

# Analysing and enriching bibliodata with AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts)

Róbert Péter

University of Szeged, Hungary

DARIAH Annual Event Workshop

25 November, 2020



# Objective

**Critical** and **interactive** analysis of bibliographic data AND texts with **data-driven** and **NLP** methods in a number of **languages**

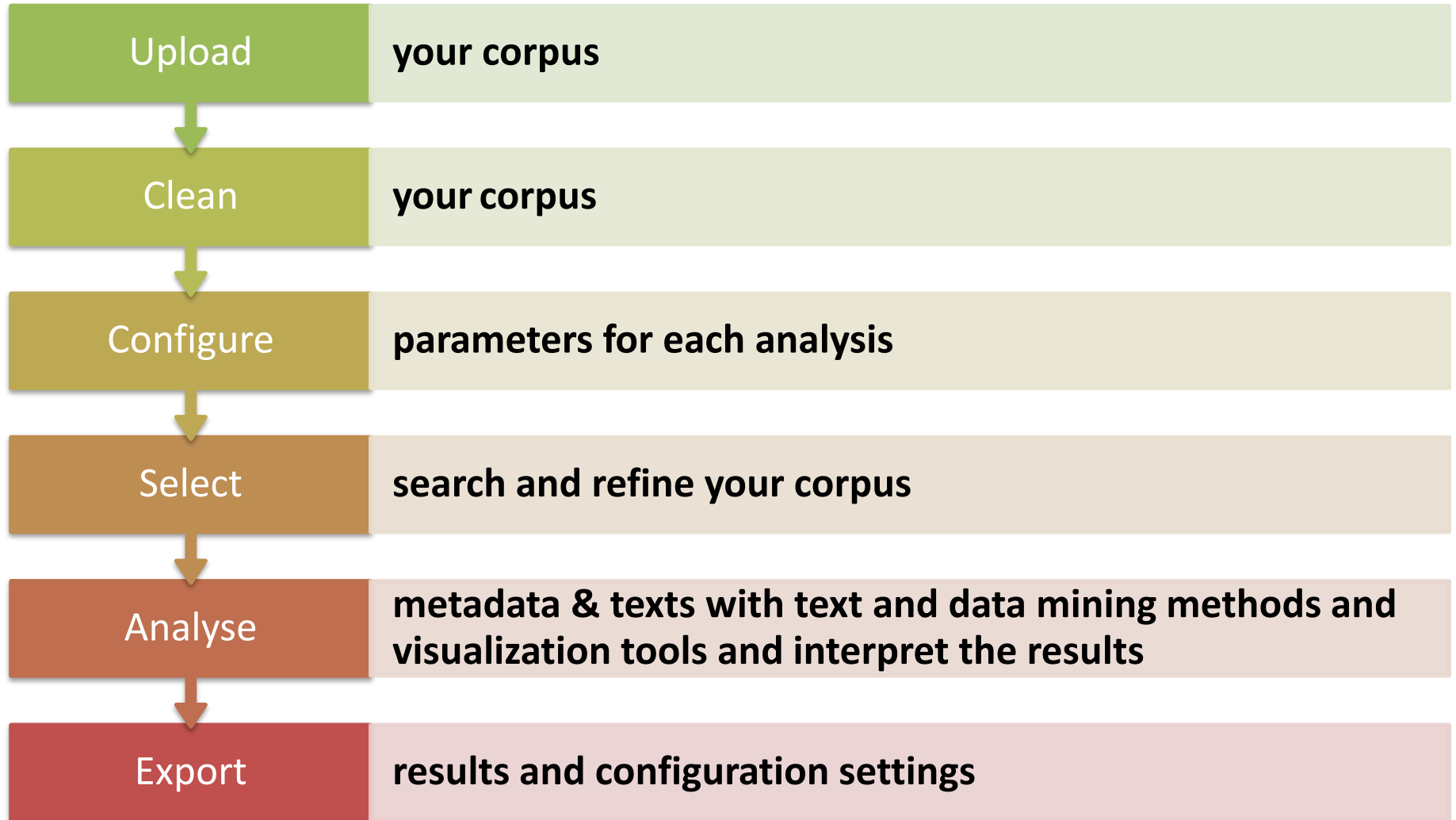


**Text and data mining  
large library & research  
databases**


# Theoretical and methodological background: the **combination** of **close** and **distant** reading



# Workflow



# AVOBMAT interface

**Databases:** 

- acta\_2020\_txt\_gender
- adop1\_zip
- aqc\_demo\_5
- c10-testz-adoption-rituals-ner\_a01\_zip
- c10-testz-adoption-rituals-ner\_a01\_zip

[Show more](#)

**Pick a date or range:**

- On
- Before
- After
- Between

[Search](#)

**Publication Year :**

- 2019 (1)
- 2018 (1)
- 2017 (2)

**Advanced search**

|   |              |                          |                              |                                     |
|---|--------------|--------------------------|------------------------------|-------------------------------------|
| Field:  | Search term: | Fuzzy: <a href="#">?</a> | Proximity: <a href="#">?</a> | Order: <a href="#">?</a>            |
| Authors   | Vajda        | 1                        | 1                            | <input checked="" type="checkbox"/> |
| <input checked="" type="radio"/> and <input type="radio"/> or <input type="radio"/> not |              |                          |                              |                                     |
| Field:  | Search term: | Fuzzy:                   | Proximity:                   | Order:                              |
| Entire document   | American     | 0                        | 1                            | <input checked="" type="checkbox"/> |

[+](#)

[Search](#) [Clear All](#)

**Commandline search (Lucene query):** [?](#)

YR:[2017 TO 2020] AND (FT:chloroquine OR FT:ivermectin) AND AB:coronavirus\*

[Search](#)

