A NLP Enhanced Visual Analytics Tool for Archives Metadata

Anıl Ozdemir, Dilara Müstecep, Orhan Ağaoğlu and Selim Balcisoy

Faculty Of Engineering And Natural Sciences, Sabanci University, Orta Mahalle, 34956 Tuzla, Istanbul, Turkey

{aozdemir,dilaramustecep,balcisoy}@sabanciuniv.edu, oagaoglu@tevitol.k12.tr

Sabancı Üniversitesi

1. Introduction

Today, almost all cultural heritage (CH) institutions are starting to digitize parts of their collections and archives to improve accessibility, preservation of originals, publicity, and visibility of the institution on the Internet. These collections are spread over more than one area of life in a vast domain, including art, history, mathematics, physics, etc. and this situation creates a substantial volume of documents digitally available. Also, it creates the need for various approaches that allow users to understand latent meanings in collections, discover and investigate relationships, and extract the necessary information from collections. To address this need, we introduce a visual exploratory tool that facilitates the uncovering of hidden information and stories underlying documents, extracting the key individuals, temporal expressions, locations, entities, and keywords within the documents, establishing a network between documents and allow researchers and archivists to form and test hypotheses and observe individual relationships, networks, and stories present in the archives metadata collections.

2. Design Rationale

Our developed exploratory tool aims to help archivists form and test hypotheses and observe certain relationships, networks, semantic and syntactic proximity, and stories present in the metadata archives. To achieve this goal, the extracted features from metadata must be displayed in various aspects comprehensively. These features include features in different domains that can best describe archive. We have divided this attribute pool into two, the ones we have obtained by ourselves using NLP methods, and those already present in the metadata. Previously available features include the year the document was published, the document's location, and the topic of the document. On the other hand, the attributes we extract with NLP and morphological methods are local locations, persons, dates, similarity, entities,keywords. We have integrated all of these attributes into the visual tool we have prepared to help an archivist analyze a document in the best possible way.

3. Use Case: Waqfs of Crete

To design such a tool, we have collaborated with archive professionals from an cultural institution, SALT (https://saltonline.org/) which focused on public service producing research-based exhibitions, publications, and digitization projects. As a result of our conversations Salt team we decided to use Waqfs of Crete which is an archive consisting of offi-cial records of Muslim inhabitants of Crete. Documents spanning the period from 1825 to 1928 in Ottoman Turkish and Greek provide an opportunity to examine the multi-layered social structure on the island, especially from a cultural and economic perspective. The metadata contains information for approximately 10 thousand documents and includes the summary of those documents, the year they were published, the location, the language used, and the documents' picture.

4. System Description *Visualization*

The designed interface consisting of six components which includes interactive map that allows the user to view documentsin different locations and view the document networks that formed by calculating total number of shared attributes betweendocuments. Remaining components include information box that contains document-specific attributes such as location, time,person, entities, and keyword, document browser that enable users and researchers to browse documents easily, individual andkeyword search menu and filtering panel. In this way, the users may find documents that are roughly related to each othervery quickly. Later, the user can browse each document on its network and view documents that have common individuals andkeywords with each other. Thus, the user may follow the interactions between documents like a story and able to do this for allthe people who lived in the 19thcentury on Crete's island.

4. System Description *Features*

From the raw metadata, we extracted the following features:

