europeana transcribe

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

ENRICHEUROPEANA+

An Approach for Curating Collections of Historical Documents with the Use of Topic Detection Technologies

Sergiu Gordea, Medina Andresel, Srdjan Stevanetic, Mina Schütz
AIT Austrian Institute of Technology GmbH

The 17th International Digital Curation Conference (IDCC 2022)

June 15th, 2022

Co-financed by the Connecting Europe Facility of the European Union

AUSTRIAN INSTITUTE OF TECHNOLOGY

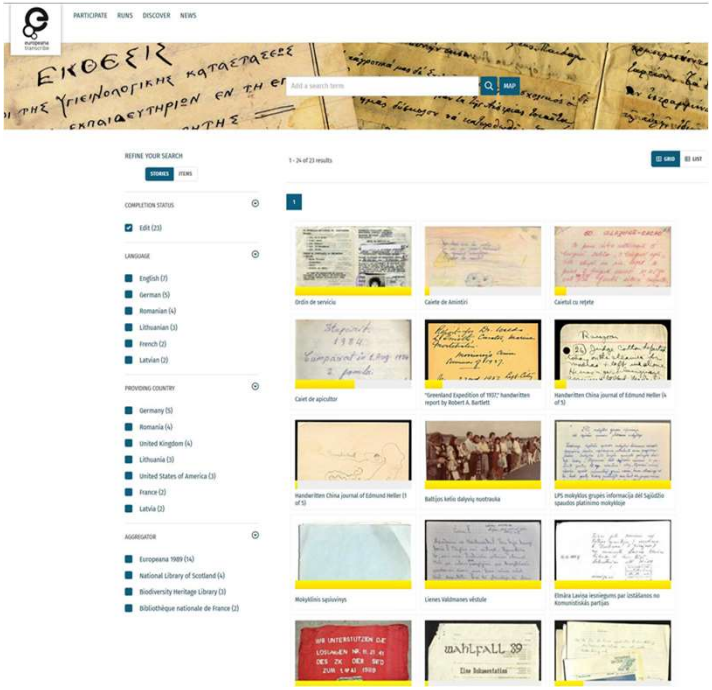
AGENDA

- Introduction
  - Transcribathon tool
- Motivation
  - Content curation for Transcribathon runs
- Proposed approach
  - Topic detection
  - Building LDA Models
  - Topic based information retrieval
- Experimental Evaluation
- Conclusions and Future Work

INTRODUCTION

Transcribathon

- A tool for transcription and enrichment of historical documents
- Using curated materials from Europeana collections

