01 | TELL AN EVENT APART
FROM ITS AVATARS: THE TITANIC CATASTROPHE AND ITS MOVIES

In this challenge, we invite you to get acquainted with the features of the impresso interface: how to read the visualisations, how to search, how to use the topic modelling and the Named Entities to narrow down or expand your research. We will also guide you to use the intermediary analysis tools such as the inspect and compare page and the creation of collection. All the answers to the questions can be found in the FAQ page or directly on the interface. Are you set? Let’s start!

1.

start with a query in the ngram viewer: where does the word titanic appear?

GOAL 1 collect articles on the event and on the movie in separate collections

GOAL 2 find movies dedicated to the event other than the one from 1998
in contrast, search for the word titanic in the general search page: what is the difference between the number of hits shown in the ngram viewer and the search page?

2. change the frequency line from sum to % what changes?

describe what you see:
when are the highest hits for the word “Titanic” and how can you explain it?
we will start now with searching for articles about the movie Titanic, with the help of **NAMED ENTITIES**: Filter by **PERSONS** related to the 1998 movie
To complement the current list of results, look for topics which relate to movies. **EXCLUDE** the collection you created in the previous step to avoid duplicates.

**save that result list in a new collection**
open the **INSPECT & COMPARE** page. In column A on the left side, click to add the current search query. Find column B on the right and select the “Titanic 1998” collection.

Scroll down and take note of the similarities and differences in columns A and B.

can you recognize information in those filters that will help you identify movies that came out before 1998 on the Titanic?

inspect the person list: what does it tell you?
compare the following queries:

**COLUMN A**: search for the keyword “titanic” and filter with the **TOPIC** “film, cinéma...”

**COLUMN B**: search for the keywords “titanic” then “film”

look at the frequencies and the topics: what do you notice? where do the queries A and B overlap? what is the difference?

scroll down to the persons and exclude on both sides the main actor of the 1998 movie and the director: what do you notice? what names do appear in column A and B?
use the metadata filters and curate the left column to focus on movies produced before 1998 on the Titanic catastrophe.
open the curated list in the SEARCH PAGE

create a new COLLECTION titled “titanic movies pre 1998” based on this result list, to collect articles about other movies than the 1998 Titanic movie
now we will look for articles dealing with the catastrophe itself:
search for the word **titanic** and select articles published in the **FRONT PAGE**

look for the relevant topics in **DE** and **FR**
open the **WORD EMBEDDING** feature to look for OCR mistakes frequently associated to the word “titanic” and other words connected to it, in **FR** and **DE**.

What do you notice? Why do the words “naufrage” and “hollywood” appear?

**what are the difference between the french and the german words surrounding Titanic?**

in the word embedding suggestions in German, select **carpathia** or **lusitania** with the same topics as filter
as an overall conclusion: if you go back to the first expectations formulated based on the frequency line and unigram viewer of the titanic query:

what has changed? what did you learn about what that result list means?
CHALLENGE COMPLETED!

You have learned:

✓ how to read the ngram and frequency lines
✓ what the impresso corpus contains
✓ how to query the corpus with key-words, persons (named entities)
✓ how to create and curate collections
✓ how to inspect and compare queries and collections
✓ how to curate a result list with topics
✓ how to expand a query with word embeddings, by exploring synonyms and surrounding words and identifying OCR mistakes

You understand how to explore the newspapers collections based on the filters and the visualisations, as a mean to search for an event and its avatars with other means that adding keywords. We have learned how the filters inform us about the content of the result list and help understand what to expect in them, before further close inspection.

NEXT CHALLENGE: understand the limits of these tools!
02 | HOW TO FIND THE RIGHT PERSON? USES AND LIMITS OF NAMED ENTITIES TO NAVIGATE DIGITISED NEWSPAPERS

Now, we propose to test the strengths and limits of the named entities that have been recognised in the impresso corpus, with the challenge of differentiating Robert Schuman - the politician, from Robert Schumann, the composer.

**GOAL 1**
create two collections of articles, one dealing only with the politician Robert Schuman and one dealing only with Robert Schumann, the composer.

**GOAL 2**
Create a sub collections with articles mentioning only Robert Schumann’s pieces, played at concerts (create a timeline/frequency graph with the concerts).

