyser

Now is also a good time to enable/load a number of different GATE plugins (namely *ANNIE* and *Tools*, and for German and French texts *TreeTagger*), as they are needed for the Keyphrase Extraction pipeline. Mark the mentioned plugins as *Load now* or *Load always* and return to the GATE main interface by clicking *OK* in the Plugin Management Console. The next steps show how to construct the keyphrase extraction processing pipeline

7. Load the required Processing Resources. To do so, right click on the *Processing Resources* field in the panel on left hand side. A menu item *New* should appear, which, when hovered over by the mouse, should trigger the display of enabled Processing Resources. Alternatively, you can click on the Menu *File > New Processing Resource.*

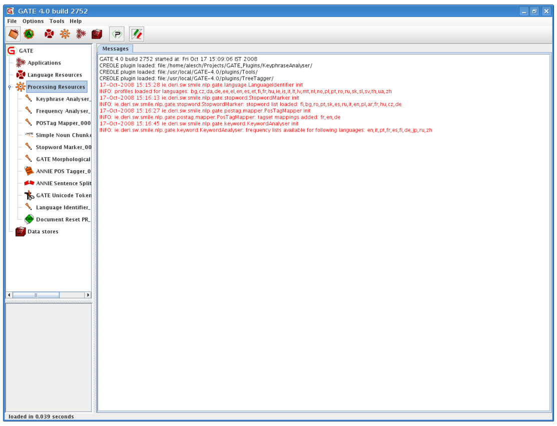


```
15:09:06 IST 2008
jects/GATE_Plugins/KeyphraseAnalyser/
4.0/plugins/Tools/
4.0/plugins/TreeTagger/
```

8. Choose the following Processing Resources (you can blindly accept/confirm the upcoming dialogue, there is no need to adjust the init-time parameters):

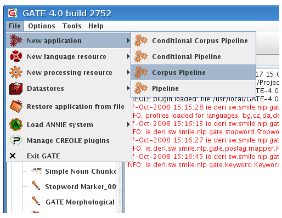
   - Document Reset PR
   - Language Identifier
   - GATE Unicode Tokeniser
   - ANNIE Sentence Splitter
   - ANNIE POS Tagger
   - GATE Morphological Analyser
   - Stopword Marker
   - Simple Noun Chunker
   - POSTag Mapper
   - Frequency Analyser
   - Keyphrase Analyser

   While loading these Processing Resources, you should be able to observe some informative output in the Messages tab.



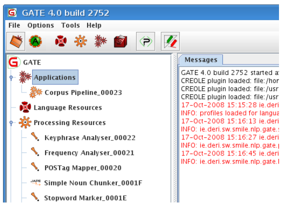
```
GATE 4.0 build 2752 started at: Fri Oct 17 15:09:06 IST 2008
CREOLE plugin loaded: file:/home/alesch/Projects/GATE_Plugins/KeyphraseAnalyser/
CREOLE plugin loaded: file:/usr/local/GATE-4.0/plugins/Tools/
CREOLE plugin loaded: file:/usr/local/GATE-4.0/plugins/TreeTagger/
17-Oct-2008 15:15:28 ie.deri.sw.smile.nlp.gate.language.LanguageIdentifier init
INFO: profiles loaded for languages: bg,cz,da,de,ee,el,en,es,et,fi,fr,hu,ie,is,it,lt,lv,mt,nl,no,pl,pt,ro,ru,sk,sl,sv,th,ua,zh
17-Oct-2008 15:16:13 ie.deri.sw.smile.nlp.gate.stopword.StopwordMarker init
INFO: ie.deri.sw.smile.nlp.gate.stopword.StopwordMarker: stopword list loaded: fi,bg,ro,pt,sk,es,ru,it,en,pl,ar,fr,hu,cz,de
17-Oct-2008 15:16:27 ie.deri.sw.smile.nlp.gate.postag.mapper.PosTagMapper init
INFO: ie.deri.sw.smile.nlp.gate.postag.mapper.PosTagMapper: tagset mappings added: fr,en,de
17-Oct-2008 15:16:45 ie.deri.sw.smile.nlp.gate.keyword.KeywordAnalyser init
INFO: ie.deri.sw.smile.nlp.gate.keyword.KeywordAnalyser: frequency lists available for following languages: en,it,pt,fr,es,fi,de,jp,ru,zh
```