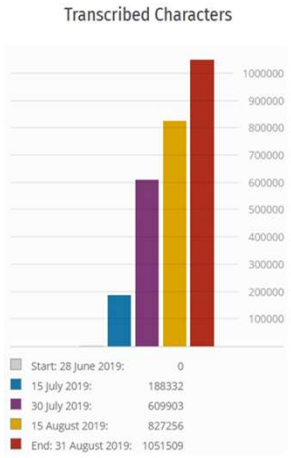


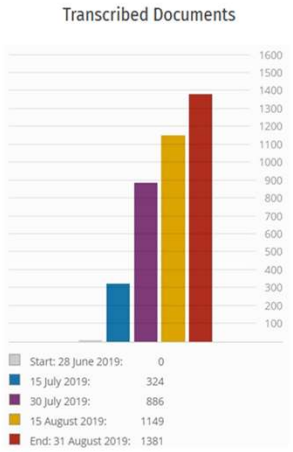
https://europeana.transcribathon.eu

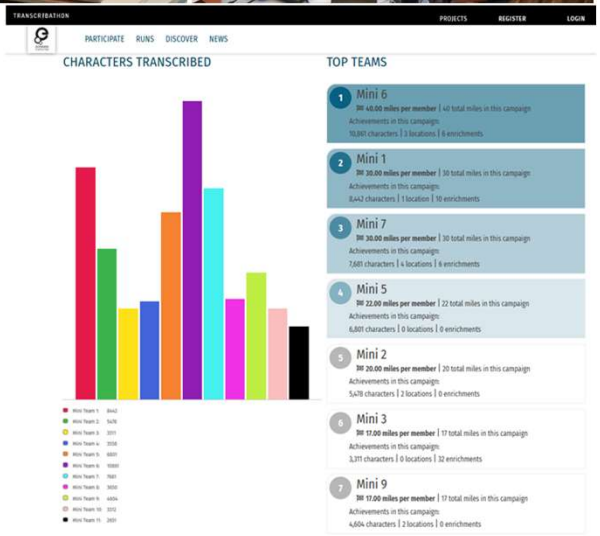
CROWDSOURCING CAMPAIGNS

Transcribathon = Transcription Marathon

- Physical, online or mixed competitions
- A gamification approach
- Focus on targeted audience







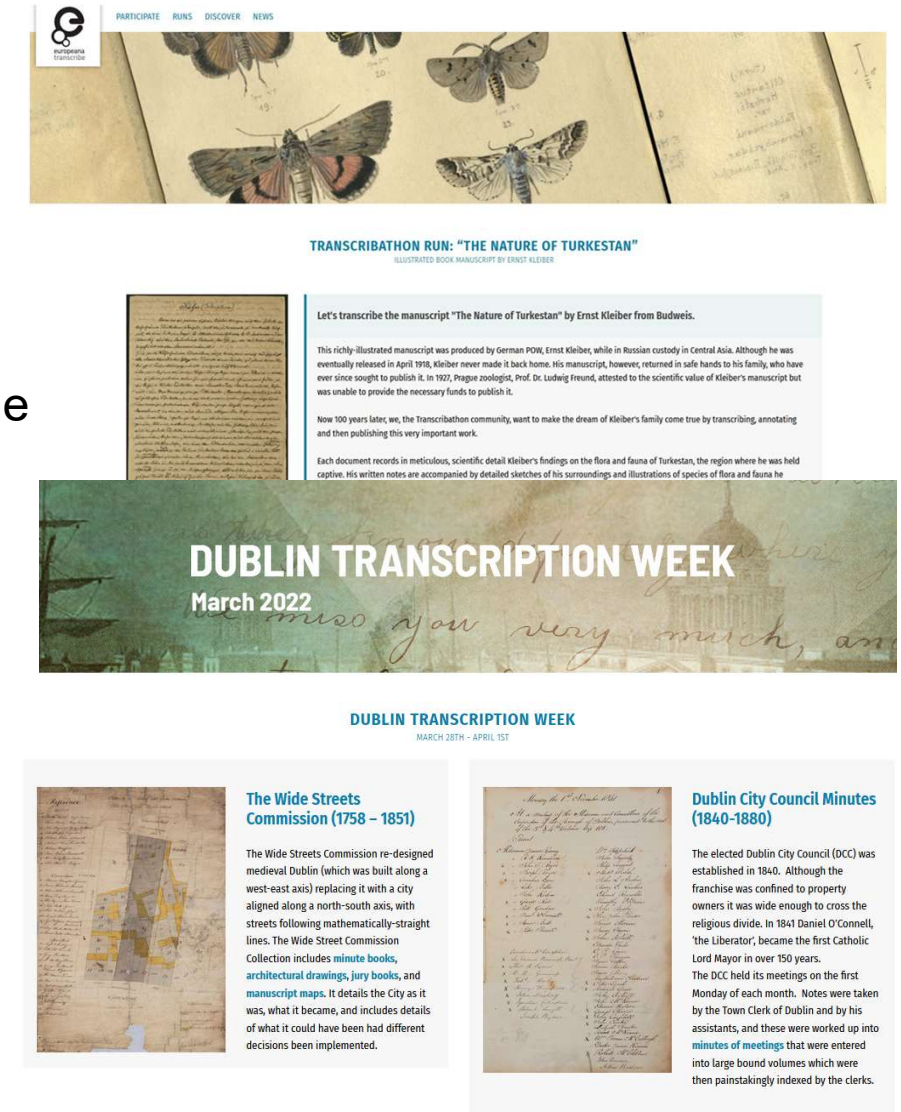
https://europeana.transcribathon.eu

THEMATIC RUNS

Curated materials on specific themes

- Memories from the First World War
- 1989 Revolution in Eastern Europe
- Industrial revolution
- Societal and urban development in the 19th century
- Also curated for local audiences
  - 20+ European Cities

https://europeana.transcribathon.eu



MOTIVATION

Automated support for curation activities

- Curate materials for Transcribathon themes
- Manual content curation is an expensive activity
- Domain expertise and knowledge in different languages is required