**Sort by:**

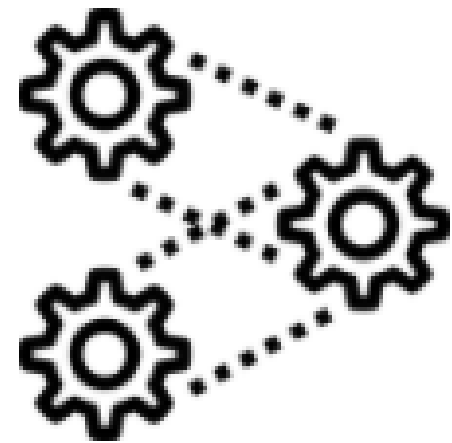
Publication date ascending



**Fast advanced, fuzzy, proximity and commandline searches**



# Preprocessing



|  |
|--|
| Language   |
| Replace text <b>61 available languages</b>             |
| Context filter <b>Regular expressions</b>              |
| Supplement/replace forenames (author gender detection) |
| Analyser configuration                                 |
| NGram viewer   |

|  |
|--|
| Topic modelling  |
| <input checked="" type="checkbox"/> Enable Topic modelling |
| <input checked="" type="checkbox"/> Lowercase              |
| <input checked="" type="checkbox"/> Remove numbers         |
| <input checked="" type="checkbox"/> Remove stopwords       |
| Additional stopword list                                   |



**Individual configuration of all analytical tools**

**Lemmatization in 29 languages**

|  |
|--|
| <input checked="" type="checkbox"/> Lemmatize                      |
| <input checked="" type="checkbox"/> Remove non-alphabetical tokens |
| <input checked="" type="checkbox"/> Remove punctuations            |
| Additional punctuation list  |

|   |
|---|
| Wordcloud - Significant text                              |
| Wordcloud - TagSpheres                                    |
| Lexical diversity   |
| Named Entity Recognition <b>NER in 16 languages</b>       |
| <b>Export Configuration</b>   <b>Import Configuration</b> |



**Reproducibility (import/export)**

# How can AVOBMAT foster critical analysis of bibliodata?

- Identify **biases** and **errors** in the databases
- **Missing values**

## Authors:

- MISSING\_VALUE (2991)
- Miklós, Kedves (241)
- Jenő, Koltay-Kastner (166)
- László, Gulyás (143)
- László, Vass (131)

[Show more](#)

## Publication Title:

- Módszertani közlemények (3781)
- Acta scientiarum mathematicarum (3247)
- Aetas (2069)

# Metadata enrichment

- Use of full texts



Automatic language  
detection



Gender analysis of  
authors

- Add your own list of **female / male names** in the Preprocessing phase



# Identify inaccurate metadata by language detection

Kurdy Fehér János:

Oleskeluharjoituksia : [vers]

Oleskeluharjoituksia

Kurdy Fehér, János and Drenko, Dénes: *Oleskeluharjoituksia : [vers]*. In: Gondolat-jel 1-2. p. 22. (1993)



Cikk, tanulmány, mű  
gondolatjel\_1993\_1\_2\_022.pdf  
[Download \(46kB\)](#) | [Preview](#)

Huusin  
sitten katsoin ympärilleni,  
missähän olen?

Seisoin erään vahvan miehen  
läheisyydessä, hän esitti minulle  
pari lentoharjoitusta.  
Hän joi ilmaa hyvin kovasti  
ja hyppeli ylöspäin  
torin kiveyksestä.

**Item Type:** Article

**Heading title:** metaszínház; Mi az, ami camp és mi az, ami nem?

**Journal or Publication Title:** Gondolat-jel

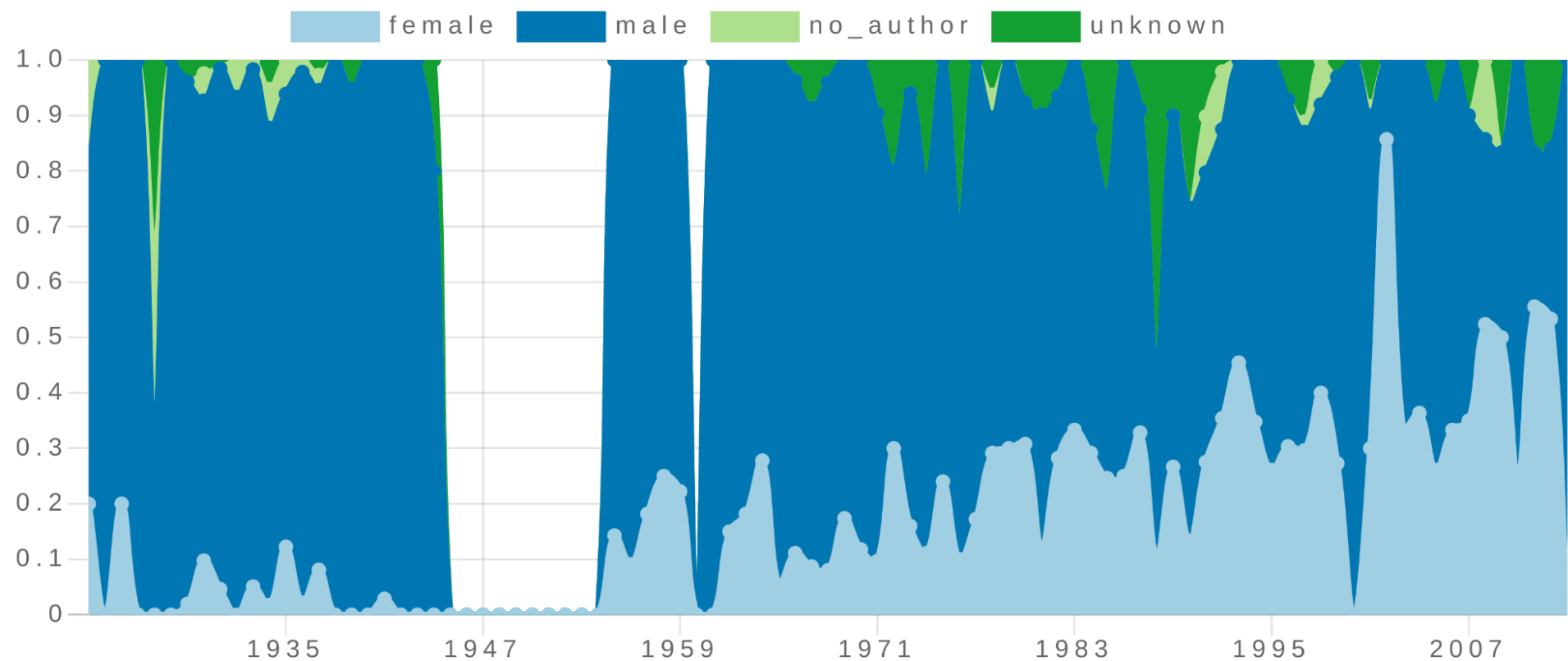
**Date:** 1993

**Number:** 1-2

**ISSN:** 1216-8033

**Page Range:** p. 22

**Language:** magyar



Distribution of **female**, **male**, **no\_author** and **authors with unidentified gender** in 1653 linguistics articles in the Acta repository (U. of Szeged)

