TermRaider

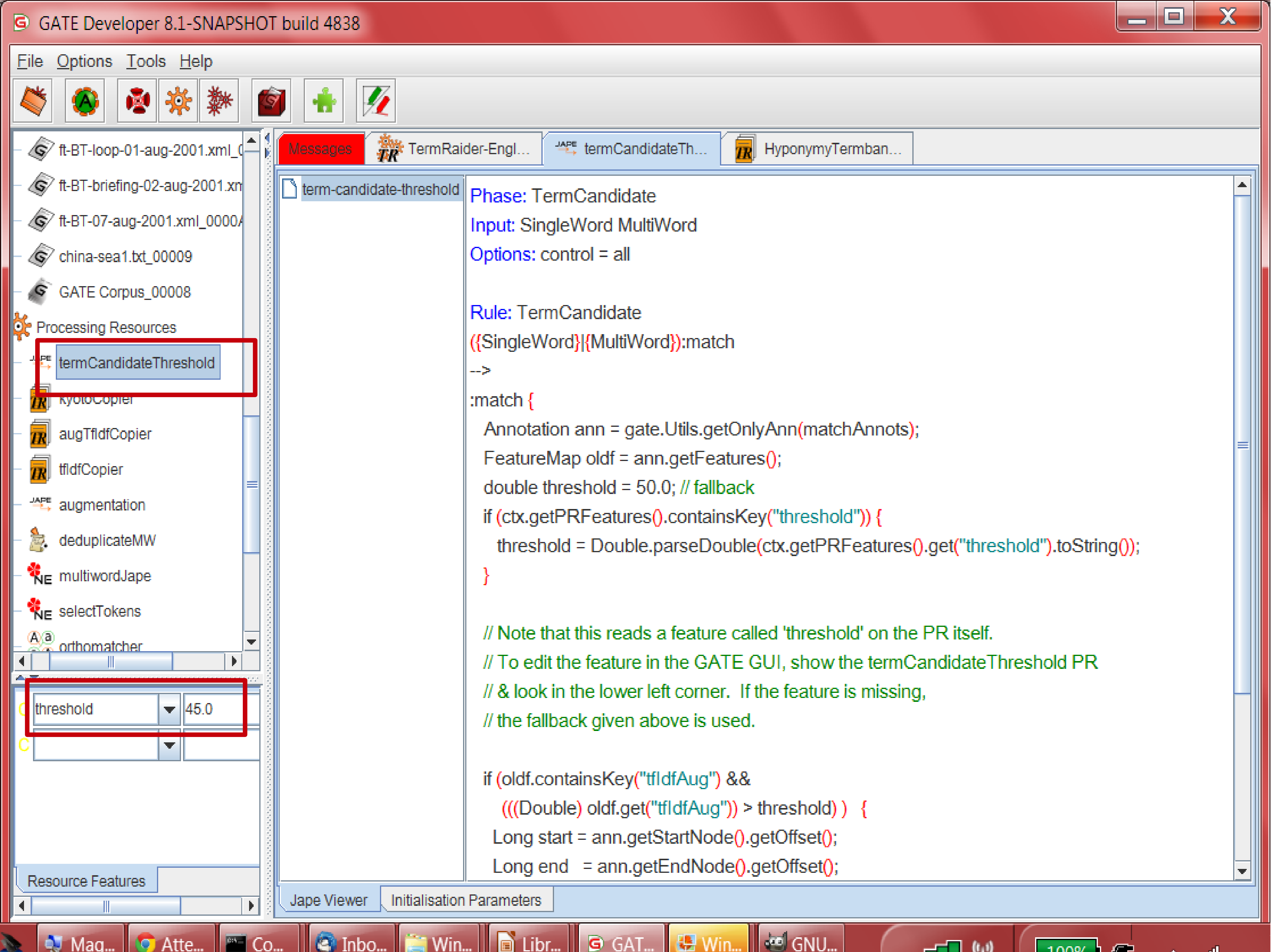
- GATE plugin for detecting single and multi-word terms
- Based originally on a simple web service - now extended to run in GATE, with visualisation tools, and extended functionality (new scoring systems, and an adaptation for German.
- Runs in GATE Developer (GUI) or on the command-line with RDF and CSV output
- Terms are ranked according to three possible scoring systems:
  - tf.idf = term frequency (nbr of times the term occurs in the corpus) divided by document frequency (nbr of documents in which the term occurs)
  - augmented tf.idf = after scoring tf.idf, the scores of hypernyms are boosted by the scores of hyponyms
  - Kyoto domain relevance = document frequency × (1 + nbr of hyponyms in the corpus), Bosma and Vossen 2010