1. search for “Robert Schumann”
2. select entity Robert Schumann
look at the search filters: who do you see?

what other named entities are mentioned in the result list?

can you spot the second Robert Schuman? there seems to be an overlap between the two entities...

the problem at stake: entity link attribution errors, misspellings by journalists

How to deal with it?
Go to **INSPECT & COMPARE** and compare the queries for each entity: Robert Schumann—the composer (in column A) and Robert Schuman—the politician (in column B).

### 5. Results in common

Lists of newspapers, named entities and topics for articles which appear both in A and B.

### 6. Newspapers

- La Liberté
- Die Tat
- La Liberté
- Neue Zürcher Zeitung
- L'Express
- L'Express
- El Imparcial
- Escher Tagblatt

Open the results in the search page.

Which topics do you see related to each entity? What are the topics of the overlapping articles?
with the help of the topics, distribute the articles between the correct historical figures and add all relevant articles to a collection dedicated to Robert Schuman—the composer and the others to a Robert Schumann—the politician collection.
Now that we have curated the list of articles mistakenly connected to both Robert Schuman(n), we will take a step back and look at the other results from the original queries.

Repeat the query for the Named Entity “Robert Schuman” and exclude the curated articles from the result list.

Next, look for the named entities that appear in the result list: do you notice any that seem not to fit?
do you see any odd names that do not fit?
what about the national lampoon's animal house?
select this entity as an additional filter (filter by person, mentioning ALL of the following)

who is the Robert Schuman mentioned in that context? what is his job? are the articles referring to the politician? if yes, let's continue with him! if not, exclude this entity from the search.

what are we looking at now? What do these 18 articles contain?
Repeat the same exercise with the query for the named entity Robert Schumann.

Curate the result list with the means of excluding non relevant named entities.

(Bonus: for instance, where do Gérard Depardieu and Robert Schumann appear together and why?)

Next, we will narrow down the result list with the help of TOPICS.

look at the topics: which could be indicative of concert programs on the radio?

(you can also filter by language, to focus either on GERMAN or FRENCH to continue)

explore the topics and select the most relevant one to find radio programs containing announces of Robert Schumann concerts, select the topic and create a collection based on that query
final step of the challenge: export the curate collection of radio programs announcing concerts with Robert Schumann’s music, as CSV file

open the zip file and extract the CSV file.

Import the CSV file in your preferred spreadsheet software, with the parameters of UTF-8 for the content and set SEMICOLON as separator

and now, you can use the metadata to create your timeline
CRITICAL USAGE UNLOCKED!

You have learned to use the features with caution:

✓ Named entities are useful to collect relevant articles at the same time, we have to keep in mind that they result from an automated process that has been validated.

✓ Other automated annotations (other named entities, topics) can help to validate one another: they function as indications rather than evidence.

✓ Named entities and topics as well as metadata can be used to create an overview and detect potentially relevant or irrelevant results.

✓ We can store the curated results in collections.

NEXT CHALLENGE: transpose historical research questions via the interface’s features.
03 | RECONSTRUCT THE MEDIA COVERAGE OF AN EVENT IN DIGITISED NEWSPAPERS: THE DREYFUS TRIAL.

In this challenge, we propose to combine the features that have been presented earlier to you, in a critical use, to try and analyse the Swiss and Luxembourgish coverage of the Dreyfus affair.

3.1/ How to find articles dealing with the Dreyfus affair? In other words, how much is the word “dreyfus” synonymous with the Dreyfus affair?

GOAL
identify articles dealing with the Dreyfus trial and revision.

Start with the NGRAM page:
what can you tell from the distribution of the word “dreyfus”?
Apply the same query in the **IMAGE SEARCH**: what can you tell from the information associated with the word “dreyfus”?

Go now to the **TOPIC PAGE**: search for “dreyfus” again: what is the result and why?
Now go to the search page, search again for “dreyfus”: what is the outcome? What are the **TOPICS** connected to that word?

What are the **ARTICLE TYPES** that may not be relevant to find articles dealing with the Dreyfus affair?

What are the **LOCATIONS** that could be relevant to find articles dealing with the Dreyfus affair?
Create a collection of articles containing “dreyfus” in their TITLES, that are NOT ADVERTISEMENTS and that contain the LOCATION “Rennes”.

What do you notice? Which topics and persons appear in the result list?
3.2/ Which are the most relevant persons mentioned with Alfred Dreyfus?

How often were Emile Zola and Alfred Dreyfus co-mentioned in the press?

