Creating Stopword Lists for Historical Languages
with the Classical Language Toolkit

Patrick J. Burns
Institute for the Study of the Ancient World
Classical Language Toolkit

Florilegia: Big Textual Data Workshop / Universität Leipzig / 11.07.17

Classical Language Toolkit

What is the Classical Language Toolkit?

The Classical Language Toolkit (CLTK) is a free and open-source Python package that offers natural language processing (NLP) support for the languages of Ancient, Classical, and Medieval Eurasia.

Language-specific tokenizers, lemmatizers, POS-taggers, morphological parsers, etc. are available, under development, or in the feature-request list. Latin and Greek functionality are currently most complete.

Who is the Classical Language Toolkit?

- Open-source community collaborating at https://github.com/cltk
- Founded by Kyle P. Johnson, Classics PhD from NYU and NLP Research Scientist at Accenture
- Academic Advisors: Gregory Crane (Leipzig/Tufts) , Neil Coffee (Buffalo), Peter Meineck (NYU), Leonard Muellner (Brandeis/CHS)
- CLTK Archive developer: Luke Hollis

CLTK Goals

- *Low: Good analysis-friendly corpora/datasets for NLP of historical languages (Latin, Ancient/Classical Greek, Egyptian hieroglyphs, Hebrew, Sanskrit, Tibetan, Classical Chinese, etc.)*

CLTK Goals

- Low: Good analysis-friendly corpora/datasets for NLP of historical languages (Latin, Ancient/Classical Greek, Egyptian hieroglyphs, Hebrew, Sanskrit, Tibetan, Classical Chinese, etc.)
- *Medium: Collect & generate linguistic data for quantified classics*

CLTK Goals

- Low: Good analysis-friendly corpora/datasets for NLP of historical languages (Latin, Ancient/Classical Greek, Egyptian hieroglyphs, Hebrew, Sanskrit, Tibetan, Classical Chinese, etc.)
- Medium: Collect & generate linguistic data for quantified classics
- *High: Framework for an integrated study of the ancient world*

CLTK Stats

- Began 2014
- 1,808 commits at https://github.com/cltk/cltk
- 54 contributors
- 52 watchers, 238 stars, 165 forks
- 54 people, 20 teams
- 56 releases (with Zenodo DOI for every release)
- 85% code coverage
- Supports POSIX OS (and partially Windows)
- 2016 & 2017 Google Summer of Code participating organization

Stoplist Research

What is a stop list?

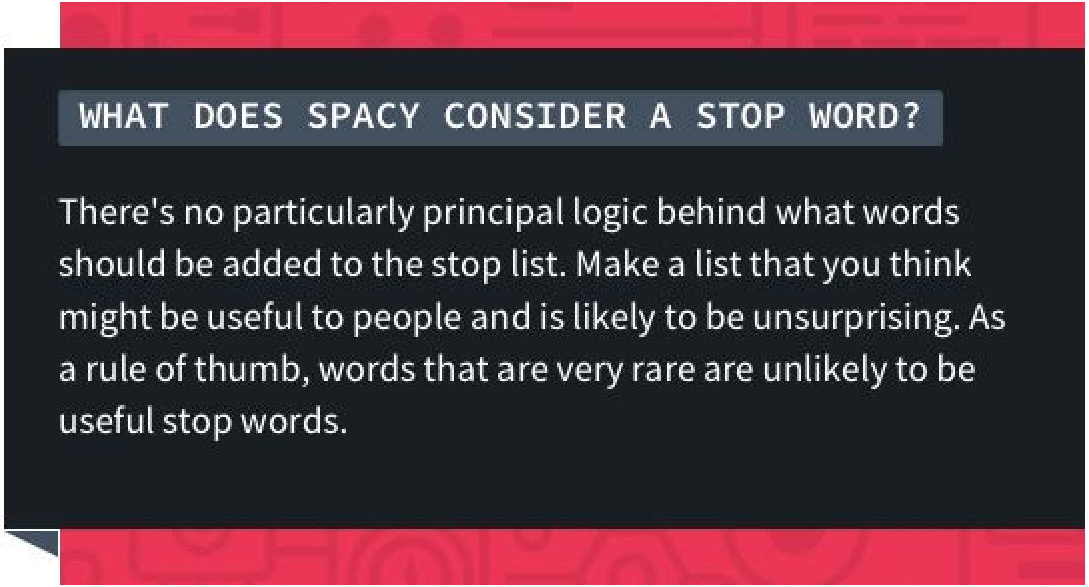
- Luhn 1957 on encoding documents:
    - “Words which fail to attain the status of major notions would not be entered.”
    - “dictionary of insignificant words”
- High frequency, low discrimination value, little semantic resolution
- Would literally be used to “stop” the physical printing of indexed abstracts of scientific papers; cf. Parkins 1963
- textual noise, “negative dictionary”
- “key parameter in the area of IR”; also now NLP, AI, text analysis, text/data mining, computational linguistics, etc.

What is a stop list?

“The general strategy for determining a stop list is to sort the terms by collection frequency, …and then take the most frequent terms, often hand-filtered for their semantic content relative to the domain of the documents being indexed, as a stop list, the members of which are then discarded during index.”—C. Manning

Manning, C., R. Prabhakar & Schütze, H. 2008. *Introduction to Information Retrieval*. Cambridge.