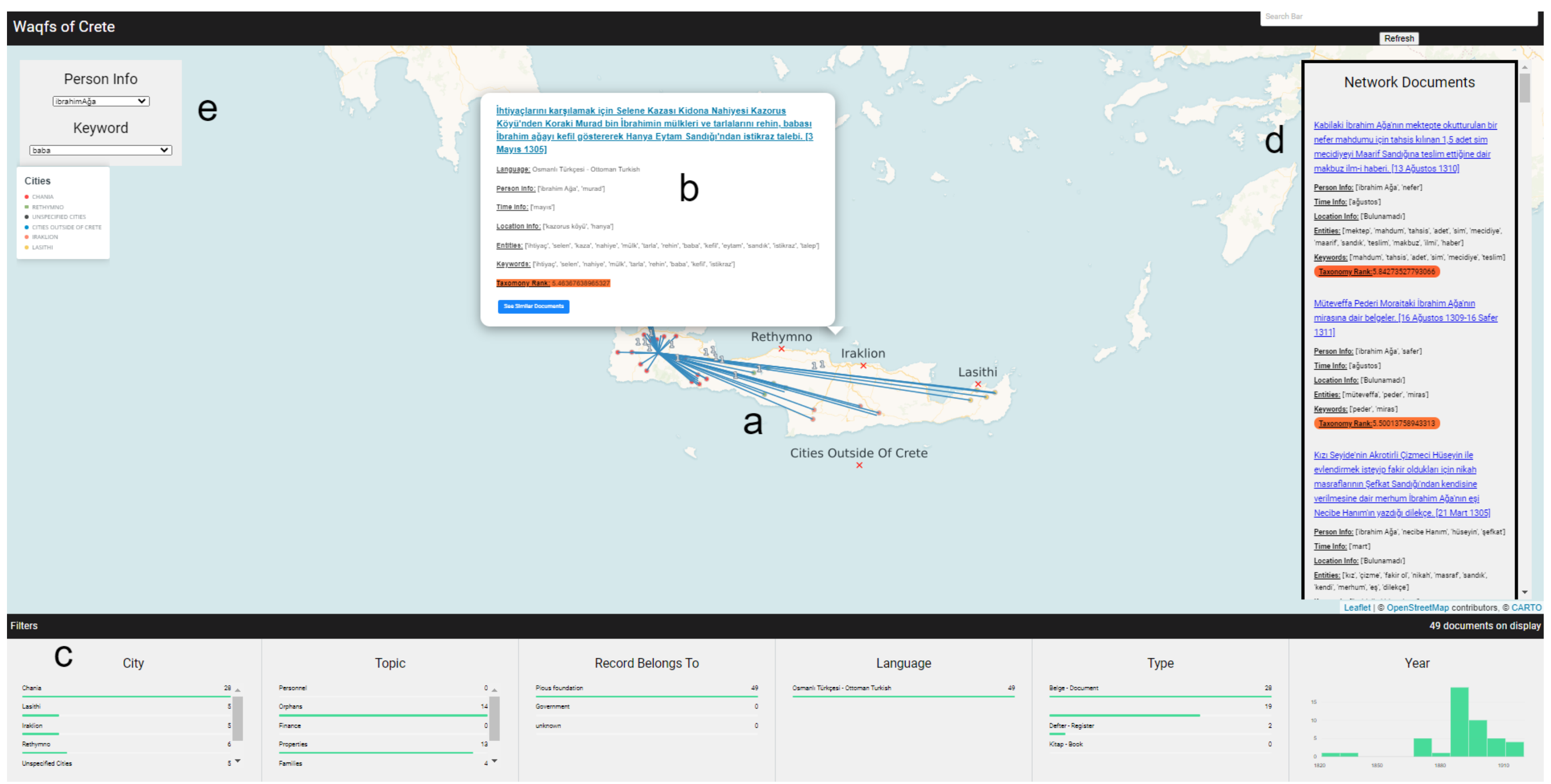
Location, Time: In this paper, we refer to villages and settlements as locations. To obtain them, we extracted location-expressing words morphologically using regular expressions [1]. Although we have the year each document was published, the concept of time in the document's depiction describes a time progression, such as morning, evening, yesterday, the next day.

Entities, keywords, and key individuals: For extracting entities, we first located nouns in the text and then determine what type of entity they refer to – such as a individual, location, or keywords with the help of regular expressions and Zemberek [2].

Similarity Between Documents:
To capture similarity between documents, we used state-of-theart word embedding models including Word2vec [3] , FastText [4] and Transformer l[5] which provides a method to compute dense vector representations for documents. Consequently, each document was represented as fixed-sized mathematical vectors as output of each model, and the similarity between documents was calculated by taking the arithmetic cosine similarities of vectors.

Network: We have integrated all extracted features into designed tool to let the user to see networks that can represent the relationship between documents, as well as easily access similar documents in the archive. In the network we demonstrated, particular nodes correspond to the documents itself. To assign an weighted edge between two documents in the network, the total number of shared individuals and keywords between documents are computed and edges are set based on a predetermined threshold value. This threshold has been found by manually tweaking both considering the speed at which the result is reflected on the application and average number of shared attributes.

