# Introducing the AVOBMAT

## (Analysis and Visualization of Bibliographic Metadata and Texts)

# Multilingual Research Tool

### Objective

This multilingual web-based digital toolkit enables researchers to perform critical and interactive analysis of bibliographic metadata and texts with data-driven and Natural Language Processing methods.

### Close reading

Focus on the specific details, complexities and nuances of a text.

+

### Distant reading

Explore the great unread: unveil and analyse repeated patterns, hidden connections, trends, themes and parallels in large quantity of texts.

=

### AVOBMAT

The combination of close and distant reading.

## Workflow & features

### 1. Upload corpus

- Import large digital collections including entire library databases
- Supported import formats include CSV (with links to full texts), RDF, EP3 XML
- Live Google Drive import

### 2. Clean your corpus

- Replace texts
- Context filter
- Remove (built-in) stopwords and punctuations in 52 languages
- Add stopwords and punctuation lists to be removed
- Remove numbers and non-alphabetical tokens
- Use of regular expressions
- Check the cleaned texts in the search results

### 3. Configure analytical tools

Change the default configuration of each tool according to your specifications including:

- Lemmatization in 20 languages
- Set N-Gram length
- Set window length for lexical diversity analysis

### 4. Select content

Search and refine your corpus:

- Faceted search
- Date (range) search
- Advanced search
- Fuzzy & proximity search
- Boolean search
- Command line search (Lucene syntax)

### 5. Analyse metadata & texts

Customizable configuration of text and data mining methods and visualization tools.

#### 5.1. Metadata analysis & interactive visualizations

- 106 metadata fields
- Metadata enrichment: automatic language detection & gender analysis of authors
- Analyse and visualize the bibliographic data chronologically in line and area charts in normalized and aggregated formats
- Create an interactive network analysis of maximum three metadata fields
- Make pie, horizontal and vertical bar charts of the bibliographic data of your choice
- Choose the metadata field(s) and the number of top items for visualization

#### 5.2. Content analysis

**Frequency analysis & word clouds** (Significant text, Tagsphere, Word count)

**Significant text analysis**

This tool for comparing corpora highlights the most related terms to a special query. Choose from 4 metric types such as chi square and set the maximum number of words and sample size.

**Tagsphere** (context of a word)

Creates tag clouds showing the co-occurring words of a specified search term in a corresponding word distance.

Besides the search term, specify the minimum frequency and the maximum distance of the co-occurring words.

Set the co-occurring words only appearing before or after the provided search term.

**Bar charts**

The results of the frequency analysis are also displayed in bar charts that the users can export.

**Keyword in context (KWIC)**

Read the context of your search terms being highlighted. Set up the length of the context and the number of documents to be displayed.

**Lexical diversity analysis**

Uses 8 metrics such as MTLD and HDD to calculate the lexical richness of texts. Set the window lengths for MSTTR and MATTR.

**N-gram viewer**

Shows the yearly count of the specified N-grams in aggregated and normalized views.

**Topic modeling**:

Cluster documents in semantic groups and find hidden semantic information.

Set the minimum frequency of words, number of topics, iterations, alpha and beta hyperparameters. Interactive removal of stopwords.

Topic documents, topic correlations and interactive time series.

**Named Entity Recognition (NER)**

It recognizes and extracts, among others, proper, common nouns and numbers from documents in **16 languages**.

The entity types include persons, places and organizations. Check the coloured predictions of the entities in the search results.

The NER output is also displayed in two different lists:

