COLING 2016

The 26th International Conference on Computational Linguistics

Proceedings of COLING 2016: System Demonstrations

December 11-16, 2016
Osaka, Japan

TopoText: Interactive Digital Mapping of Literary Text

Randa El Khatib
Department of English
University of Victoria
British Columbia, Canada
khatib@uvic.ca

Julia El Zini
Department of Computer Science
American University of Beirut
Beirut, Lebanon
jwe0@4aub.edu.lb

David Wrisley
Department of English
American University of Beirut
Beirut, Lebanon
dw04@aub.edu.lb

Mohamad Jaber
Department of Computer Science
American University of Beirut
Beirut, Lebanon
mj54@aub.edu.lb

Shady Elbassuoni
Department of Computer Science
American University of Beirut
Beirut, Lebanon
se58@aub.edu.lb

Abstract

We demonstrate TopoText, an interactive tool for digital mapping of literary text. TopoText takes as input a literary piece of text such as a novel or a biography article and automatically extracts all place names in the text. The identified places are then geoparsed and displayed on an interactive map. TopoText calculates the number of times a place was mentioned in the text, which is then reflected on the map allowing the end-user to grasp the importance of the different places within the text. It also displays the most frequent words mentioned within a specified proximity of a place name in context or across the entire text. This can also be faceted according to part of speech tags. Finally, TopoText keeps the human in the loop by allowing the end-user to disambiguate places and to provide specific place annotations. All extracted information such as geolocations, place frequencies, as well as all user-provided annotations can be automatically exported as a CSV file that can be imported later by the same user or other users.

1 Overview

Spatial humanities researchers have long been utilizing digital mapping techniques in digital humanities. These visualizations are of interest because they uncover the internal spatial construction of works and often evoke arguments through patterns that may have otherwise eluded the reader through traditional close reading techniques. In this paper, we demonstrate TopoText, an interactive tool for digital mapping of literary texts in various languages such as English, German and Spanish. TopoText combines many NLP tools to provide the user with a comprehensive location-centered summary of a given text. First, it extracts all place names in the given text, which are then geoparsed and displayed on a map. In case of ambiguous place names, it provides a list of all the alternative locations that a place may be referring to (such as London England vs. London Ontario) with their geo-coordinates. The user can then pick the correct location the place is referring to, thereby introducing human intervention into automatic mapping in order to create the most accurate map possible. Moreover, TopoText calculates the number of times a place is mentioned in the text and plots it onto the map by having the points appear in different sizes. This results in a more meaningful map that is reflective of the content of the work rather than creating the illusion that all the places carry the same importance in the text.

TopoText goes beyond simply creating maps by instantly providing the user with word-place collocations that contextualize a place by offering the most recurring words related to it. These words can be faceted by part of speech tagging, which is useful because when tagging for nouns, for example, the resulting words would likely point to the general themes associated to this place. The word cloud can

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations,
pages 189–193, Osaka, Japan, December 11-17 2016.

either show the word-place collocation in a specific passage (such as, the most frequent words around London in a user-highlighted text) or in the entire work (the most frequent words around London in Charles Dickenss Oliver Twist, or in his entire corpus). These word-place collocations can show how spatial representation of various places changes over time and across authors.

TopoText also allows the user to annotate places and the annotations are directly displayed on that place on the map. The annotations can range all the way from merely extracting specific passages to personal responses and analysis. This subjective element is crucial to humanities work and really opens up the system to many fields and drastically increases its scope. Finally, TopoText exports all the automatically geoparsed data, including place names, their geo-coordinates, and other attributes such as the frequencies and annotations, into a separate file that can be reused on other mapping platforms. This functionality counters commercial GIS tools and aligns with the open-source values that lie at the core of digital humanities practices. TopoText is fully implemented and is available as a free open-source tool under the GNU General Public License at `https://github.com/rkhatib/topotext/tree/v2`.