AI opportunities for CH

- Machine translation supported for all European languages
- Natural language processing and machine learning approaches available
  - Classification / clustering / recommendations / event prediction …

PROBLEM DESCRIPTION

Proposed Approach

- Use of topic detection technology to support digital curation activities
- Use unsupervised learning approach – LDA (Latent Dirichlet Allocation)

Goals:

- Organize materials available in the Transcribathon tool in several groups of closely related documents (group by topics)
- Implement efficient and scalable solutions for searching related content in Europeana (based on detected topics)

Challenges:

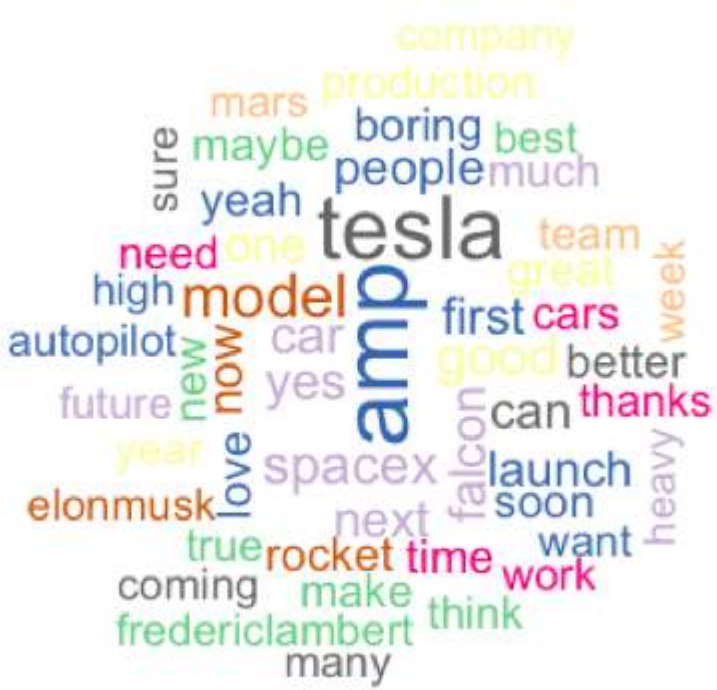
- Ensure appropriate and qualitative data for model learning
- Identify best-suited topic detection models
- Efficient implementation for topic-based search and categorization/clustering

TOPIC DETECTION

“Topic analysis (also called topic detection, topic modeling, or topic extraction) is a **machine learning technique that organizes and understands large collections of text data**, by assigning “tags” or categories according to each individual text's topic or theme.”