Real world considerations



WHAT DOES SPACY CONSIDER A STOP WORD?

There's no particularly principal logic behind what words should be added to the stop list. Make a list that you think might be useful to people and is likely to be unsurprising. As a rule of thumb, words that are very rare are unlikely to be useful stop words.

https://spacy.io/docs/usage/adding-languages

Acronymic considerations

- GIGO
- CACE

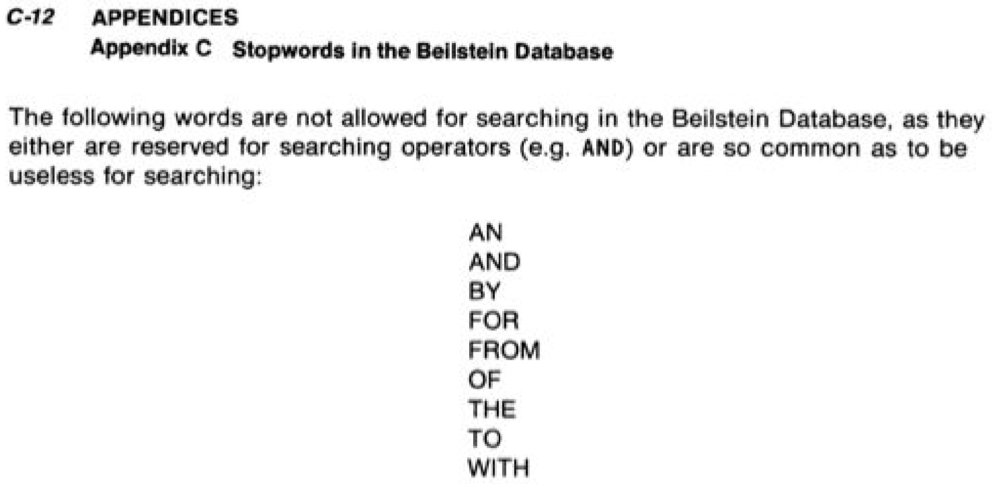
Flexible stoplist creation

- Two kinds of stoplists
    - Non-tailored (arbitrary, generic)
        - Based on corpus generalization
        - Static reference
    - Tailored (domain-specific)
        - Based on a specific corpus/dataset
        - Flexible resource

How long is a stop list?

- Number varies widely by application
- Number varies by language/script (cf. Saini & Rakholia 2016)
- Some examples
    - Fox 1989 (“A Stoplist for General Text) has 421 English words based on Brown Corpus
    - DIALOG stops total 9 words

“…useless for searching.”



C-12 APPENDICES

Appendix C Stopwords in the Beilstein Database

The following words are not allowed for searching in the Beilstein Database, as they either are reserved for searching operators (e.g. AND) or are so common as to be useless for searching:

AN
AND
BY
FOR
FROM
OF
THE
TO
WITH

Heller & Milne. 1991 *Online Searching in DIALOG*. Springer.

Zou et al. 2006 Aggregate Stoplist Model

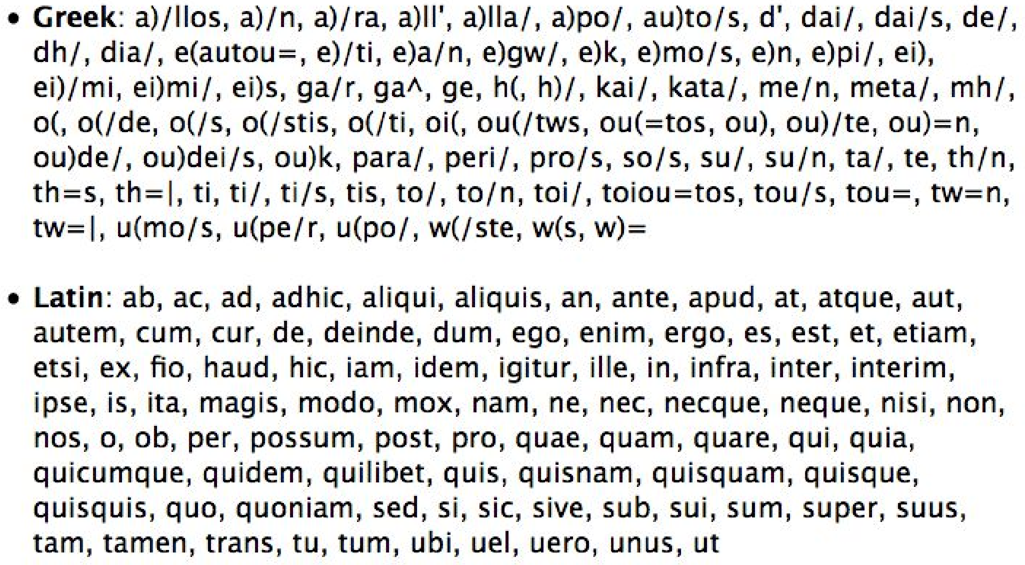
- Statistical Model
  - Frequency: Zipfian measures
  - Weighted frequency: Mean probability
  - Distribution: Variance probability
- Information Retrieval Model
  - Information value: Entropy
- Aggregate Model
  - Borda sort of above models

Zou, F., F. L. Wang, X. Deng, S. Han, and L. S. Wang. “Automatic Construction of Chinese Stop Word List.” In *Proceedings of the 5th WSEAS International Conference on Applied Computer Science*, 1010–15, 2006.