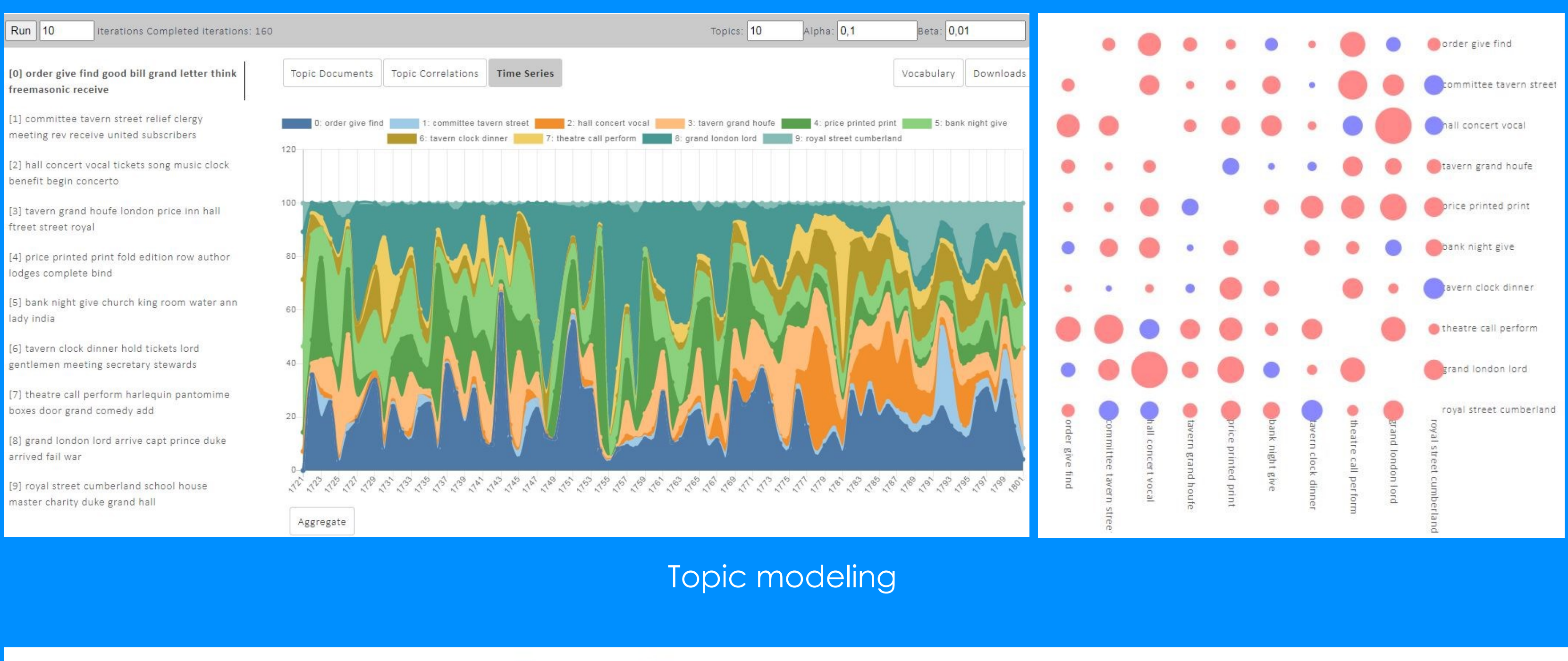
- Entities in all documents showing the list of top entities by type in all documents
- Entities by documents showing the number of entities found in specific documents

### 6. Export results and settings

All the statistics and visualizations can be downloaded in CSV and PNG formats.

For reasons of reproducibility and transparency, preprocessing parameters and configuration settings can be exported and imported.