Source: https://monkeylearn.com/topic-analysis/

TOPIC DETECTION ON TWITTER



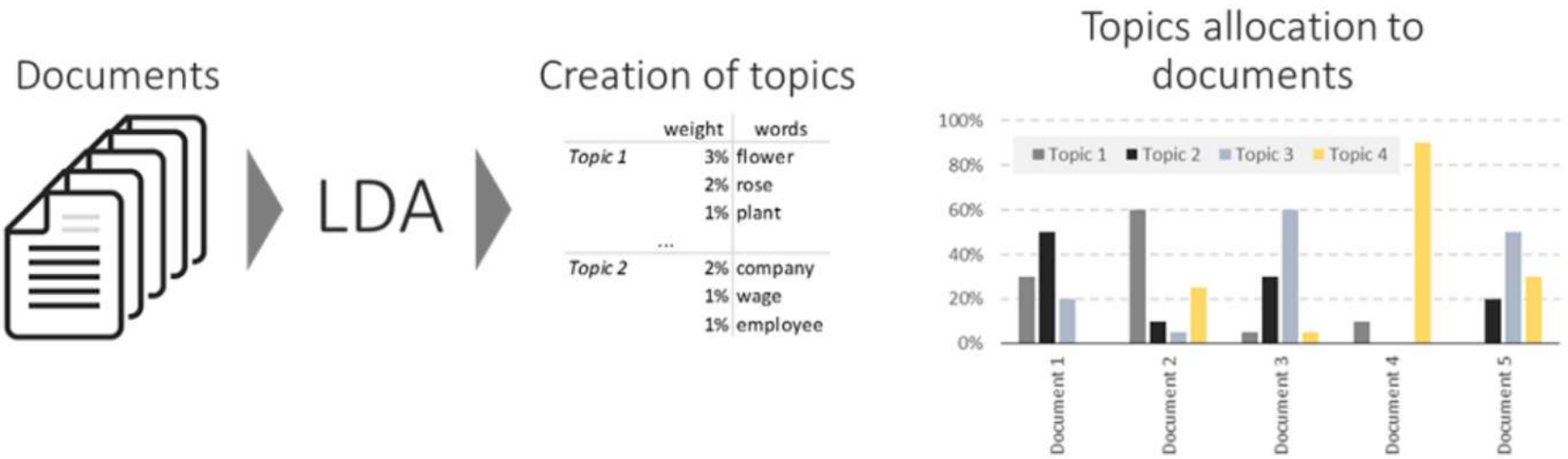
Word Cloud of Elon Musk's tweets

"Elon Musk uses his twitter account to talk about his work in **Tesla**, **The Boring Company**, and **SpaceX** as evident in this word cloud."

Source: https://towardsdatascience.com/elon-musk-twitter-adf324120b3f

TOPIC DETECTION - LDA

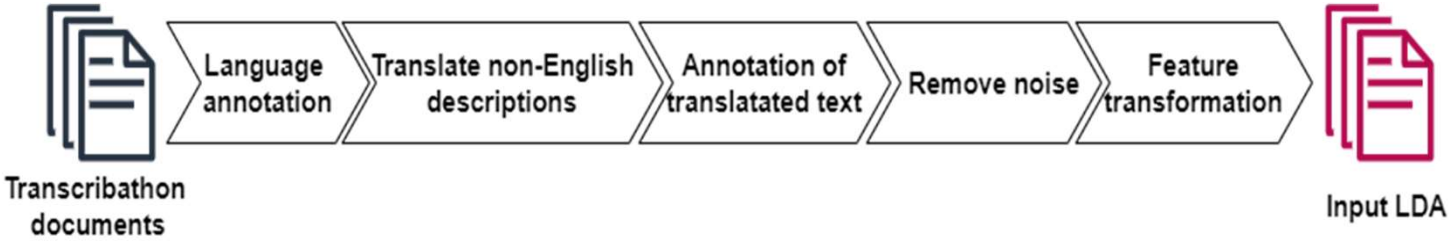
- LDA is one of the most popular unsupervised topic modelling methods
- Each document is made up of various words, and each topic also has various words (together with their weights) belonging to it. The aim of LDA is to find topics a document belongs to, based on the words in it.



Source: https://towardsdatascience.com/the-complete-guide-for-topics-extraction-in-python-a6aaa6cedbbc