9. Now, right click on *Applications* in the panel on left hand side, and select *New > Corpus Pipeline.* Alternatively, you can also select from the *Menu File > New Application > Corpus Pipeline.*



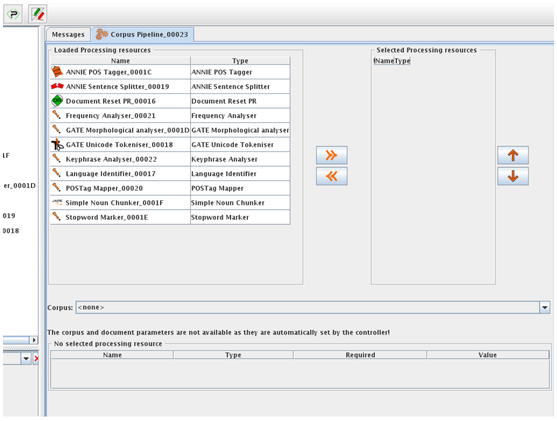
```
17 15:0
/Projec
TE-4.0
EOLE plugin loaded: file:/usr/local/GATE-4.0
7-Oct-2008 15:15:28 ie.deri.sw.smile.nlp.gat
FO: profiles loaded for languages: bg,cz,da,de
7-Oct-2008 15:16:13 ie.deri.sw.smile.nlp.gat
FO: ie.deri.sw.smile.nlp.gate.stopword.Stopwo
7-Oct-2008 15:16:27 ie.deri.sw.smile.nlp.gat
FO: ie.deri.sw.smile.nlp.gate.postag.mapper.P
7-Oct-2008 15:16:45 ie.deri.sw.smile.nlp.gat
INFO: ie.deri.sw.smile.nlp.gate.keyword.Keyword
```

You may name the pipeline and confirm, after that an instance of the Corpus Pipeline with the assigned name should appear in the panel on the left hand side.



```
GATE 4.0 build 2752 started a
CREOLE plugin loaded: file:/hor
CREOLE plugin loaded: file:/usr
CREOLE plugin loaded: file:/usr
17-Oct-2008 15:15:28 ie.deri
INFO: profiles loaded for langua
17-Oct-2008 15:16:13 ie.deri
INFO: ie.deri.sw.smile.nlp.gate.s
17-Oct-2008 15:16:27 ie.deri
INFO: ie.deri.sw.smile.nlp.gate.p
17-Oct-2008 15:16:45 ie.deri
INFO: ie.deri.sw.smile.nlp.gate.k
```

10. Double click the newly created instance of Corpus Pipeline, and a tab opens in the central area of the screen, which provides a view of the *Loaded Processing Resources* (left hand side) and the *Selected Processing Resources* (right hand side). Currently, no Processing Resource is selected.



Messages | Corpus Pipeline_00023

Loaded Processing resources

Name	Type
ANNIE POS Tagger_0001C	ANNIE POS Tagger
ANNIE Sentence Splitter_00019	ANNIE Sentence Splitter
Document Reset PR_00016	Document Reset PR
Frequency Analyser_00021	Frequency Analyser
GATE Morphological analyser_0001D	GATE Morphological analyser
GATE Unicode Tokeniser_00018	GATE Unicode Tokeniser
Keyphrase Analyser_00022	Keyphrase Analyser
Language Identifier_00017	Language Identifier
POSTag Mapper_00020	POSTag Mapper
Simple Noun Chunker_0001F	Simple Noun Chunker
Stopword Marker_0001E	Stopword Marker

Selected Processing resources

!NameType

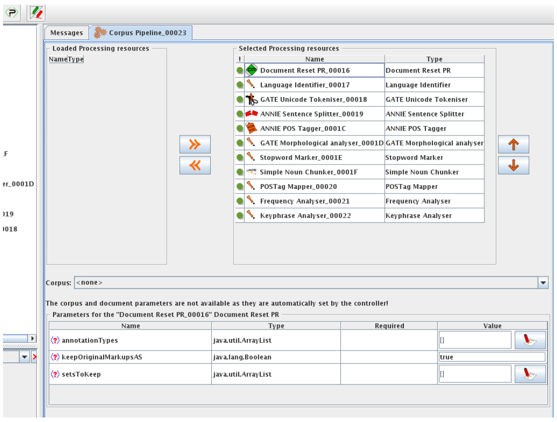
Corpus: <none>

The corpus and document parameters are not available as they are automatically set by the controller!