TermRaider: Methodology

- After linguistic pre-processing (tokenisation, lemmatisation, POS tagging etc.), nouns and noun phrases are identified as initial term candidates
- Noun phrases include post-modifiers such as prepositional phrases, and are marked with head information for determining hyponymy. Nested nouns and noun phrases are all marked as candidates.
- Term candidates are then scored in 3 ways.
- The results can be viewed in the GATE GUI, exported as RDF according to the ARCOMEM data model, or saved as CSV files
- The viewer can be used to adjust the cutoff parameter. This is used to determine the score threshold for a term to be considered valid
- Terms can also be shown as a tag cloud

Deciding what is a term

- Because TermRaider ranks every possible candidate term, you probably don't want to use all candidate terms if you're annotating terms in a text
- We therefore provide a cutoff mechanism to select what score should determine whether something is a term or not
- The last PR in TermRaider is a JAPE grammar which takes a feature “threshold” and a value, by default set to 45, and annotates candidates as “Term” only if the value of the augmented tf.idf is above the threshold.



threshold 45.0

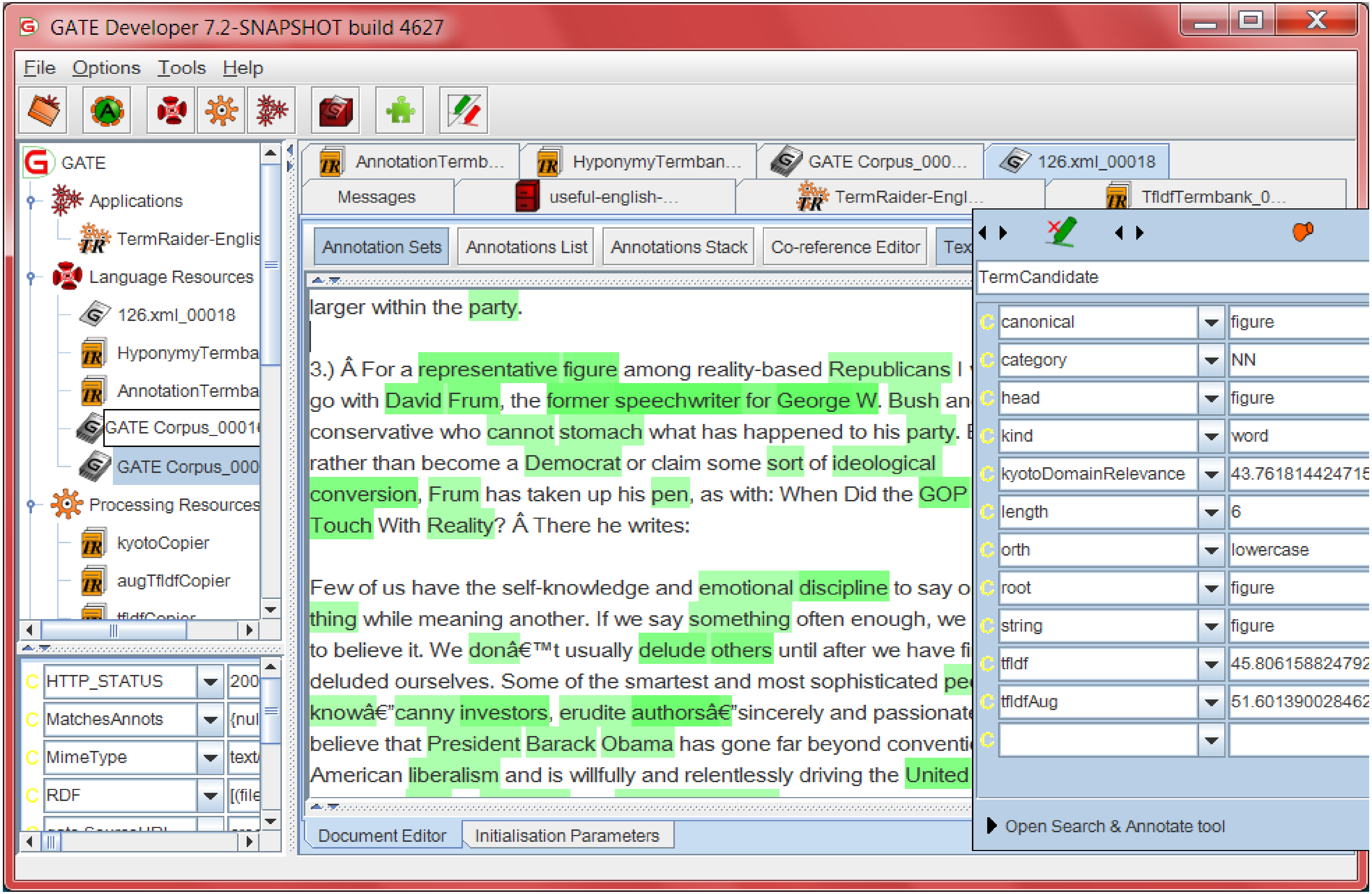
```
Phase: TermCandidate
Input: SingleWord MultiWord
Options: control = all

Rule: TermCandidate
({SingleWord}|{MultiWord}):match
-->
:match {
  Annotation ann = gate.Utils.getOnlyAnn(matchAnnots);
  FeatureMap oldf = ann.getFeatures();
  double threshold = 50.0; // fallback
  if (ctx.getPRFeatures().containsKey("threshold")) {
    threshold = Double.parseDouble(ctx.getPRFeatures().get("threshold").toString());
  }

  // Note that this reads a feature called 'threshold' on the PR itself.
  // To edit the feature in the GATE GUI, show the termCandidateThreshold PR
  // & look in the lower left corner.  If the feature is missing,
  // the fallback given above is used.

  if (oldf.containsKey("tfIdfAug") &&
     (((Double) oldf.get("tfIdfAug")) > threshold) )   {
   Long start = ann.getStartNode().getOffset();
   Long end   = ann.getEndNode().getOffset();
```