## Examples



[0] order give find good bill grand letter think freemasonic receive

[1] committee tavern street relief clergy meeting rev receive united subscribers

[2] hall concert vocal tickets song music clock benefit begin concerto

[3] tavern grand houfe london price inn hall ftreet street royal

[4] price printed print fold edition row author lodges complete bind

[5] bank night give church king room water ann lady india

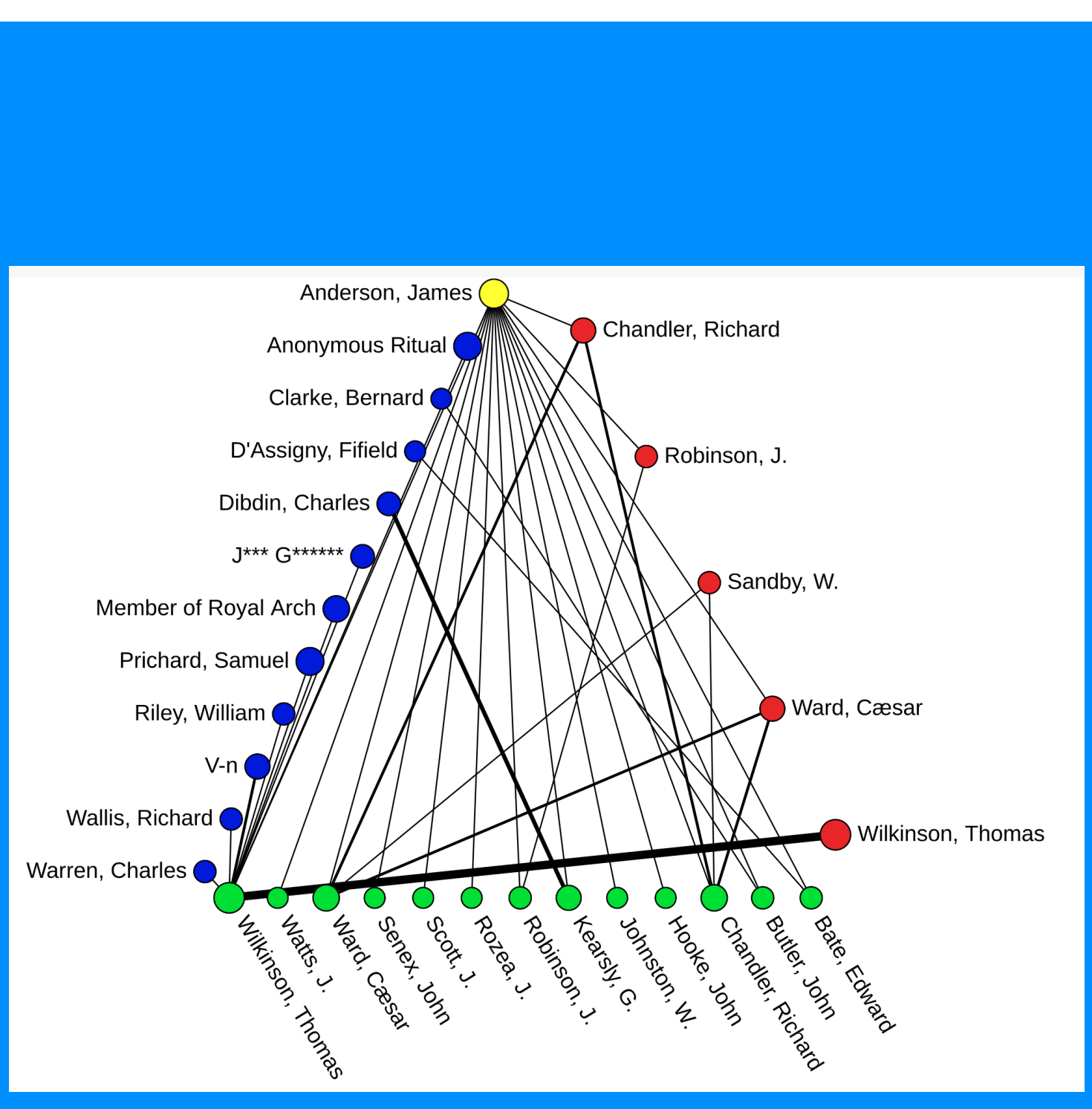
[6] tavern clock dinner hold tickets lord gentlemen meeting secretary stewards

[7] theatre call perform harlequin pantomime boxes door grand comedy add

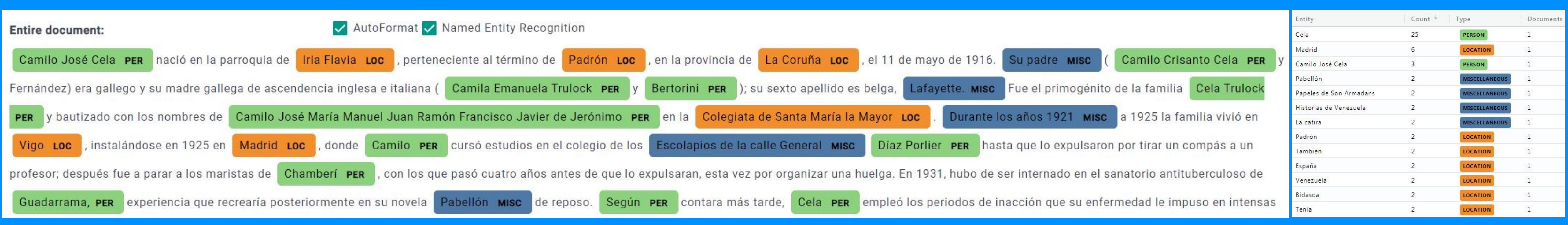
[8] grand london lord arrive capt prince duke arrived fail war