# Interactive metadata visualizations

Visualize the (filtered) collection by selecting the metadata field(s) and the type of chart.

Choose diagram type

Network

Choose metadata field for visualization

Authors



number of top items per metafield

20



Choose metadata field for visualization

Publisher



number of top items per metafield

20



Choose metadata field for visualization

Automatic Tags



number of top items per metafield

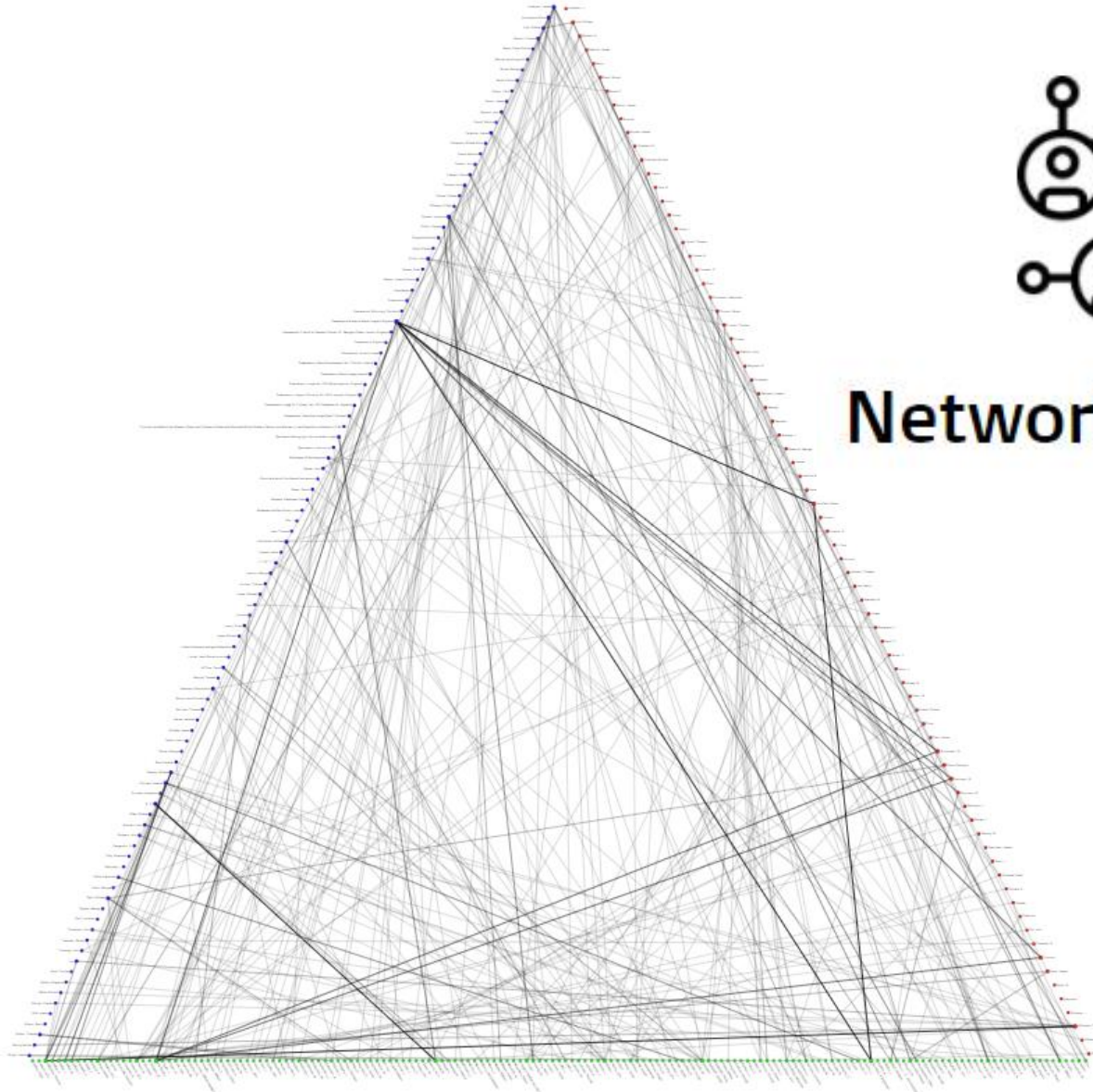
20



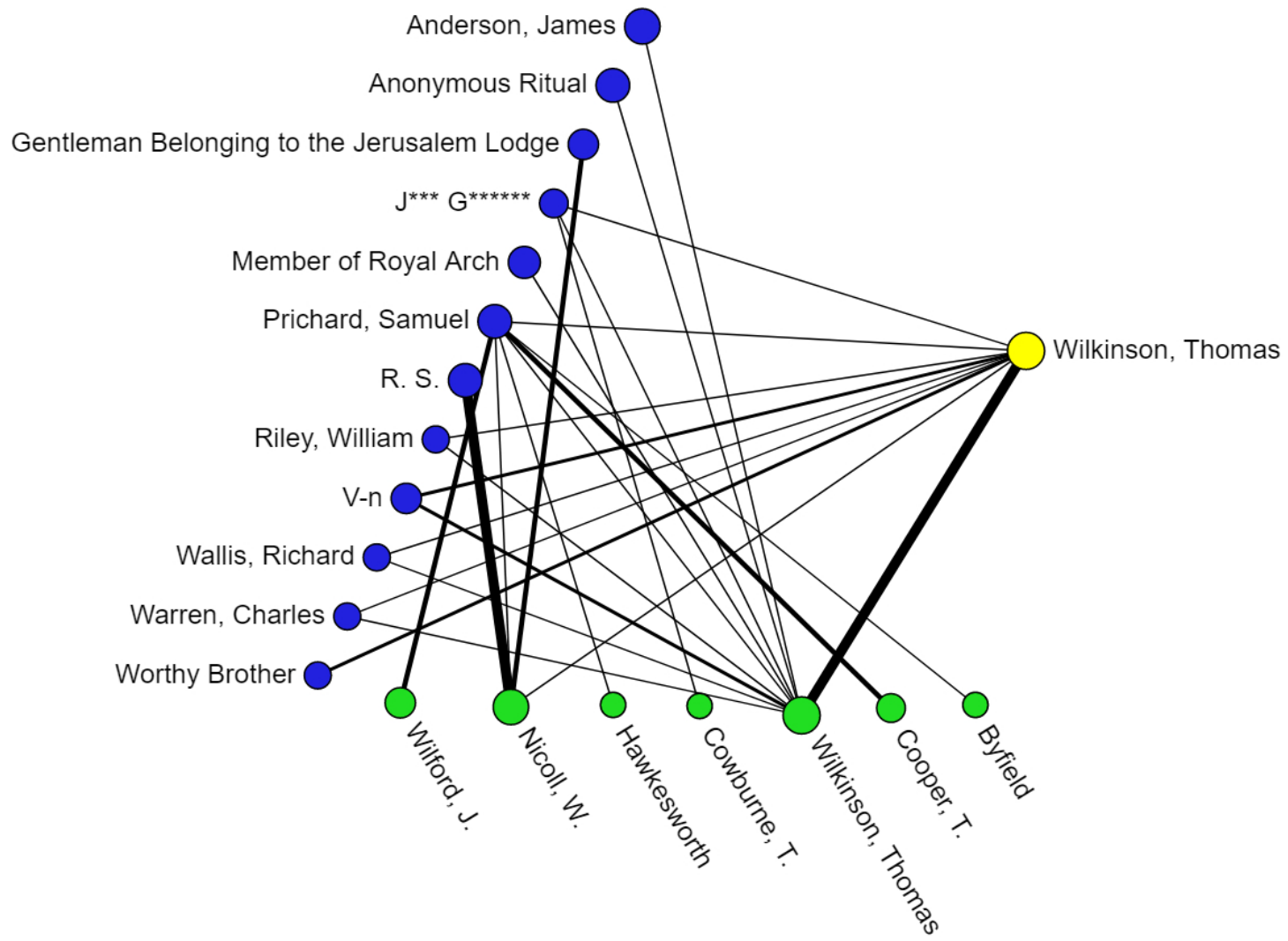