“Our stop word extraction algorithm is a promising technique, which saves the time for manual generation and constructs a standard. It could be applied to other languages in the future. ”—Zou et al.

Latin Stopword Lists

Perseus Greek and Latin stopword lists

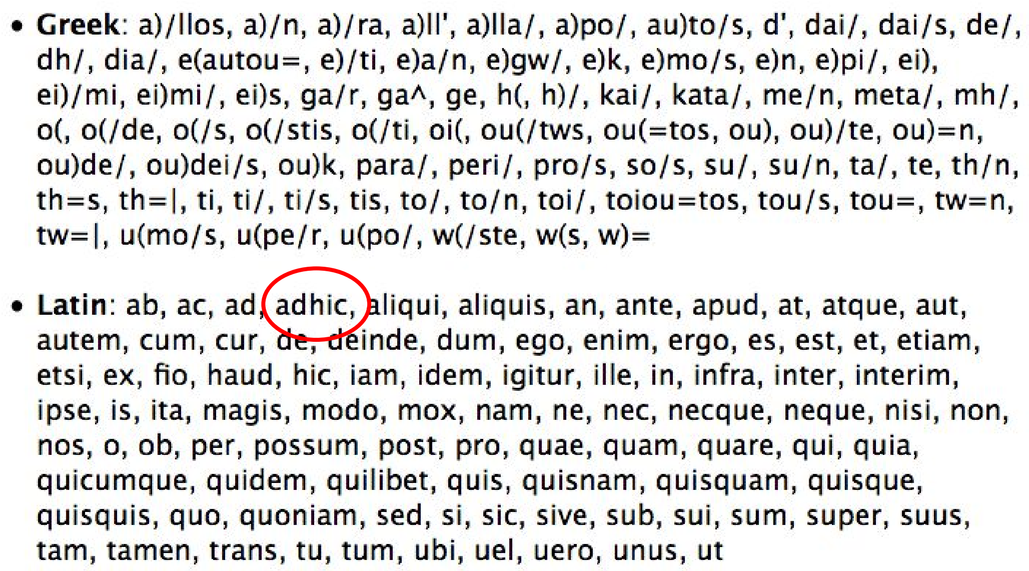


- **Greek**: a)/llos, a)/n, a)/ra, a)ll', a)lla/, a)po/, au)to/s, d', dai/, dai/s, de/, dh/, dia/, e(autou=, e)/ti, e)a/n, e)gw/, e)k, e)mo/s, e)n, e)pi/, ei), ei)/mi, ei)mi/, ei)s, ga/r, ga^, ge, h(, h)/, kai/, kata/, me/n, meta/, mh/, o(, o(/de, o(/s, o(/stis, o(/ti, oi(, ou(/tws, ou(=tos, ou), ou)/te, ou)=n, ou)de/, ou)dei/s, ou)k, para/, peri/, pro/s, so/s, su/, su/n, ta/, te, th/n, th=s, th=|, ti, ti/, ti/s, tis, to/, to/n, toi/, toiou=tos, tou/s, tou=, tw=n, tw=|, u(mo/s, u(pe/r, u(po/, w(/ste, w(s, w)=

- **Latin**: ab, ac, ad, adhic, aliqui, aliquis, an, ante, apud, at, atque, aut, autem, cum, cur, de, deinde, dum, ego, enim, ergo, es, est, et, etiam, etsi, ex, fio, haud, hic, iam, idem, igitur, ille, in, infra, inter, interim, ipse, is, ita, magis, modo, mox, nam, ne, nec, necque, neque, nisi, non, nos, o, ob, per, possum, post, pro, quae, quam, quare, qui, quia, quicumque, quidem, quilibet, quis, quisnam, quisquam, quisque, quisquis, quo, quoniam, sed, si, sic, sive, sub, sui, sum, super, suus, tam, tamen, trans, tu, tum, ubi, uel, uero, unus, ut

Perseus stopword lists also available for English, Italian, German, and French.