No selected processing resource

Name	Type	Required	Value

Mark all of the previously loaded Processing Resources, and use the arrow button pointing to the right hand side to assign the selected Processing Resources to the Processing Pipeline. Now, with the up and down arrows, assemble the Processing Resources to the order given at point 8.



Messages | Corpus Pipeline_00023

Loaded Processing resources

NameType

Selected Processing resources

!	Name	Type
●	Document Reset PR_00016	Document Reset PR
●	Language Identifier_00017	Language Identifier
●	GATE Unicode Tokeniser_00018	GATE Unicode Tokeniser
●	ANNIE Sentence Splitter_00019	ANNIE Sentence Splitter
●	ANNIE POS Tagger_0001C	ANNIE POS Tagger
●	GATE Morphological analyser_0001D	GATE Morphological analyser
●	Stopword Marker_0001E	Stopword Marker
●	Simple Noun Chunker_0001F	Simple Noun Chunker
●	POSTag Mapper_00020	POSTag Mapper
●	Frequency Analyser_00021	Frequency Analyser
●	Keyphrase Analyser_00022	Keyphrase Analyser

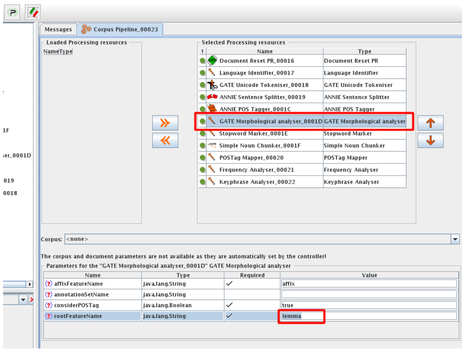
Corpus: <none>

The corpus and document parameters are not available as they are automatically set by the controller!

Parameters for the "Document Reset PR_00016" Document Reset PR

Name	Type	Required	Value
annotationTypes	java.util.ArrayList		[]
keepOriginalMarkupsAS	java.lang.Boolean		true
setsToKeep	java.util.ArrayList		[]

11. Still in the Corpus Pipeline view, click on the *GATE Morphological Analyser*. In the lower part of the view, you can see the default runtime parameters. Click on the value *rootFeatureName*, which is set to "root" by default. You need to change this value to "lemma".



Messages | Corpus Pipeline_00023

Loaded Processing resources

NameType

Selected Processing resources

!	Name	Type
●	Document Reset PR_00016	Document Reset PR
●	Language Identifier_00017	Language Identifier
●	GATE Unicode Tokeniser_00018	GATE Unicode Tokeniser
●	ANNIE Sentence Splitter_00019	ANNIE Sentence Splitter
●	ANNIE POS Tagger_0001C	ANNIE POS Tagger
●	GATE Morphological analyser_0001D	GATE Morphological analyser
●	Stopword Marker_0001E	Stopword Marker
●	Simple Noun Chunker_0001F	Simple Noun Chunker
●	POSTag Mapper_00020	POSTag Mapper
●	Frequency Analyser_00021	Frequency Analyser
●	Keyphrase Analyser_00022	Keyphrase Analyser

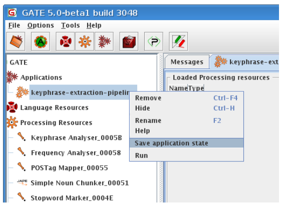
Corpus: <none>

The corpus and document parameters are not available as they are automatically set by the controller!

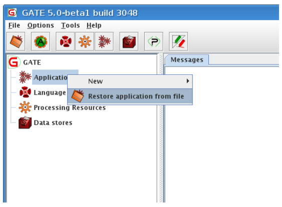
Parameters for the "GATE Morphological analyser_0001D" GATE Morphological analyser

Name	Type	Required	Value
affixFeatureName	java.lang.String	✓	affix
annotationSetName	java.lang.String		
considerPOSTag	java.lang.Boolean	✓	true
rootFeatureName	java.lang.String	✓	lemma

12. For convenience reasons, you can now save the application state to a file, in order to easily load the pipeline at a later point in time. This way you can avoid having to load every processing resource and the configuration process. To do so, simply right-click on the corpus pipeline created in the previous step, select *Save application state* and enter a file name, followed by the ".gapp" (or ".xgapp") extension.