Show visualization

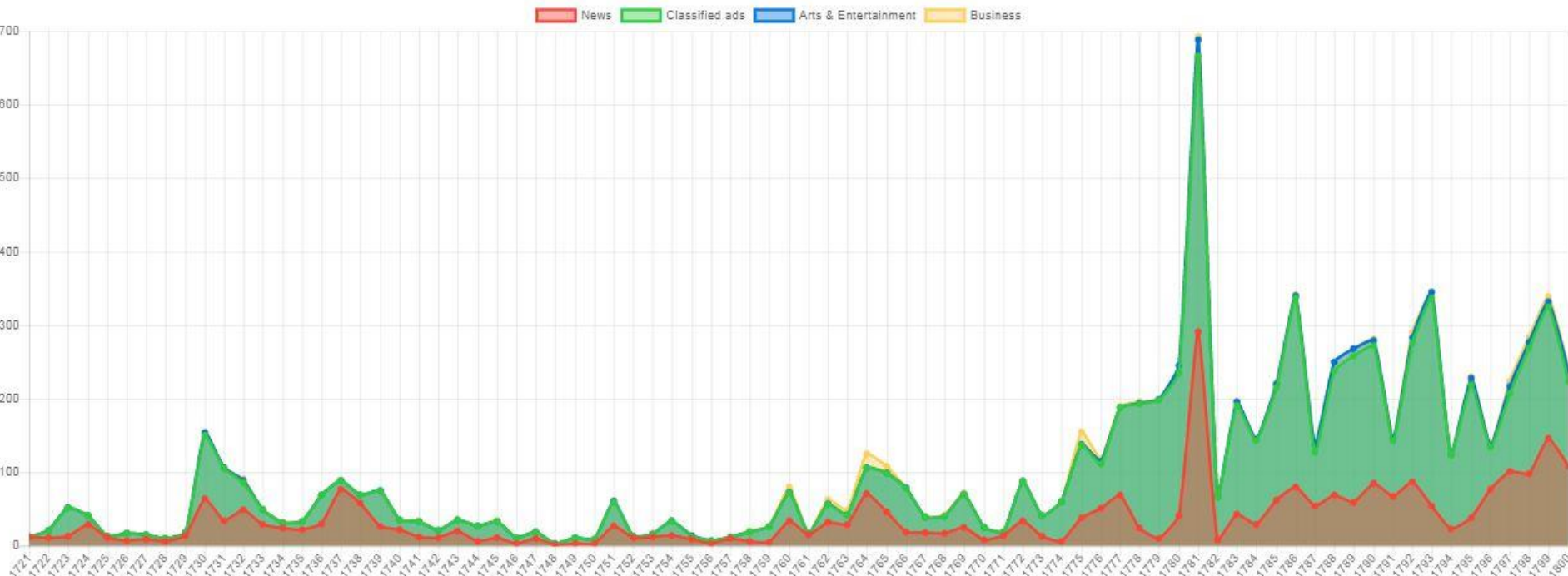
Cancel

# Author-publisher-bookseller network



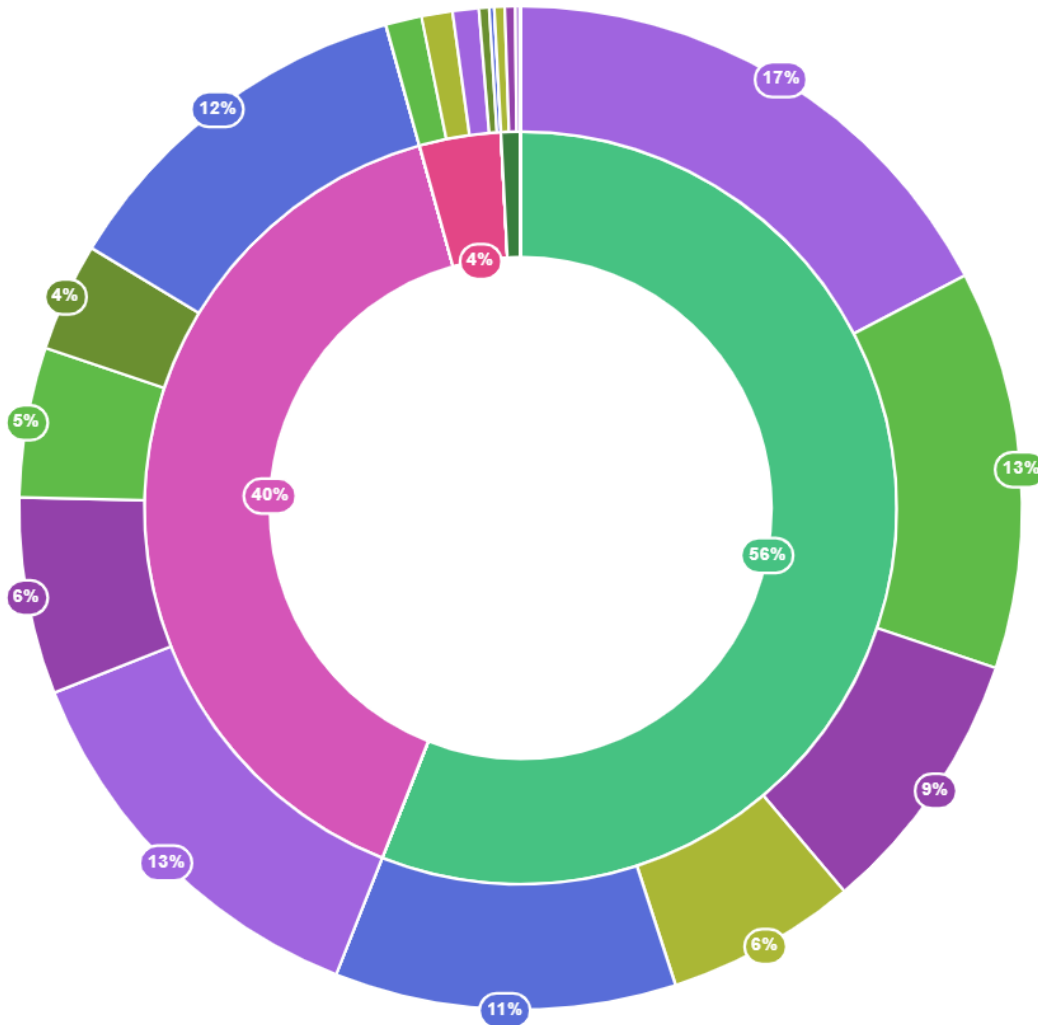
**Network analysis**





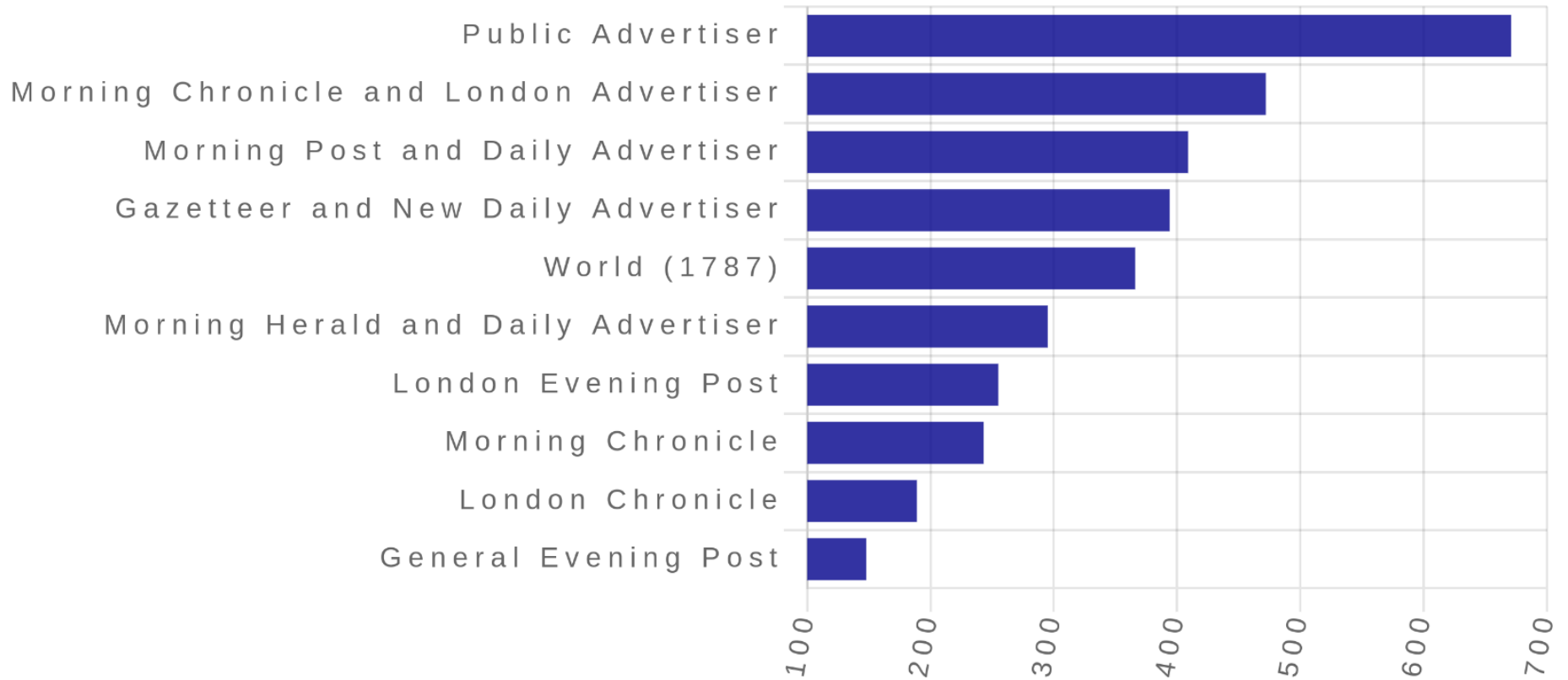
**Interactive metadata  
analysis & visualizations**