Perseus Greek and Latin stopword lists

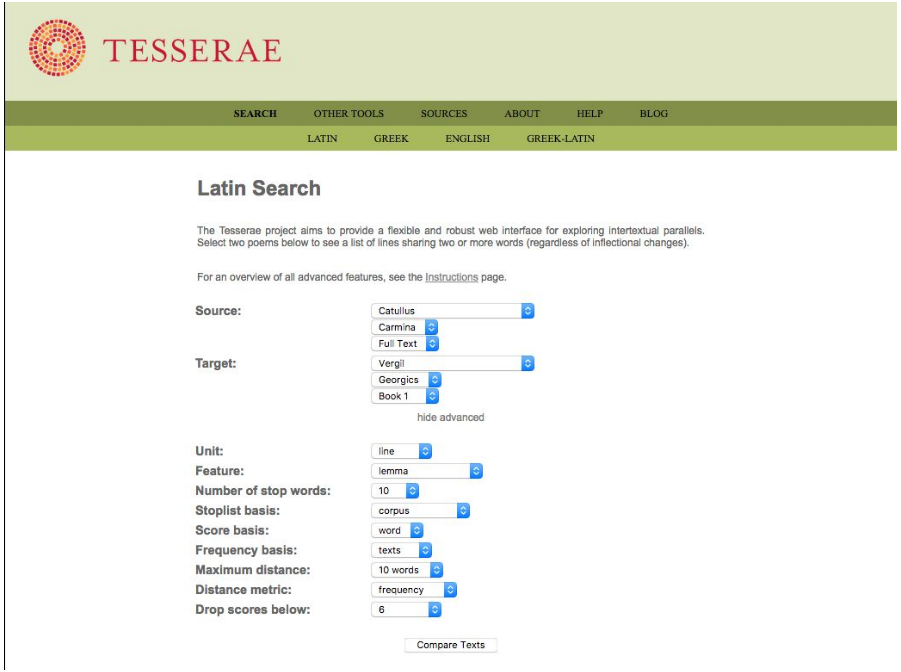


- **Greek**: a)/llos, a)/n, a)/ra, a)ll', a)lla/, a)po/, au)to/s, d', dai/, dai/s, de/, dh/, dia/, e(autou=, e)/ti, e)a/n, e)gw/, e)k, e)mo/s, e)n, e)pi/, ei), ei)/mi, ei)mi/, ei)s, ga/r, ga^, ge, h(, h)/, kai/, kata/, me/n, meta/, mh/, o(, o(/de, o(/s, o(/stis, o(/ti, oi(, ou(/tws, ou(=tos, ou), ou)/te, ou)=n, ou)de/, ou)dei/s, ou)k, para/, peri/, pro/s, so/s, su/, su/n, ta/, te, th/n, th=s, th=|, ti, ti/, ti/s, tis, to/, to/n, toi/, toiou=tos, tou/s, tou=, tw=n, tw=|, u(mo/s, u(pe/r, u(po/, w(/ste, w(s, w)=

- **Latin**: ab, ac, ad, adhic, aliqui, aliquis, an, ante, apud, at, atque, aut, autem, cum, cur, de, deinde, dum, ego, enim, ergo, es, est, et, etiam, etsi, ex, fio, haud, hic, iam, idem, igitur, ille, in, infra, inter, interim, ipse, is, ita, magis, modo, mox, nam, ne, nec, necque, neque, nisi, non, nos, o, ob, per, possum, post, pro, quae, quam, quare, qui, quia, quicumque, quidem, quilibet, quis, quisnam, quisquam, quisque, quisquis, quo, quoniam, sed, si, sic, sive, sub, sui, sum, super, suus, tam, tamen, trans, tu, tum, ubi, uel, uero, unus, ut

Perseus stopword lists also available for English, Italian, German, and French.

Stopwords in Tesserae Search



Latin Search

The Tesserae project aims to provide a flexible and robust web interface for exploring intertextual parallels. Select two poems below to see a list of lines sharing two or more words (regardless of inflectional changes).

For an overview of all advanced features, see the Instructions page.

Source: Catullus / Carmina / Full Text

Target: Vergil / Georgics / Book 1

hide advanced

Unit: line

Feature: lemma

Number of stop words: 10

Stoplist basis: corpus

Score basis: word

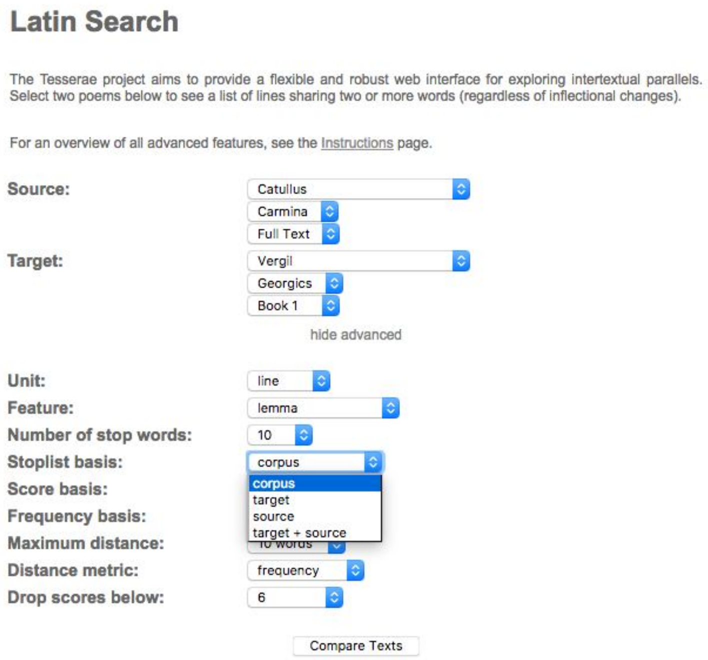
Frequency basis: texts

Maximum distance: 10 words

Distance metric: frequency

Drop scores below: 6

Stopwords in Tesserae Search



Latin Search

The Tesserae project aims to provide a flexible and robust web interface for exploring intertextual parallels. Select two poems below to see a list of lines sharing two or more words (regardless of inflectional changes).

For an overview of all advanced features, see the Instructions page.