LEARNING TOPIC MODELS

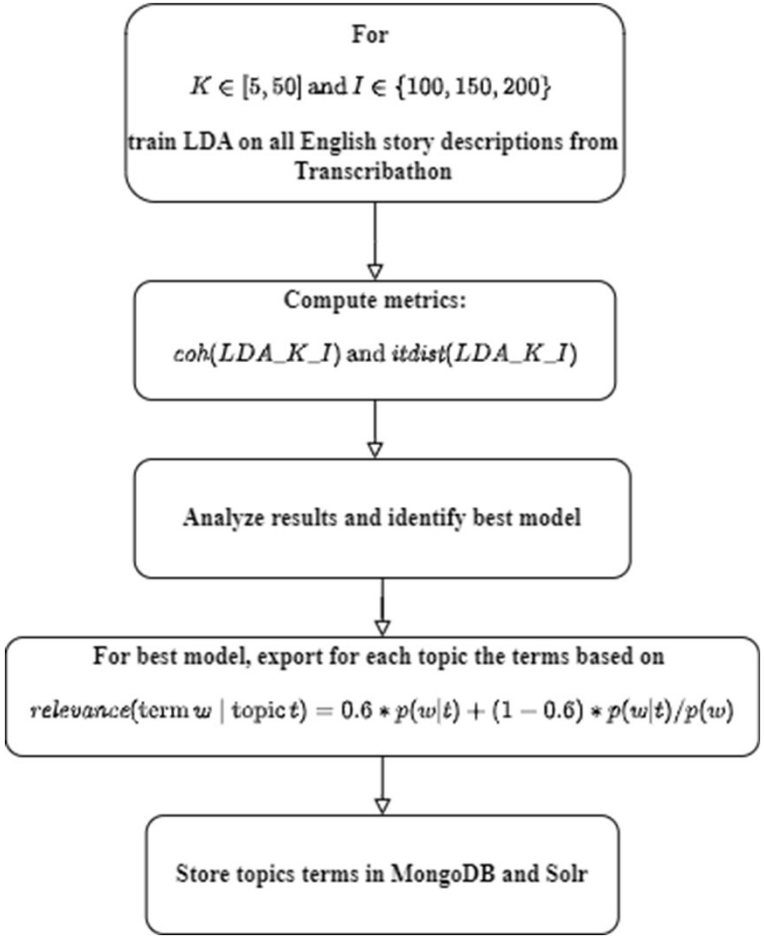
Pre-processing pipeline



Learning Topic Models

- LDA configuration paramerers: ***K*** – the number of topics to be learned,
  I – the number of iterations to run
- Which are the best LDA Models?
- Metrics for assessing the quality of the learned topics:
  - *Coherence* (Shaheen Syed, 2017)
  - *Inter-Topic Distance* (Carson S., 2014)

TOPIC MODEL SELECTION



1. Run LDA with different values for ***K*** and ***I***
2. Compute coherence and inter-topic distance metrics
3. Identify best model (expert intervention)
4. Export most representative topic terms and their relevancy
5. Store topics in the databases

TOPIC-BASED INFORMATION RETRIEVAL