Term candidates in a document



larger within the party.

3.) Â For a representative figure among reality-based Republicans I w
go with David Frum, the former speechwriter for George W. Bush an
conservative who cannot stomach what has happened to his party. B
rather than become a Democrat or claim some sort of ideological
conversion, Frum has taken up his pen, as with: When Did the GOP
Touch With Reality? Â There he writes:

Few of us have the self-knowledge and emotional discipline to say o
thing while meaning another. If we say something often enough, we
to believe it. We donâ€™t usually delude others until after we have fi
deluded ourselves. Some of the smartest and most sophisticated pe
knowâ€"canny investors, erudite authorsâ€"sincerely and passionate
believe that President Barack Obama has gone far beyond conventi
American liberalism and is willfully and relentlessly driving the United

TermCandidate

canonical	figure
category	NN
head	figure
kind	word
kyotoDomainRelevance	43.761814424715
length	6
orth	lowercase
root	figure
string	figure
tfIdf	45.806158824792
tfIdfAug	51.601390028462

HTTP_STATUS	200
MatchesAnnots	{nul
MimeType	text
RDF	[(file

Try TermRaider in GATE

- Load the TermRaider plugin in GATE
- Load a corpus (around 20-100 documents on a similar topic is ideal, e.g. the news texts from the hands-on file that you have used previously in Module 1)
- Load TermRaider from the “Ready-made Applications” and run it on the corpus
- Inspect the results (click on “SingleWord”, “MultiWord” or “Candidate Term” in the document viewer)
- Try the Term Cloud viewer
- Change the threshold (open the termCandidateThreshold PR in GATE and then modify the value of “threshold” in the box in the bottom left corner). See what happens when you re-run the application.