Source: Catullus, Carmina, Full Text

Target: Vergil, Georgics, Book 1

Unit: line

Feature: lemma

Number of stop words: 10

Stoplist basis: corpus

Score basis:

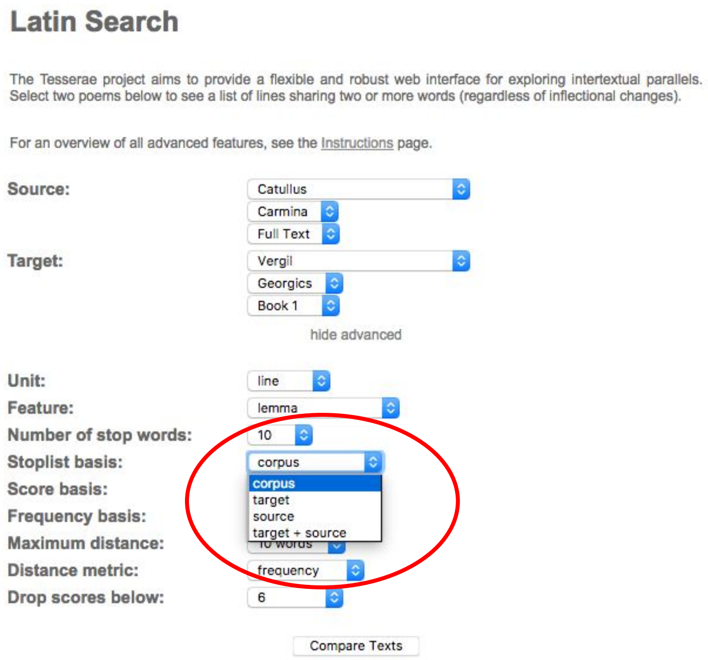
Frequency basis:

Maximum distance: 10 words

Distance metric: frequency

Drop scores below: 6

Stopwords in Tesserae Search



Latin Search

The Tesserae project aims to provide a flexible and robust web interface for exploring intertextual parallels. Select two poems below to see a list of lines sharing two or more words (regardless of inflectional changes).

For an overview of all advanced features, see the Instructions page.

Source: Catullus
Carmina
Full Text

Target: Vergil
Georgics
Book 1

hide advanced

Unit: line

Feature: lemma

Number of stop words: 10

Stoplist basis: corpus

Score basis:

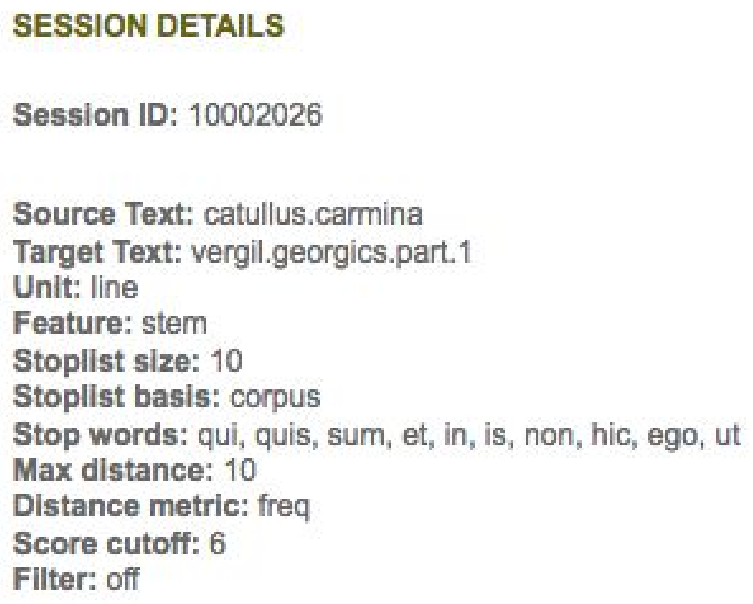
Frequency basis:

Maximum distance: 10 words

Distance metric: frequency

Drop scores below: 6

Stopwords in Tesserae Search

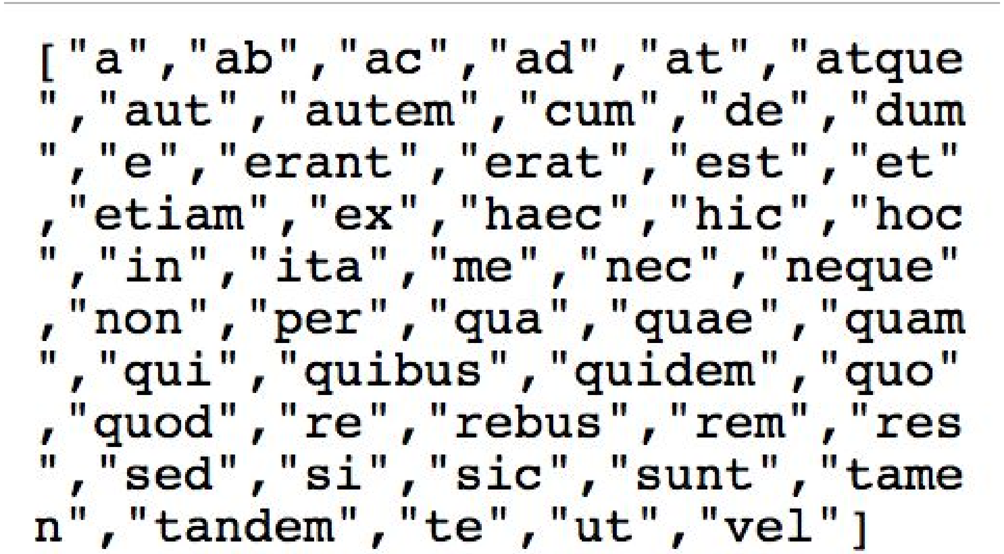


SESSION DETAILS

Session ID: 10002026

Source Text: catullus.carmina
Target Text: vergil.georgics.part.1
Unit: line
Feature: stem
Stoplist size: 10
Stoplist basis: corpus
Stop words: qui, quis, sum, et, in, is, non, hic, ego, ut
Max distance: 10
Distance metric: freq
Score cutoff: 6
Filter: off