# Pie chart



- Classified ads
- Classified ads - Morning Herald and Daily Ad...
- Classified ads - Morning Chronicle and Londo...
- Classified ads - Public Advertiser
- Classified ads - London Courant and Westmins...
- Classified ads - Others values
- Classified ads - Missing values
- News
- News - Morning Herald and Daily Ad...
- News - Public Advertiser
- News - Morning Chronicle and Londo...
- News - London Chronicle
- News - Others values
- News - Missing values
- Arts & Entertainment
- Arts & Entertainment - Morning Chronicle and Londo...
- Arts & Entertainment - London Courant and Westmins...
- Arts & Entertainment - Morning Herald and Daily Ad...
- Arts & Entertainment - London Chronicle
- Arts & Entertainment - Others values
- Arts & Entertainment - Missing values
- Business
- Business - London Courant and Westmins...
- Business - Public Advertiser
- Business - Aurora and Universal Advert...
- Business - Others values
- Business - Missing values
- Others values
- Missing values

# Bar chart





# Text analysis





# Significant words in a subcorpus

| 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%  |

# Frequency analysis 2

## Wordcloud

Choose visualization type

Tagspheres ▼



keyword

freemason

# TagSpheres (context of a word)

maximum word distance

3

minimum frequency of words

5

Context:  Left  Right  Both

Show visualization

Cancel

pantomime  
procession  
night pantomime  
tomime pantomime  
pan night procession time  
inn harle hall harlequin time time  
lequin conclude letter grand  
**freemason**  
queen har mime harlequin  
informed ladies quin new  
repeated great tavern morrow new  
given called called panto  
principal performed  
principal called

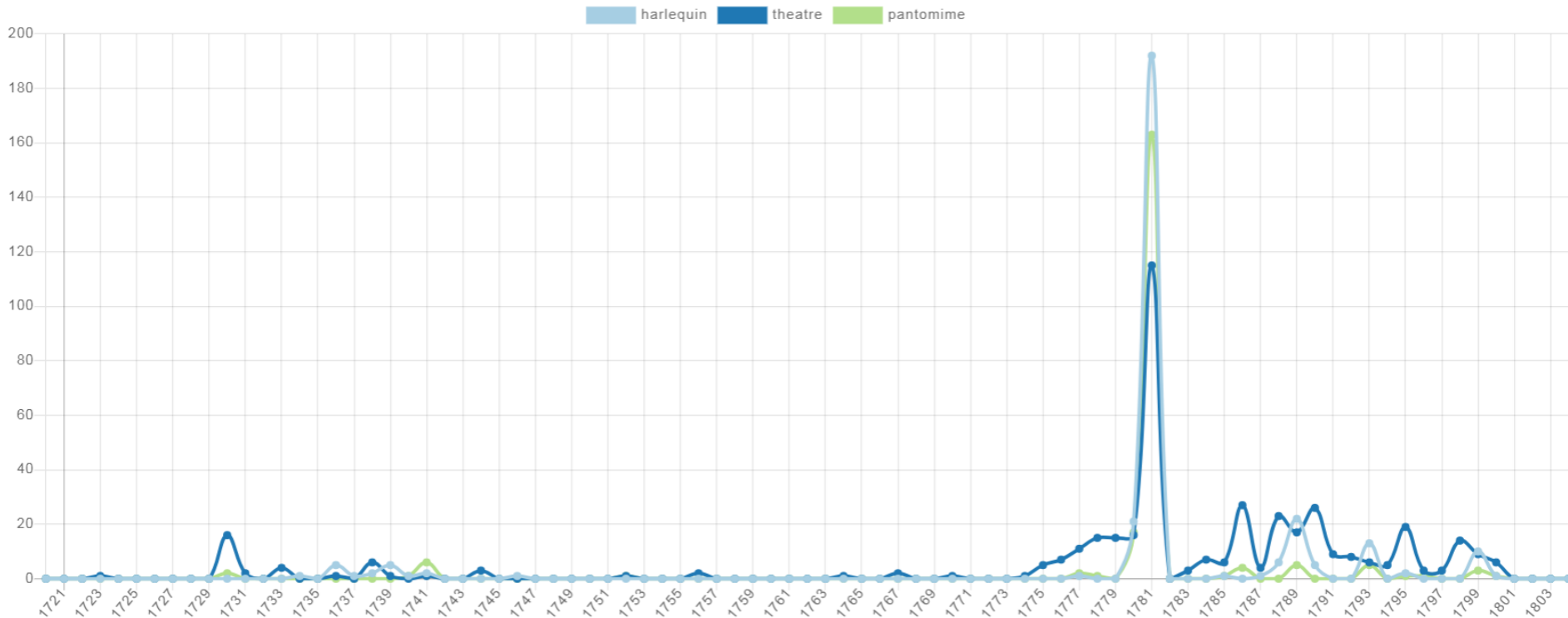
# Most frequent words in a given word distance

| Word       | Text | Count | %       |
|------------|------|-------|---------|
| freemason  | 0    | 281   | 100.00% |
| harlequin  | 1    | 182   | 64.77%  |
| conclude   | 1    | 78    | 27.76%  |
| hall       | 1    | 17    | 6.05%   |
| tavern     | 1    | 12    | 4.27%   |
| harleqqjin | 1    | 11    | 3.91%   |