Content curation from external repositories

- Topic assignment for new ingested materials
  - Supported directly by the LDA implementation

- Recommending new materials for a given Topic:
  - Retrieve candidate documents based on topic terms
  - Re-rank documents using the LDA Model
  - Build recommendation list

- How to ensure qualitative document recommendations?
  - Number of documents to search
  - Accuracy of recommendations list (Top N)

EXPERIMENTAL EVALUATION

Answering the following research questions:

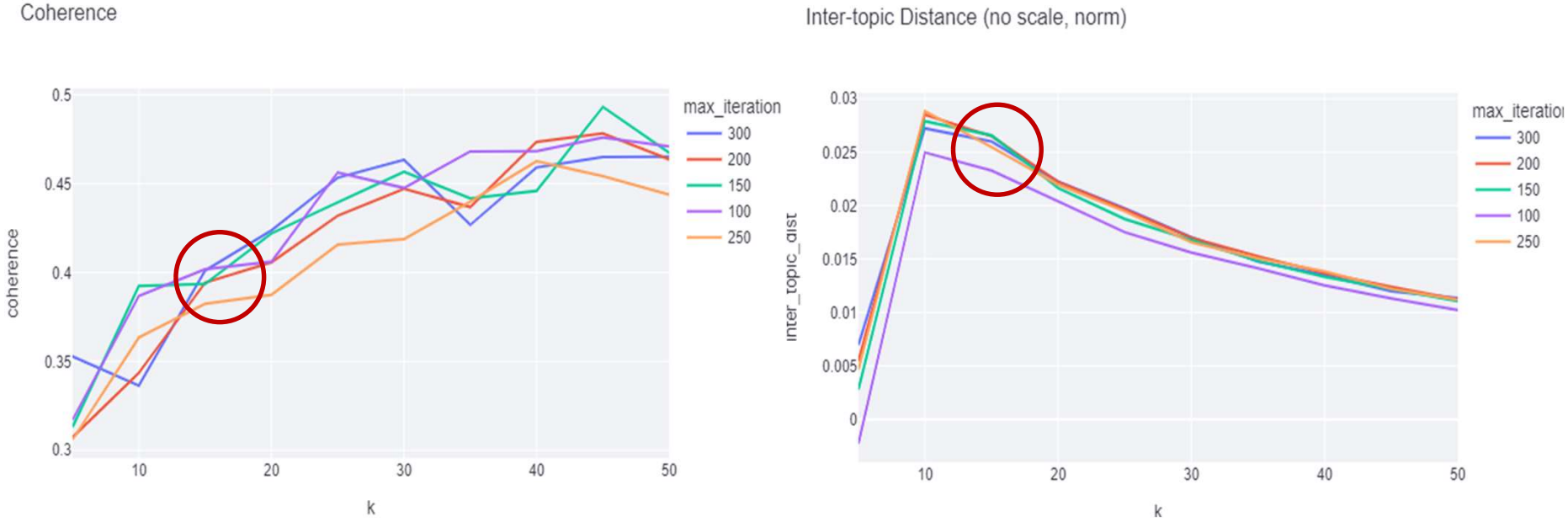
1. Which are the most appropriate ***K*** and ***I*** parameters for learning a good LDA Model on the complete Transcribathon dataset?