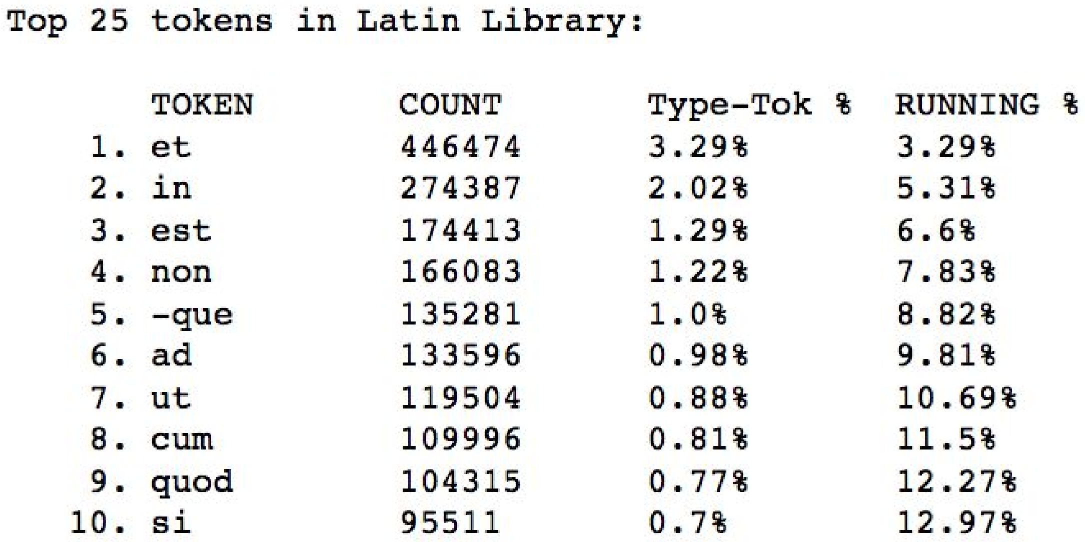
stopwords-json Latin stops



```
["a","ab","ac","ad","at","atque
","aut","autem","cum","de","dum
","e","erant","erat","est","et"
,"etiam","ex","haec","hic","hoc
","in","ita","me","nec","neque"
,"non","per","qua","quae","quam
","qui","quibus","quidem","quo"
,"quod","re","rebus","rem","res
","sed","si","sic","sunt","tame
n","tandem","te","ut","vel"]
```

https://github.com/6/stopwords-json/blob/master/dist/la.json

Top tokens in Latin Library

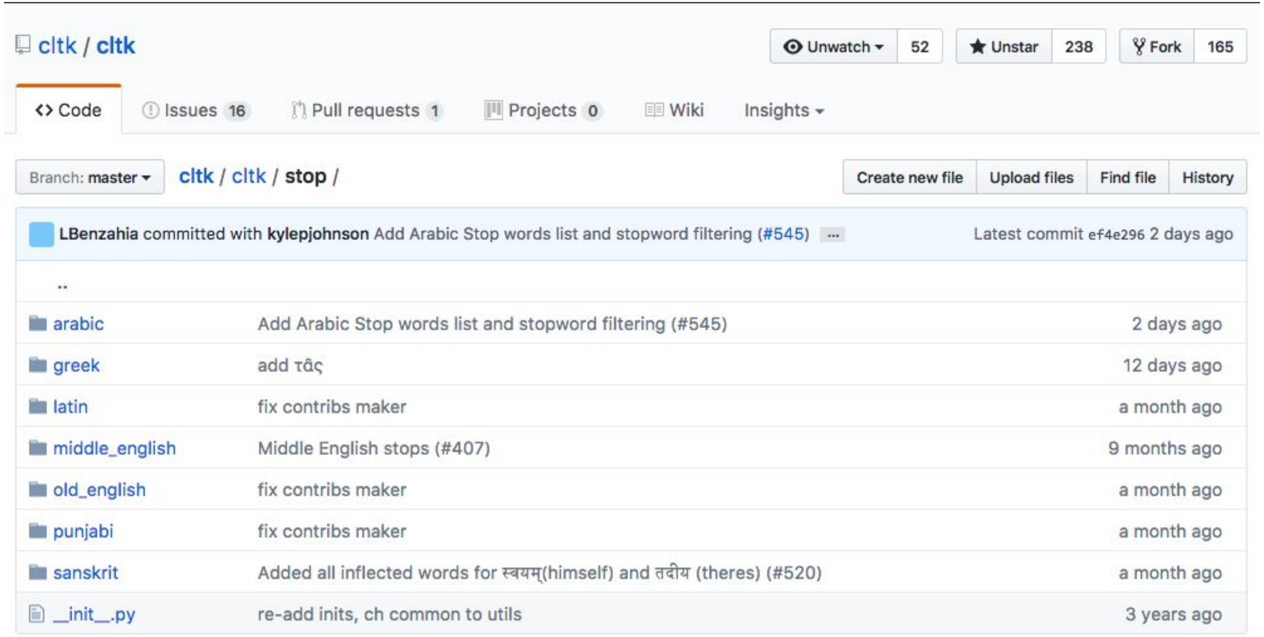


```
Top 25 tokens in Latin Library:

       TOKEN          COUNT          Type-Tok %   RUNNING %
    1. et             446474         3.29%        3.29%
    2. in             274387         2.02%        5.31%
    3. est            174413         1.29%        6.6%
    4. non            166083         1.22%        7.83%
    5. -que           135281         1.0%         8.82%
    6. ad             133596         0.98%        9.81%
    7. ut             119504         0.88%        10.69%
    8. cum            109996         0.81%        11.5%
    9. quod           104315         0.77%        12.27%
   10. si             95511          0.7%         12.97%
```

