# Lexos

## Summary

Lexos is a suite of tools designed to facilitate the computational analysis of literary and historical texts. It offers an integrated workflow in which the pre-processing ("scrubbing"), analysis, and visualization steps can be accomplished in a single, web-based environment. Scrubbing features include handling punctuation, stop words, markup tags, and character consolidations, as well as document segmentation, culling, and n-gram tokenization. Analytical tools include basic document statistics, hierarchical and k-means cluster analysis, cosine similarity ranking, and z-score analysis. Visualizations include word and bubble clouds, comparative "multiclouds", and rolling window analysis. Analytical tools produce line, PCA, Voronoi cell, and dendrogram graphs. Each of the tools has export functionality.

Lexos is aimed at entry-level users as well as advanced scholars using small to medium-sized text corpora. It places particular emphasis on the processing of ancient and non-standard languages, as well as non-Western languages that do not use the Roman alphabet.

Lexos is produced by the Lexomics Research Group. An online version of Lexos v3.0 is available at http://lexos.wheatoncollege.edu/.

## Release history

This repo reflects ongoing development since our Summer 2018: Lexos v3.2.

### New Features in v3.2

* Lexos now uses Plotly on many pages for better interaction in graphs.
* Bootstrap modals are now used consistently for all error messages, and error messages have been improved for greater clarity.
* New video introductions have been embedded for the Analyze tools.
* The Statistics page layout has been re-designed with a new Plotly box plot graph.
* The Hierarchical Clustering tool now uses Plotly for plotting dendrograms.
* The K-Means Clustering tool now uses Plotly for Voronoi cell and 2D scatter plots. A new 3D scatter plot has been added.
* The Topword tool has an improved interface for showing the user the existing document classes.
* The Rolling Window Analysis tool now uses Plotly graphs. Users now can add multiple milestones. Also the download result button is fixed.

### New Beta Tools in v3.2

* Bootstrap Consensus Trees provides a measure of the stability of cluster analyses, as discussed by M. Eder, "Computational stylistics and biblical translation: how reliable can a dendrogram be?" In T. Piotrowski and Ł. Grabowski, editors, The Translator and the Computer, pages 155–170. WSF Press, Wrocław, 2012.
* Content Analysis provides a method of comparing the presenence of terms in documents according to user defined criteria. The tool can be used for applications as divers as opinion mining, determining organizational hardiness in stock broker reports, and sentiment analysis.

### Removed Features in v3.2

* The Grey word feature has been removed.
* The "topic clouds" feature in the Multicloud tool, which can be used to analyze data from MALLET-produced topic models, has been temporarily removed. We hope to re-introduce it in the next release.

## Installation

Installation instructions for Lexos v3.2 are available in the project Wiki.

* Windows Install Guide
* macOS Install Guide
* Linux Install Guide

### System Architecture (in brief)

Lexos v3.2 is written in Python 3.6 (as distributed in Anaconda 5.2) using theFlask microframework, based on Werkzeug and Jinja2.

The front end is designed using jQuery and the Bootstrap 3 framework, with a few functions derived from jQuery UI and DataTables. We increasingly incorporate the wiz fromD3.js and the Plotly Python graphing library in our visualizations and the power in thescikit-learn modules for text and statistical processing.

The directions for setting up the development environment for testing (using localhost:5000) on your local machine are stored in the 0\_InstallGuides directory.

## Dependencies

Lexos requires the following Python packages:

biopython, chardet, colorlover, flask, gensim, matplotlib, natsort, numpy, pandas, pip, plotly, scikit-bio, scikit-learn, scipy, pip, requests

On MacOS, the PDF Viewer extension needs to be enabled in the Chrome browser.

On Windows, the scikit-bio package requires Microsoft Visual C++ 14.0.

Lexos works on Chrome and Firefox. Other browsers are not supported, and some features may not function.

## License Information

See the file LICENSE for information on the Terms & Conditions for usage and a DISCLAIMER OF ALL WARRANTIES.

## Citation Information

Kleinman, S., LeBlanc, M.D., Drout, M., and Feng, W. (2018). *Lexos*. v3.2. https://github.com/WheatonCS/Lexos/. doi:10.5281/zenodo.1215821.