5. Visual Exploratory Tool & Conclusion



İhtiyaçlarını karşılamak için Selene Kazası Kidona Nahiyesi Kazorus Köyü'nden Koraki Murad bin İbrahimin mülkleri ve tarlalarını rehin, babası İbrahim ağayı kefil göstererek Hanya Eytam Sandığı'ndan istikraz talebi. [3 Mayıs 1305]

Language: Osmanlı Türkçesi - Ottoman Turkish

Person Info: ['İbrahim Ağa', 'murad']

Time Info: ['mayıs']

Location Info: ['kazorus köyü', 'hanya']

Entities: ['ihtiyaç', 'selen', 'kaza', 'nahiye', 'mülk', 'tarla', 'rehin', 'baba', 'kefil', 'eytam', 'sandık', 'istikraz', 'talep']

Keywords: ['ihtiyaç', 'selen', 'nahiye', 'mülk', 'tarla', 'rehin', 'baba', 'kefil', 'istikraz']

Taxonomy Rank: 5.46367638965327

Kabilaki İbrahim Ağa'nın mektepte okutturulan bir nefer mahdumu için tahsis kılınan 1,5 adet sim mecidiyeyi Maarif Sandığına teslim ettiğine dair makbuz ilm-i haberi. [13 Ağustos 1310]

Person Info: ['İbrahim Ağa', 'nefer']

Time Info: ['ağustos']

Location Info: ['Bulunamadı']

Entities: ['mektep', 'mahdum', 'tahsis', 'adet', 'sim', 'mecidiye', 'maarif', 'sandık', 'teslim', 'makbuz', 'ilmi', 'haber']

Keywords: ['mahdum', 'tahsis', 'adet', 'sim', 'mecidiye', 'teslim']

Taxonomy Rank: 5.84273527793066

Müteveffa Pederi Moraitaki İbrahim Ağa'nın mirasına dair belgeler. [16 Ağustos 1309-16 Safer 1311]

Person Info: ['İbrahim Ağa', 'safer']

Time Info: ['ağustos']

Location Info: ['Bulunamadı']

Entities: ['müteveffa', 'peder', 'miras']

Keywords: ['peder', 'miras']

Taxonomy Rank: 5.50013758943313

Kızı Seyide'nin Akrotirli Çizmeci Hüseyin ile evlendirmek isteyip fakir oldukları için nikah masraflarının Şefkat Sandığı'ndan kendisine verilmesine dair merhum İbrahim Ağa'nın eşi Necibe Hanım'ın yazdığı dilekçe. [21 Mart 1305]

Person Info: ['İbrahim Ağa', 'necibe Hanım', 'hüseyin', 'şefkat']

Time Info: ['mart']

Location Info: ['Bulunamadı']

Entities: ['kız', 'çizme', 'fakir ol', 'nikah', 'masraf', 'sandık', 'kendi', 'merhum', 'eş', 'dilekçe']

City		Topic		Record Belongs To		Language		Type	
Chania	28	Personnel	0	Pious foundation	49	Osmanlı Türkçesi - Ottoman Turkish	49	Belge - Document	28
Lasithi	5	Orphans	14	Government	0				19
Iraklion	5	Finance	0	unknown	0			Defter - Register	2
Rethymno	6	Properties	13					Kitap - Book	0
Unspecified Cities	5	Families	4						

Figure 1: The designed user interface. a) Interactive map. b) Information box, document-specific attributes were placed on this box. c) The filtering panel. d) The document browser that integrated enables users to browse documents easily. e) Individual and keyword search menu.

The proposed visual exploratory tool is demonstrated to uncover hidden information and stories underlying documents, extracting the key attributes within the documents, and establishing a network between documents using various NLP and visualization methods. With developed tool we tested on the Archive of Waqfs of Crete users may observe people's specific actions in the documents and get a chance to witness a person living in the 1900s life. The application was only tested and designed on the Waqfs of Crete archive, but it needs to be tested on many more collections in different domains. In this way, the relationships between documents in different domains are more noticeable.

6. References

[1] Alfred V Aho and Margaret J Corasick. Efficient string matching: an aid to bibliographic search. *Communications of the ACM*, 18(6):333–340, 1975.

[2] Ahmet Afsin Akın and Mehmet Dündar Akın. Zemberek, an open source nlp framework for turkic languages. *Structure*, 10:1–5, 2007.

[3] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[4] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

[5] Stefan Schweter. Berturk - bert models for turkish, April 2020.