[9] royal street cumberland school house master charity duke grand hall

Topic modeling



Network analysis

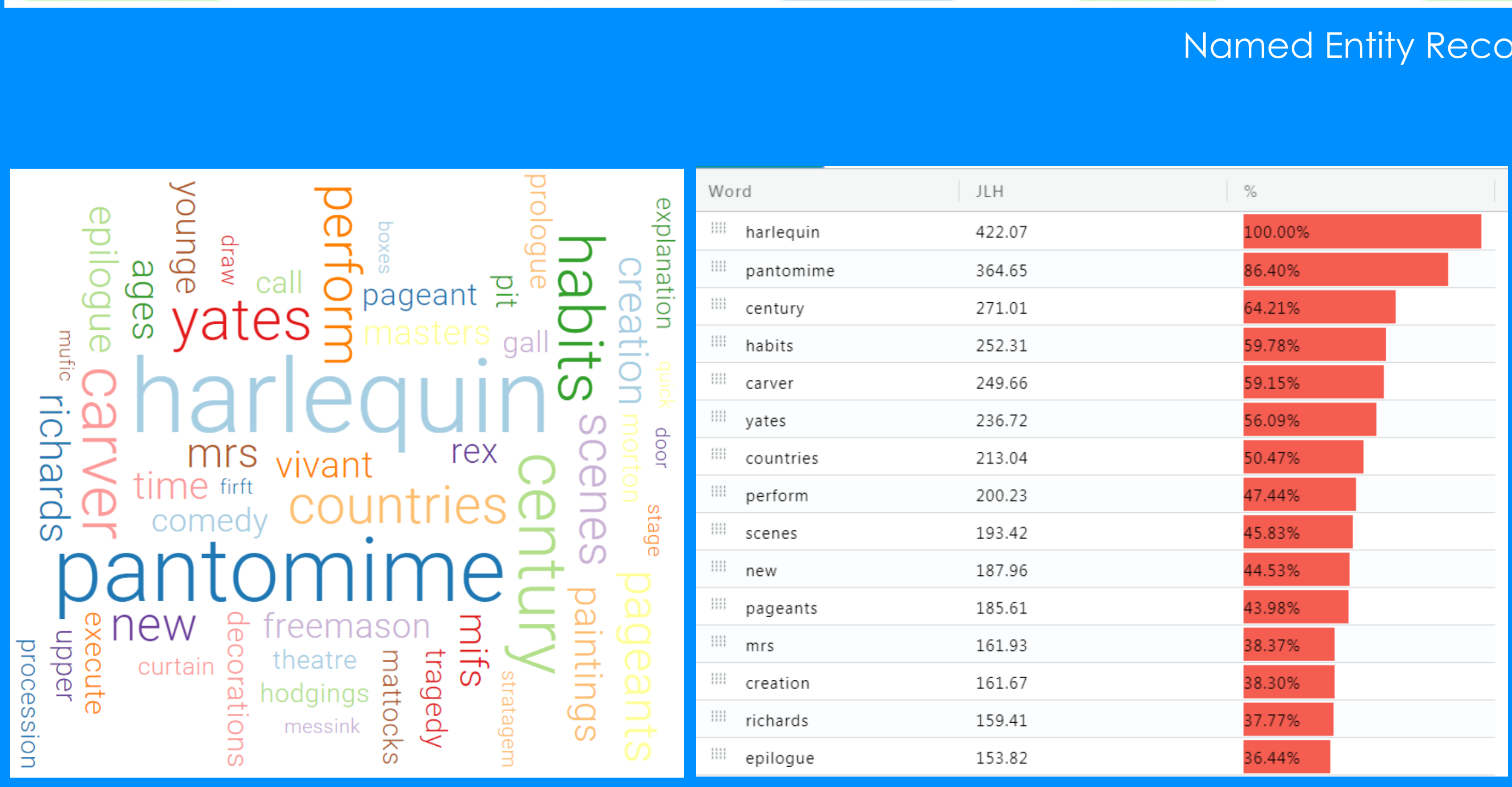


**Entire document:** ☑ AutoFormat ☑ Named Entity Recognition

Camilo José Cela PER nació en la parroquia de Iria Flavia LOC , perteneciente al término de Padrón LOC , en la provincia de La Coruña LOC , el 11 de mayo de 1916. Su padre MISC ( Camilo Crisanto Cela PER y Fernández) era gallego y su madre gallega de ascendencia inglesa e italiana ( Camila Emanuela Trulock PER y Bertorini PER ); su sexto apellido es belga, Lafayette. MISC Fue el primogénito de la familia Cela Trulock PER y bautizado con los nombres de Camilo José María Manuel Juan Ramón Francisco Javier de Jerónimo PER en la Colegiata de Santa María la Mayor LOC . Durante los años 1921 MISC a 1925 la familia vivió en Vigo LOC , instalándose en 1925 en Madrid LOC , donde Camilo PER cursó estudios en el colegio de los Escolapios de la calle General MISC Díaz Porlier PER hasta que lo expulsaron por tirar un compás a un profesor; después fue a parar a los maristas de Chamberí PER , con los que pasó cuatro años antes de que lo expulsaran, esta vez por organizar una huelga. En 1931, hubo de ser internado en el sanatorio antituberculoso de Guadarrama, PER experiencia que recrearía posteriormente en su novela Pabellón MISC de reposo. Según PER contara más tarde, Cela PER empleó los periodos de inacción que su enfermedad le impuso en intensas

| Entity | Count | Type | Documents |
|---|---|---|---|
| Cela | 25 | PERSON | 1 |
| Madrid | 6 | LOCATION | 1 |
| Camilo José Cela | 3 | PERSON | 1 |
| Pabellón | 2 | MISCELLANEOUS | 1 |
| Papeles de Son Armadans | 2 | MISCELLANEOUS | 1 |
| Historias de Venezuela | 2 | MISCELLANEOUS | 1 |
| La catira | 2 | MISCELLANEOUS | 1 |
| Padrón | 2 | LOCATION | 1 |
| También | 2 | LOCATION | 1 |
| España | 2 | LOCATION | 1 |
| Venezuela | 2 | LOCATION | 1 |