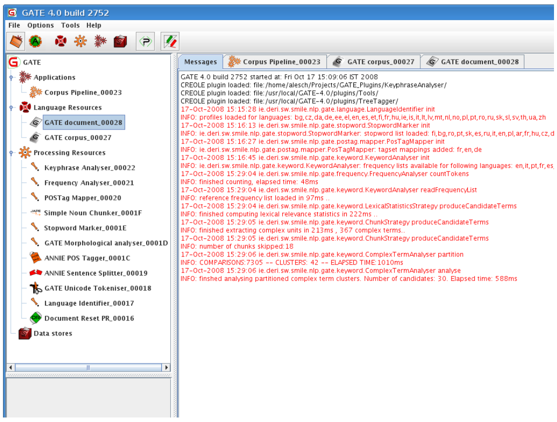


To load a previously saved application pipeline, select *File* > *Restore application from state*, and choose the desired *.gapp (or *.xgapp) file.



13. Create a corpus, populate it with some documents, and set it as the corpus to be processed by the Processing Pipeline.

14. When you run the previously constructed Processing Pipeline over your corpus, each document will be annotated with a number of additional features, which you can observe in the lower, left-hand corner after double-clicking on a document.



```
GATE 4.0 build 2752 started at: Fri Oct 17 15:09:06 IST 2008
CREOLE plugin loaded: file:/home/alesch/Projects/GATE_Plugins/KeyphraseAnalyser/
CREOLE plugin loaded: file:/usr/local/GATE-4.0/plugins/Tools/
CREOLE plugin loaded: file:/usr/local/GATE-4.0/plugins/TreeTagger/
17-Oct-2008 15:15:28 ie.deri.sw.smile.nlp.gate.language.LanguageIdentifier init
INFO: profiles loaded for languages: bg,cz,da,de,ee,el,en,es,et,fi,fr,hu,ie,is,it,lt,lv,mt,nl,no,pl,pt,ro,ru,sk,sl,sv,th,ua,zh
17-Oct-2008 15:16:13 ie.deri.sw.smile.nlp.gate.stopword.StopwordMarker init
INFO: ie.deri.sw.smile.nlp.gate.stopword.StopwordMarker: stopword list loaded: fi,bg,ro,pt,sk,es,ru,it,en,pl,ar,fr,hu,cz,d
17-Oct-2008 15:16:27 ie.deri.sw.smile.nlp.gate.postag.mapper.PosTagMapper init
INFO: ie.deri.sw.smile.nlp.gate.postag.mapper.PosTagMapper: tagset mappings added: fr,en,de
17-Oct-2008 15:16:45 ie.deri.sw.smile.nlp.gate.keyword.KeywordAnalyser init
INFO: ie.deri.sw.smile.nlp.gate.keyword.KeywordAnalyser: frequency lists available for following languages: en,it,pt,fr,es,
17-Oct-2008 15:29:04 ie.deri.sw.smile.nlp.gate.frequency.FrequencyAnalyser countTokens
INFO: finished counting, elapsed time: 48ms
17-Oct-2008 15:29:04 ie.deri.sw.smile.nlp.gate.keyword.KeywordAnalyser readFrequencyList
INFO: reference frequency list loaded in 97ms ..
17-Oct-2008 15:29:04 ie.deri.sw.smile.nlp.gate.keyword.LexicalStatisticsStrategy produceCandidateTerms
INFO: finished computing lexical relevance statistics in 222ms ..
17-Oct-2008 15:29:05 ie.deri.sw.smile.nlp.gate.keyword.ChunkStrategy produceCandidateTerms
INFO: finished extracting complex units in 213ms , 367 complex terms..
17-Oct-2008 15:29:05 ie.deri.sw.smile.nlp.gate.keyword.ChunkStrategy produceCandidateTerms
INFO: number of chunks skipped:18
17-Oct-2008 15:29:06 ie.deri.sw.smile.nlp.gate.keyword.ComplexTermAnalyser partition
INFO: COMPARISONS:7305 -- CLUSTERS: 42 -- ELAPSED TIME:1010ms
17-Oct-2008 15:29:06 ie.deri.sw.smile.nlp.gate.keyword.ComplexTermAnalyser analyse
INFO: finshed analysing partitioned complex term clusters. Number of candidates: 30. Elapsed time: 588ms
```