# Close reading – KWIC

| Authors         | Title          | Pu... | Text   |
|-----------------|----------------|-------|--|
| Without authors | News           | 1781  | tragedy is in re hearfal and will be brought out as fpcedily as po ilible Covent garden Theatre So many Ladies and Gentlemen having been difappointed of places on each night represen tation of the new pantomime call Harle quin <b>freemason</b> thoeie who have fecurcd places for the enfuing nights arc leipcilfully acquainted that entertainment will certainly be performed every night this week and the next This Evening by defire the Belle Stratagem And on Monday next the tragedy of Jane Shore The th night of the new opera call the Iflanders will be on Thurfday next Meetings of Creditors This Day at Guildhall    |
| Without authors | News           | 1781  | week And new Tragedy is in rehcaffjl and will be brought out as fpocdily as poffible Ceyatt Gof ien So many ladies and gentlemen having been difappoiatcd of places on each night represen tation of the new Pantomime called Harlequin <b>freemason</b> thofe who have fecured places for the cnfiling night are repedfully acquainted that entertainment Jwrjll ceraiuly be performed every night this week and the next This en ing by delire the Belles Stratagem and on Monday next the tragedy of Ja ie Shore TJic th night of the bew opera callcd thc IflanJcrd Will be on Thurfiay next The Piles arc diferdcrc to which        |
| Without authors | Classified ads | 1781  | that accompanied the moft illuftrious actions of their lives The hint indeed is almoft imnceflary as gratitude is one of the chief charaferiftics of every Free and ac cepted Mafon When the Clown in the new pantomime cal led Harlequin <b>freemason</b> takes the pig out of the baOcct fome of the audience feel nn eafy poflibly from an idea that the animal is hurt In fatifsaction to their feelings it is right to fay that the pig receives not the frnal left injury but fqucaks as all pigs do when they arc touched ever fo gently from mere apprc henfion Frequent complaint has been made of the theatre and particularly |
| Without authors | Classified ads | 1781  | du Mefnil the beft performer in many refpefts ever fecnand will be produced few day new tragedy is in rchearfal and will be brought out as fpcdily as poffible Covent gardcn Theatre Ladies and gentlemen who have feeured places for thcnfuing representations of the new pan tonine call Harlequin <b>freemason</b> are refpeit fully informed that it will be repeated ever evening this week This Evening with the tra gedy oi Jane Shore To morrow with the Beg gar Opera On Wedncfday not ailed this Icafon the Comedy of Errors The th night of the new opera call the Iflanders will be on Thurfday next new tragedy is          |
| Without authors | News           | 1781  | performed in few dots new tragedy is in rehcarl al and will be brought out as fpcdily as pofiblic Consent Garden Theatre Ladies and gentlemen who have fecurcd placcs for the enfuing rcprentatioiis of the new pan tomime called Harlequin <b>freemason</b> are rc fpc fully informed that it will be repeated every evening this week this evening with the tragedy of Jane Shore and to   |

# N-gram viewer



Aggregated and normalized views



# Topic modeling (LDA)

Run  iterations Completed iterations: 160

Topics:  Alpha:  Beta:

[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

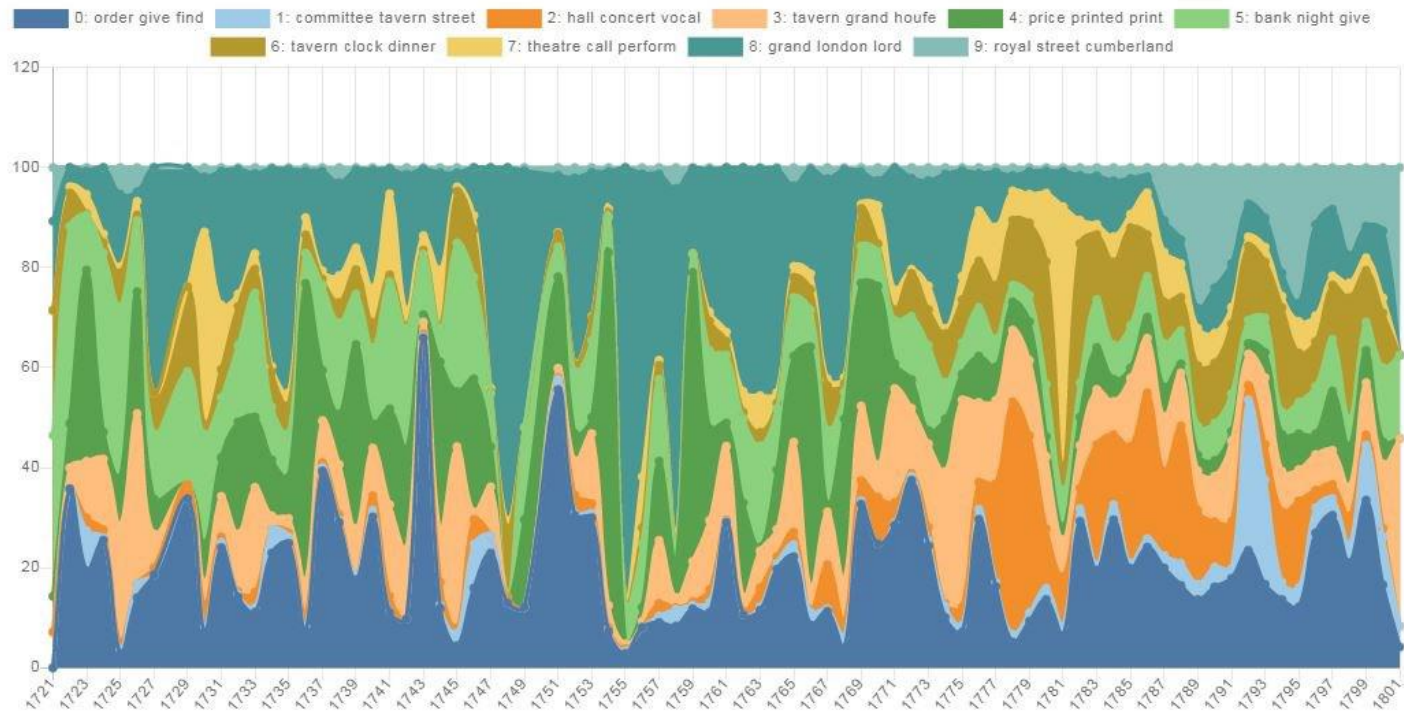
[7] theatre call perform harlequin pantomime  
boxes door grand comedy add