https://github.com/diyclassics/ll-experiments/blob/master/ll-10000.ipynb

Stopword Lists in CLTK



cltk / cltk / stop /

LBenzahia committed with kylepjohnson Add Arabic Stop words list and stopword filtering (#545) … Latest commit ef4e296 2 days ago

..		
arabic	Add Arabic Stop words list and stopword filtering (#545)	2 days ago
greek	add τᾶς	12 days ago
latin	fix contribs maker	a month ago
middle_english	Middle English stops (#407)	9 months ago
old_english	fix contribs maker	a month ago
punjabi	fix contribs maker	a month ago
sanskrit	Added all inflected words for स्वयम्(himself) and तदीय (theres) (#520)	a month ago
__init__.py	re-add inits, ch common to utils	3 years ago

Language coverage of current CLTK stop module

> “There is a growing awareness in the linguistic community that a non-trivial amount of *standardization* across resources is necessary…to be useful for all kinds of comparative linguistic research.”—D. Haug

Haug, D. “Standardizing Treebanks for Historical Indo-European Languages”

“Creating an infrastructure for…linguistic research means building **resources** by structuring data following a given **standard** with the help of **tools** both in the establishing phase and the later dissemination.”—D. Haug

Haug, D. “Standardizing Treebanks for Historical Indo-European Languages”

CLTK Stoplist WIP

- Jupyter Notebook with sample code/workflow for generating Latin stoplist:
    - https://github.com/diyclassics/stopwords

CLTK Latin Stoplist WIP: First Pass

- Data
    - CLTK Latin Library corpus
        - 2164 files
        - 13.1M tokens
        - 425701 unique tokens
        - 194347 hapaxes

CLTK Latin Stoplist WIP: First Pass

- 100 word stoplist by mean probability
  - ['et', 'in', 'est', 'non', 'ad', 'ut', 'cum', 'quod', 'qui', 'sed', 'si', 'de', 'quae', 'quam', 'per', 'ex', 'nec', 'sunt', 'esse', 'se', 'hoc', 'enim', 'ab', 'aut', 'autem', 'etiam', 'quid', 'te', 'atque', 'uel', 'eius', 'me', 'quo', 'sit', 'iam', 'quia', 'ne', 'haec', 'mihi', 'tamen', 'ac', 'tibi', 'nam', 'sic', 'ita', 'id', 'pro', 'eo', 'nunc', 'uero', 'neque', 'inter', 'quem', 'erat', 'ille', 'ergo', 'ipse', 'eum', 'quibus', 'quoque', 'sibi', 'ego', 'quidem', 'nisi', 'qua', 'omnia', 'hic', 'post', 'fuit', 'tu', 'nihil', 'ea', 'illa', 'his', 'omnes', 'nos', 'esset', 'modo', 'dum', 'sine', 'quis', 'ubi', 'sicut', 'ante', 'sub', 'tam', 'secundum', 'deus', 'potest', 'dei', 'nobis', 'quos', 'igitur', 'ei', 'omnibus', 'res', 'cui', 'sua', 'apud', 'eorum']

CLTK Latin Stoplist WIP: First Pass

- 100 word stoplist by variance probability
  - ['et', 'in', 'est', 'non', 'quod', 'ad', 'ut', 'cum', 'qui', 'de', 'si', 'sed', 'quae', 'per', 'ex', 'quam', 'esse', 'nec', 'te', 'sunt', 'autem', 'me', 'enim', 'se', 'dig', 'hoc', 'aut', 'ab', 'bibit', 'quid', 'uel', 'atque', 'mihi', 'eius', 'quaestio', 'pro', 'etiam', 'tibi', 'quia', 'sit', 'iam', 'secundum', 'quo', 'ac', 'ne', 'ergo', 'od', 'nihil', 'tu', 'haec', 'sic', 'id', 'nam', 'ego', 'neque', 'tamen', 'eum', 'deus', 'nunc', 'dei', 'ita', 'eo', 'uero', 'sicut', 'uos', 'hic', 'erat', 'nouus', 'fuit', 'nos', 'ille', 'inter', 'dum', 'quem', 'quoque', 'quidem', 'esset', 'bellum', 'ipse', 'sibi', 'nummus', 'anno', 'quibus', 'post', 'his', 'omnia', 'ea', 'super', 'qua', 'sub', 'illa', 'dominus', 'deo', 'rex', 'nisi', 'totus', 'dixit', 'dicitur', 'ed', 'ante']