The additional feature names are as follows, and they can be accessed programmatically via API as well:

keywords: a sorted map of <keyphrase:confidence > pairs, where confidence is an double value between 0 and 1. High confidence keyphrases have a value closer to 1.

language: the language of the document

document_size: the number of words/tokens in the document

lexicon_size: the number of different lemmas in the document

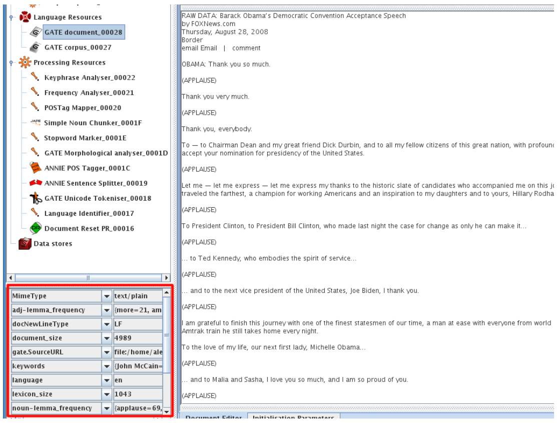
overall-token_frequency: the frequencies of the tokens in the document, implemented as a sorted map

overall-token_frequency: the frequencies of the tokens in the document, implemented as a sorted map

noun-lemma_frequency: the lemma frequencies of the nouns in the document, implemented as a sorted map

verb-lemma_frequency: the lemma frequencies of the verbs in the document, implemented as a sorted map

adj-lemma_frequency: the lemma frequencies of the adjectives in the document, implemented as a sorted map



MimeType	text/plain
adj-lemma_frequency	{more=21, am
docNewLineType	LF
document_size	4989
gate.SourceURL	file:/home/ale
keywords	{John McCain=
language	en
lexicon_size	1043
noun-lemma_frequency	{applause=69,

RAW DATA: Barack Obama's Democratic Convention Acceptance Speech
by FOXNews.com
Thursday, August 28, 2008
Border
email Email | comment

OBAMA: Thank you so much.

(APPLAUSE)

Thank you very much.

(APPLAUSE)

Thank you, everybody.

To — to Chairman Dean and my great friend Dick Durbin, and to all my fellow citizens of this great nation, with profound
accept your nomination for presidency of the United States.

(APPLAUSE)

Let me — let me express — let me express my thanks to the historic slate of candidates who accompanied me on this jo
traveled the farthest, a champion for working Americans and an inspiration to my daughters and to yours, Hillary Rodha

(APPLAUSE)

To President Clinton, to President Bill Clinton, who made last night the case for change as only he can make it...

(APPLAUSE)

... to Ted Kennedy, who embodies the spirit of service...

(APPLAUSE)

... and to the next vice president of the United States, Joe Biden, I thank you.

(APPLAUSE)

I am grateful to finish this journey with one of the finest statesmen of our time, a man at ease with everyone from world
Amtrak train he still takes home every night.

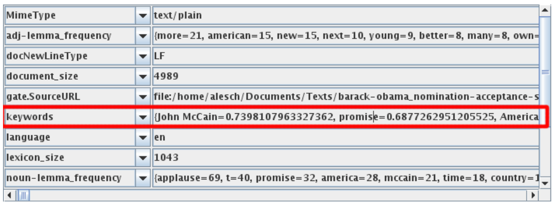
To the love of my life, our next first lady, Michelle Obama...

(APPLAUSE)

... and to Malia and Sasha, I love you so much, and I am so proud of you.

(APPLAUSE)

15. The results of the keyphrase extraction can be inspected by looking at the keywords feature of the document.



MimeType	text/plain
adj-lemma_frequency	{more=21, american=15, new=15, next=10, young=9, better=8, many=8, own=
docNewLineType	LF
document_size	4989
gate.SourceURL	file:/home/alesch/Documents/Texts/barack-obama_nomination-acceptance-s
keywords	{John McCain=0.7398107963327362, promise=0.6877262951205525, America
language	en
lexicon_size	1043
noun-lemma_frequency	{applause=69, t=40, promise=32, america=28, mccain=21, time=18, country=1