2. Which are the configurations required to obtain a good Precision at Top 10 (i.e., ***Precision@10***)?

Dataset:

- 31,957 documents available in Transcribathon (i.e. stories)
- Multilingual dataset (only a very small fraction have English description)
- ~ 28,000 words in vocabulary

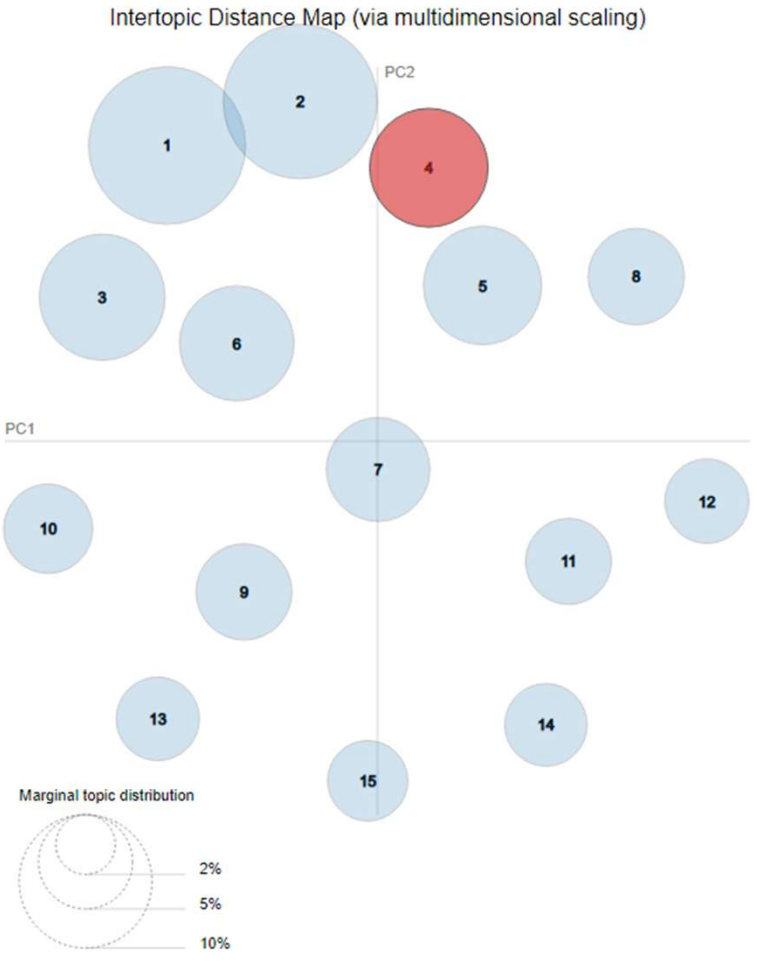
SELECTION OF MOST OPTIMAL LDA MODEL

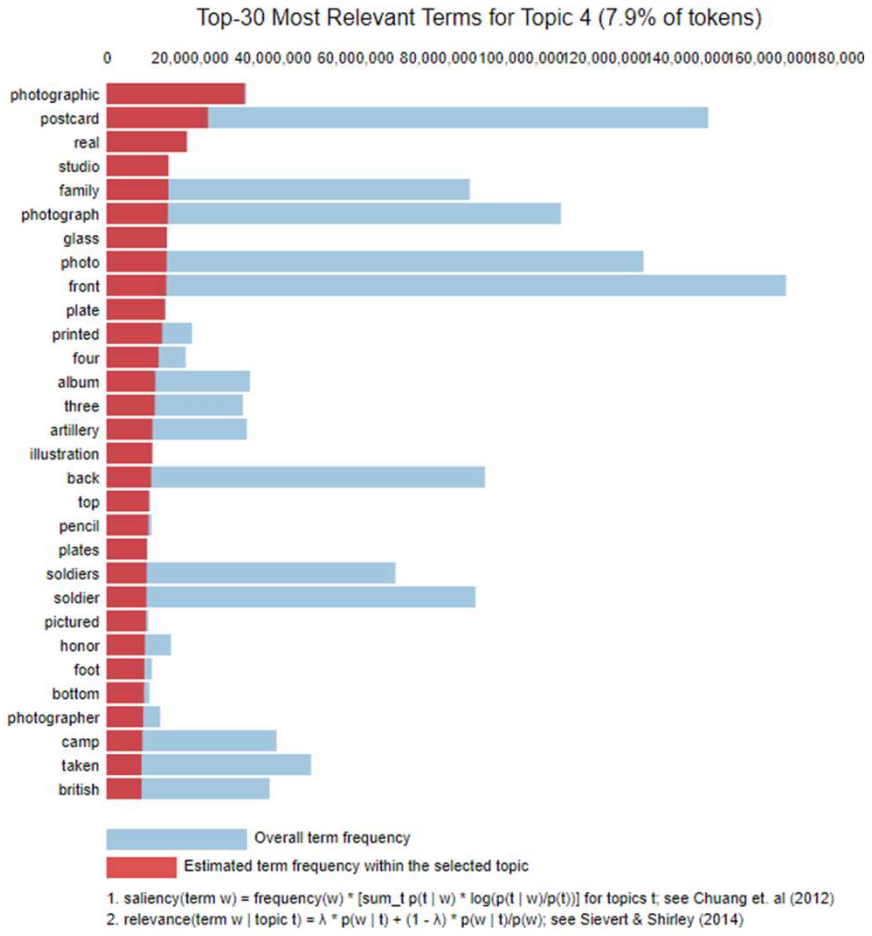


Observations:

1. Model training riches a saturation when ***I*** >= 150 (norm. inter topic distance)
2. Model overfitting for K > 25 (Coherence)
3. Performance for most of the models converges for ***K***=15 (similar performance of Models independent from I)
4. Most appropriate model for this dataset is considered ***LDA_15_150***