| Bidasoa | 2 | LOCATION | 1 |
| Tenía | 2 | LOCATION | 1 |

Named Entity Recognition



| Word | JLH | % |
|---|---|---|
| harlequin | 422.07 | 100.00% |
| pantomime | 364.65 | 86.40% |
| century | 271.01 | 64.21% |
| habits | 252.31 | 59.78% |
| carver | 249.66 | 59.15% |
| yates | 236.72 | 56.09% |
| countries | 213.04 | 50.47% |
| perform | 200.23 | 47.44% |
| scenes | 193.42 | 45.83% |
| new | 187.96 | 44.53% |
| pageants | 185.61 | 43.98% |
| mrs | 161.93 | 38.37% |
| creation | 161.67 | 38.30% |
| richards | 159.41 | 37.77% |
| epilogue | 153.82 | 36.44% |

Significant text analysis

### Technical specifications

elasticsearch

spaCy

LemmaGen Multilingual Open Source Lemmatisation

python

### What can AVOBMAT offer? How can it help your research?

- Explore your large digital collections in innovative & interactive ways with customizable preprocessing, analysis & visualization tools
- Discover new insights, unveil overlooked connections, themes, trends & patterns
- Critically analyse & interpret texts, (meta)data & visualizations
- Identify missing values, biases and errors in your databases at scale (e.g. selection, metadata, classification) to make more informed decisions about your research (questions)
- Discover novel type of evidence & test old hypotheses

### Try it out!

Try the limited beta version of AVOBMAT with a COVID-19 dataset of scholarly articles at avobmat.hu

Used in 44 countries.

Please note that at the moment AVOBMAT is hosted on a virtual machine with basic parameters.

https://www.avobmat.hu/

Zsolt Szántó, József Seres, Vilmos Bilicki, Bendegúz M. Bendicsek, Gábor Berend, Róbert Péter

Poster designed by Miki Csiky.

If you have any questions or want to use AVOBMAT in your library or project, please contact: Róbert Péter (University of Szeged) at avobmat@gmail.com

If you would like to be notified about the development and release of AVOBMAT, sign up here: https://avobmat.hu/contact/