When was Zola mentioned first in the press, as available in the impresso corpus? When was he mentioned the most? When was he mentioned first in the context of the Dreyfus affair?

go to the INSPECT AND COMPARE page and compare the coverage of the words “dreyfus” and “zola”

what can you say about the TOPICS related to each word?
do the same with the named entities Alfred Dreyfus and Emile Zola

**Topics**

<table>
<thead>
<tr>
<th>Topic</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR tribunal - affaire - prison - procès - mois</td>
<td>459</td>
</tr>
<tr>
<td>FR presse - journal - lettre - affaire - article</td>
<td>371</td>
</tr>
<tr>
<td>FR armée - colonel - service - major - chef</td>
<td>365</td>
</tr>
<tr>
<td>FR gouvernement - général - ministre - guerre - ordre</td>
<td>338</td>
</tr>
<tr>
<td>FR vie - monde - mort - foi - peuple</td>
<td>213</td>
</tr>
<tr>
<td>FR conseil - commission - projet - loi - rapport</td>
<td>197</td>
</tr>
<tr>
<td>DE regierung - paris - frankreich - minister - kammer</td>
<td>164</td>
</tr>
<tr>
<td>FR président - fête - ville - lieu - discours</td>
<td>152</td>
</tr>
<tr>
<td>FR président - ministre - gouvernement - général - chef</td>
<td>152</td>
</tr>
<tr>
<td>FR question - point - pays - conseil</td>
<td>149</td>
</tr>
</tbody>
</table>

**407 results in common**

Lists of newspapers, named entities and topics for articles which appear both in A and B.

**Topics**

<table>
<thead>
<tr>
<th>Topic</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR tribunal - affaire - prison - procès - mois</td>
<td>44</td>
</tr>
<tr>
<td>FR presse - journal - lettre - affaire - article</td>
<td>39</td>
</tr>
<tr>
<td>FR armée - colonel - service - major - chef</td>
<td>36</td>
</tr>
<tr>
<td>FR gouvernement - général - ministre - guerre - ordre</td>
<td>26</td>
</tr>
<tr>
<td>FR vie - monde - mort - foi - peuple</td>
<td>25</td>
</tr>
<tr>
<td>FR vie - monde - auteur - foi - sen</td>
<td>18</td>
</tr>
<tr>
<td>FR président - fête - ville - lieu - discours</td>
<td>17</td>
</tr>
<tr>
<td>FR fille - main - voix - foi - temps</td>
<td>15</td>
</tr>
<tr>
<td>FR livre - auteur - ouvrage - histoire - vol</td>
<td>15</td>
</tr>
<tr>
<td>FR président - ministre - gouvernement - général - chef</td>
<td>14</td>
</tr>
</tbody>
</table>

**Topics**

<table>
<thead>
<tr>
<th>Topic</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR film - cinéma - semaine - jean - john</td>
<td>978</td>
</tr>
<tr>
<td>FR vie - monde - auteur - foi - sen</td>
<td>972</td>
</tr>
<tr>
<td>FR livre - auteur - ouvrage - histoire - vol</td>
<td>859</td>
</tr>
<tr>
<td>FR vie - monde - mort - foi - peuple</td>
<td>771</td>
</tr>
<tr>
<td>FR presse - journal - lettre - affaire - article</td>
<td>687</td>
</tr>
<tr>
<td>FR scène - théâtre - spectacle - pièce - public</td>
<td>667</td>
</tr>
<tr>
<td>FR tribunal - affaire - prison - procès - mois</td>
<td>573</td>
</tr>
<tr>
<td>FR gouvernement - général - ministre - guerre - ordre</td>
<td>547</td>
</tr>
<tr>
<td>FR fille - main - voix - foi - temps</td>
<td>538</td>
</tr>
<tr>
<td>FR président - fête - ville - lieu - discours</td>
<td>521</td>
</tr>
</tbody>
</table>

look at the articles which contain both named entities: how many are actually covering the affair?
You know now how to use Natural Language Processing and visualisations to query and analyse digitised historical material:

- You learned how to interpret a peak in frequencies: how to differentiate keyword hits from the news coverage of an event (Dreyfus as a string of character vs. the Dreyfus Affair)
- You learned how the topic models were computed, what they entail and what not (nouns vs. names)
- You learned how to use historical information to find relevant articles (location of an event, persons related to specific dimension of an event)

**CONGRATULATIONS!** you are now an expert user of digitised newspapers interfaces!