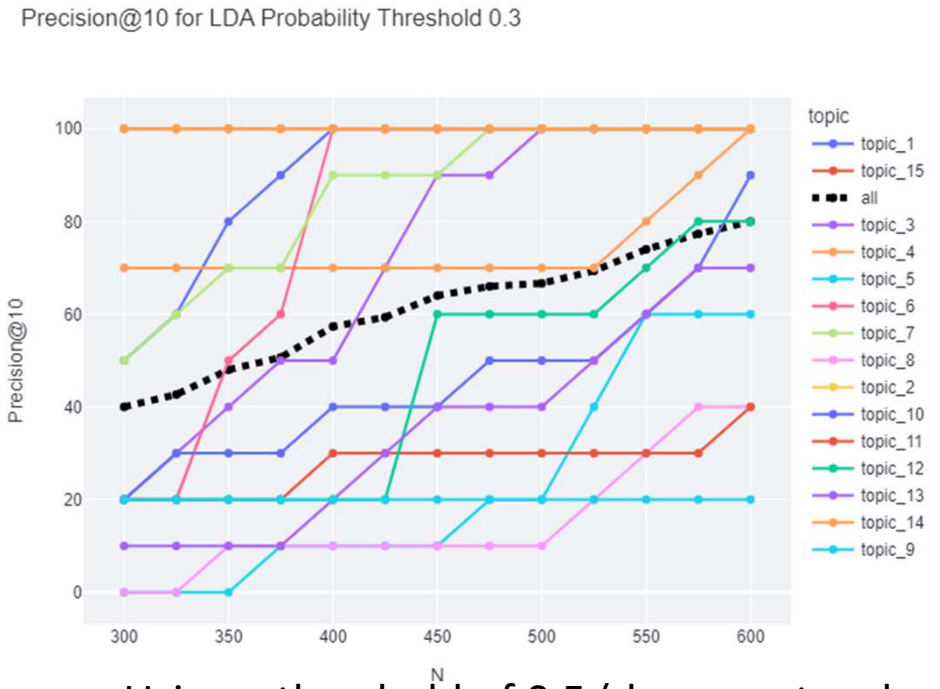
VISUALIZATION OF LDA MODELS

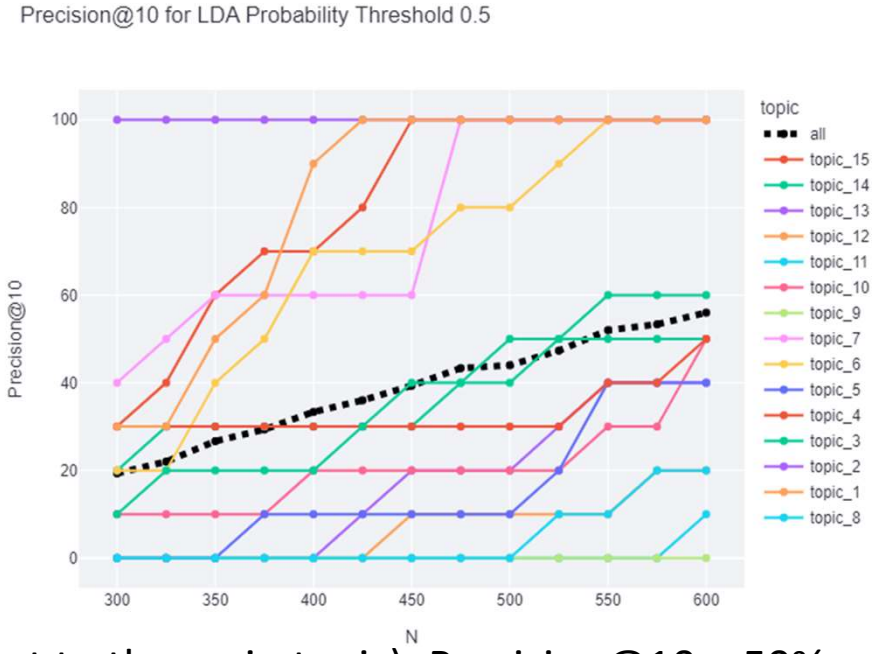




TOPIC-BASED CONTENT RECOMMENDATION

Precision@10 for different number of candidate documents - N and the relevant document thresholds of 0.3 and 0.5





- Using a threshold of 0.5 (documents relevant to the main topic): Precision@10 > 50% for N=550 (for N=600, Precision@10 = 56%).
- Using a threshold of 0.3 (documents relevant to their first and second topic) Precision@10 > 80% for N = 600 is obtained

CONCLUSIONS AND FUTURE WORK

- We propose:
  - Scalable approach for clustering large corpora of historical documents in finer grade collections.
  - Well-defined procedure for learning and choosing the best LDA topic model to support the curation of new materials for Transcribathon campaigns.
  - Search functionality from large CH repositories like Europeana to reduce the computation efforts required by LDA based document clustering.

- Future work:
  - Implementation of topic assignment for a given document using Solr
  - Investigation for using other topic modeling techniques such as BERTopic
  - Evaluate supervised machine learning approaches starting from the previously curated datasets

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

THANK YOU!

Sergiu Gordea

AUSTRIAN INSTITUTE OF TECHNOLOGY

THEMES IN EUROPEANA 1914-1918