CLTK Latin Stoplist WIP: First Pass

- 100 word stoplist by entropy
  - ['et', 'in', 'est', 'non', 'ad', 'ut', 'cum', 'quod', 'qui', 'sed', 'si', 'de', 'quae', 'quam', 'per', 'ex', 'nec', 'sunt', 'esse', 'se', 'hoc', 'ab', 'enim', 'aut', 'autem', 'etiam', 'quid', 'quo', 'atque', 'eius', 'te', 'uel', 'sit', 'me', 'iam', 'ne', 'haec', 'quia', 'tamen', 'nam', 'ac', 'mihi', 'ita', 'sic', 'tibi', 'id', 'pro', 'eo', 'inter', 'nunc', 'quem', 'ipse', 'uero', 'neque', 'quibus', 'ille', 'erat', 'eum', 'sibi', 'qua', 'nisi', 'quoque', 'ergo', 'quidem', 'omnia', 'post', 'hic', 'fuit', 'ego', 'ea', 'nihil', 'omnes', 'his', 'illa', 'modo', 'tu', 'esset', 'sine', 'nos', 'dum', 'ubi', 'ante', 'quis', 'tam', 'sub', 'sicut', 'quos', 'omnibus', 'potest', 'nobis', 'sua', 'cui', 'igitur', 'res', 'ei', 'tantum', 'cuius', 'apud', 'contra', 'magis']

CLTK Latin Stoplist WIP: First Pass

- 100-word aggregate stoplist by borda sort
    - ['et', 'in', 'est', 'non', 'ad', 'ut', 'quod', 'cum', 'qui', 'si', 'sed', 'de', 'quae', 'quam', 'per', 'ex', 'nec', 'esse', 'sunt', 'se', 'hoc', 'enim', 'autem', 'ab', 'aut', 'te', 'quid', 'uel', 'etiam', 'atque', 'me', 'eius', 'quo', 'sit', 'quia', 'iam', 'ne', 'ac', 'mihi', 'haec', 'tamen', 'tibi', 'pro', 'nam', 'id', 'ita', 'sic', 'eo', 'neque', 'uero', 'eum', 'nunc', 'inter', 'ergo', 'erat', 'quem', 'ipse', 'ego', 'quibus', 'nihil', 'ille', 'quoque', 'quidem', 'sibi', 'dig', 'nisi', 'qua', 'post', 'ea', 'tu', 'hic', 'fuit', 'omnia', 'his', 'esset', 'nos', 'sicut', 'illa', 'omnes', 'sine', 'secundum', 'bibit', 'modo', 'dum', 'quis', 'quaestio', 'ubi', 'deus', 'od', 'ante', 'dei', 'potest', 'tam', 'sub', 'ei', 'uos', 'nouus', 'quos', 'nobis', 'bellum’]

Future Directions

Plan to improve CLTK Stopword Lists

- Reverse current logic for `stop` module.
    - Add code for systematic creation of stop word lists.
    - Publish reference stop word lists for CLTK languages based on code
    - i.e. maintain balance of static reference and flexible resource

Plan to improve CLTK Stoplists

- Ensure that the stoplists are:
  - properly documented
  - version controlled
  - citable
- RAD paradigm (cf. Haswell 2005), that is…
  - replicable
  - aggregable
  - data-driven
- And as such will be better suited to be included in other text analysis projects.

Toward a Historical Language BLARK

- CLTK as basic language resource kit (Krauwer 2003)
  - "minimum general text corpus required to be able to do any precompetitive research for the language at all"
  - "collection of basic tools to manipulate and analyse the corpora"
  - "collection of skills that constitute the minimal starting point for the development of a competitive NL/Speech technology industry."

Krauwer, S. 2003. “The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap.” Proceedings of the 2003 International Workshop on Speech and Computer (SPECOM 2003) : 8-15.

CLTK’s BLARK in Progress

	Arabic	Akkadian	Bengali	Chinese	Classical Hindi	Coptic	Egyptian	Greek	Hebrew	Hindi	Javanese	Latin	Malayalam	Middle English	Old Church Slavonic	Old English	Old French	Old Norse	Pali	Persian	Prakrit	Punjabi	Sanskrit	Telugu	Tibetan	Urdu
Corpora	●		●	●	●	●	●	●	●	●	●	●	●		●	●	○	●	●	●	●	●	●	●	●	●
Stoplist	●							●				●		●		●	○					●	●			
Sentence Tokenizer								●				●					○									
Word Tokenizer	●											●					○						●			
Stemmer		●										●											●			
Lemmatizer								●				●														
POS-Tagger								●				●														
Prosody Tagger								●				●														
NER								●				●					○									

Proposing—a CLTK BLARK-a-thon

- What I'd like to see happen in the next year:
    - Pick a day/week(?)/month(?)
    - Choose a "basic" tool/resource/etc.
    - Get basic "basic" coverage for as many historical languages as possible
    - Repeat
- Stoplists seem like a good place to start
    - Only a few languages covered now
    - Have a theoretical basis and a systematic way of bootstrapping stoplists

Creating Stopword Lists for Historical Languages
with the Classical Language Toolkit

Patrick J. Burns

Institute for the Study of the Ancient World

Classical Language Toolkit

Florilegia: Big Textual Data Workshop / Universität Leipzig / 11.07.17