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

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

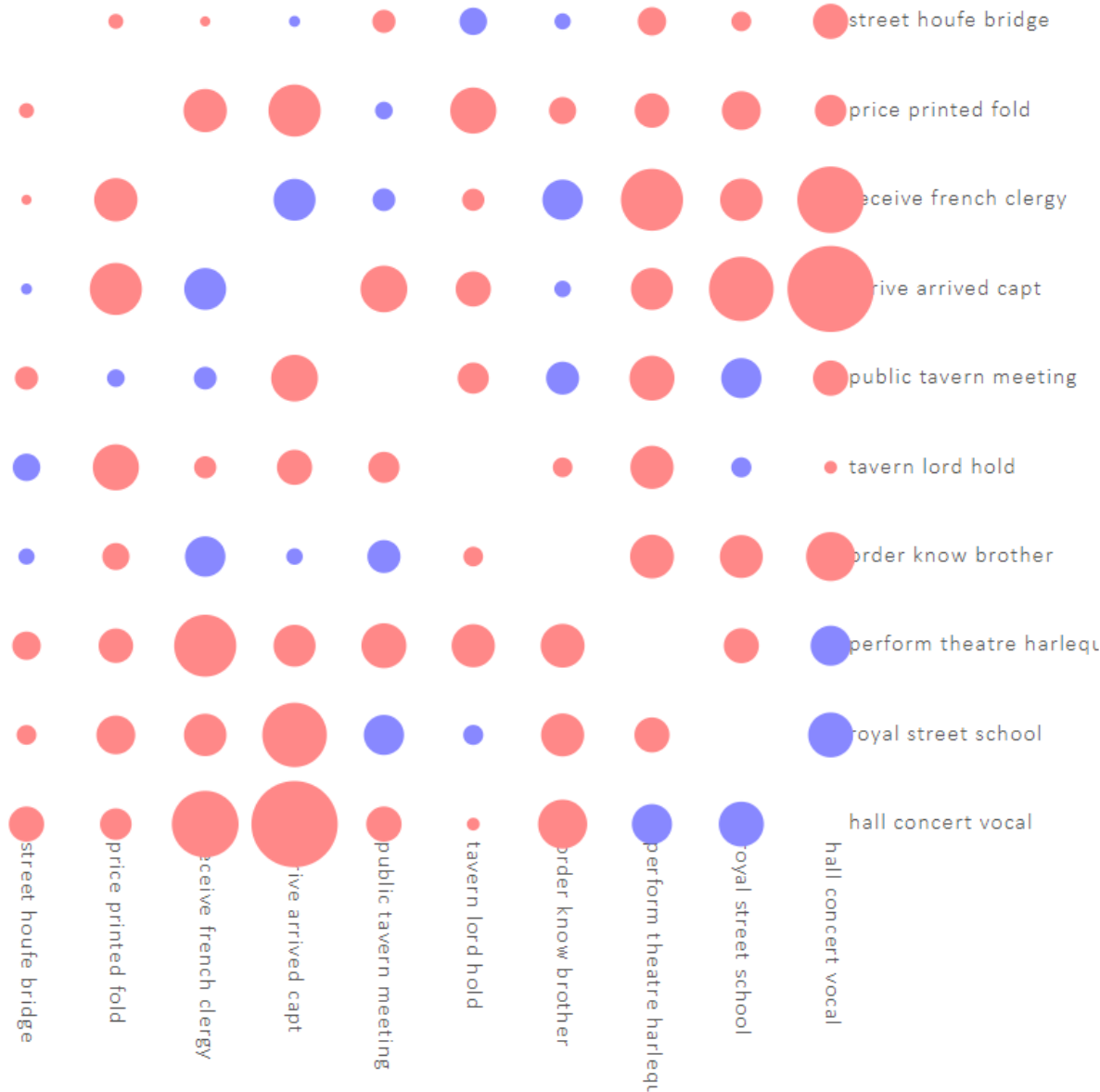
Topic Documents Topic Correlations **Time Series**

Vocabulary Downloads



Aggregate

# Topic correlations



# Named Entity Recognition

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      | <b>PERSON</b>        | 1         |
| Madrid                  | 6       | <b>LOCATION</b>      | 1         |
| Camilo José Cela        | 3       | <b>PERSON</b>        | 1         |
| Pabellón                | 2       | <b>MISCELLANEOUS</b> | 1         |
| Papeles de Son Armadans | 2       | <b>MISCELLANEOUS</b> | 1         |
| Historias de Venezuela  | 2       | <b>MISCELLANEOUS</b> | 1         |
| La catira               | 2       | <b>MISCELLANEOUS</b> | 1         |
| Padrón                  | 2       | <b>LOCATION</b>      | 1         |
| También                 | 2       | <b>LOCATION</b>      | 1         |
| España                  | 2       | <b>LOCATION</b>      | 1         |
| Venezuela               | 2       | <b>LOCATION</b>      | 1         |
| Bidasoa                 | 2       | <b>LOCATION</b>      | 1         |
| Tenia                   | 2       | <b>LOCATION</b>      | 1         |

NER of abstracts and publication titles →  
**Metadata enrichment**

# Poster, news & try it!

- AVOBMAT – DARIAH poster
- News about the development and release  
<http://bit.ly/avobmat>
- Try a limited beta version with a COVID-19 dataset: [avobmat.hu](http://avobmat.hu)



Covid-19 Database  
in  
**AVOBMAT**

**Thank you for  
your attention!**