2 System Architecture

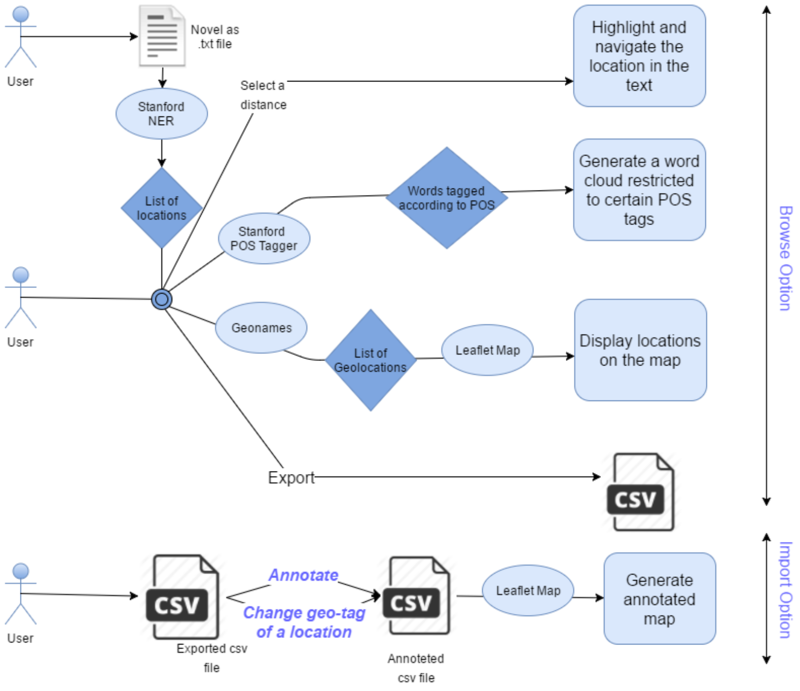


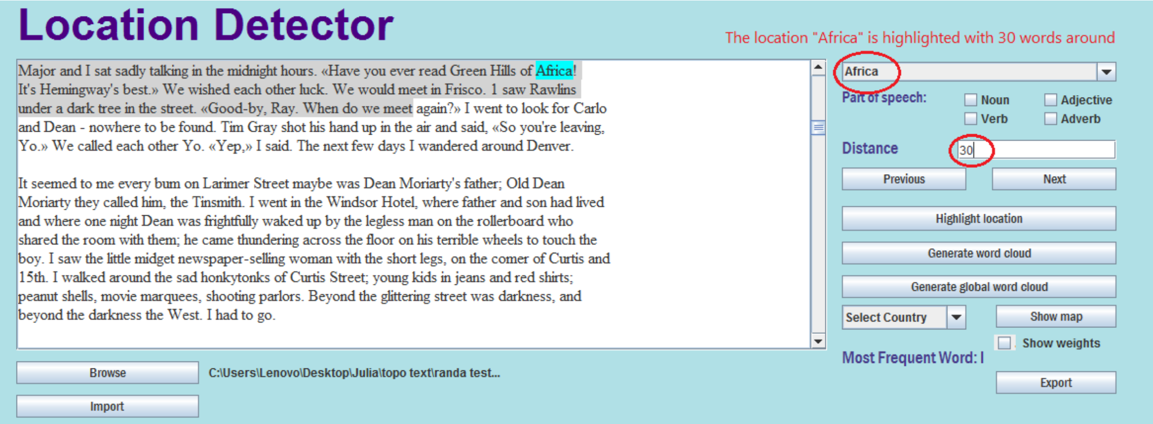
Figure 1: TopoText System Architecture

Figure 1 depicts the system architecture for TopoText. TopoText consists of two main components: a browse component and an import component. We will describe each separately next.

2.1 Browse Component

In the browse component, the user provides a piece of text, such as a novel or a biography article as a .txt file. The input text is then passed to the Stanford Named Entity Recognition (NER) Classifier (Finkel et al., 2005) which extracts all named entities in the text belonging to one of three classes: PERSON, ORGANIZATION and LOCATION. The list of places recognized (i.e. names tagged as LOCATION) are then extracted and provided back to the user, who has the following ways to explore these places:

- The user can select a place and track its occurrences in the text. TopoText highlights the user-selected place in the text and n words around it (distance in Figure 1). The user can also navigate to the next and previous occurrence of the place in the text. Figure 2 shows a snapshot of this feature.



Location Detector

The location "Africa" is highlighted with 30 words around

Major and I sat sadly talking in the midnight hours. «Have you ever read Green Hills of Africa! It's Hemingway's best.» We wished each other luck. We would meet in Frisco. 1 saw Rawlins under a dark tree in the street. «Good-by, Ray. When do we meet again?» I went to look for Carlo and Dean - nowhere to be found. Tim Gray shot his hand up in the air and said, «So you're leaving, Yo.» We called each other Yo. «Yep,» I said. The next few days I wandered around Denver.

It seemed to me every bum on Larimer Street maybe was Dean Moriarty's father; Old Dean Moriarty they called him, the Tinsmith. I went in the Windsor Hotel, where father and son had lived and where one night Dean was frightfully waked up by the legless man on the rollerboard who shared the room with them; he came thundering across the floor on his terrible wheels to touch the boy. I saw the little midget newspaper-selling woman with the short legs, on the comer of Curtis and 15th. I walked around the sad honkytonks of Curtis Street; young kids in jeans and red shirts; peanut shells, movie marquees, shooting parlors. Beyond the glittering street was darkness, and beyond the darkness the West. I had to go.

Africa

Part of speech: □ Noun □ Adjective □ Verb □ Adverb

Distance 30

□ Show weights

Most Frequent Word: I

C:\Users\Lenovo\Desktop\Julia\topo text\randa test...

Figure 2: The selected place Africa is highlighted in the provided text along with a context of 30 words.

- The user can generate a map of all extracted places using Leaflet, which is an open-source JavaScript library for interactive maps (lea, 2016). To be able to do this, the extracted places by the NER tool are first geoparsed through the GeoNames web service (geo, 2016). GeoNames is the largest open gazetteer that is historical, multilingual, and provides alternative spellings for place names. This differentiates TopoText from many other mapping tools that rely on modern gazetteers such as Google Maps, thus excluding historical place names, alternative spellings (a common historical occurrence before the standardization of spelling and in works translated from other languages), and works written in other languages. In contrast, by relying on GeoNames, TopoText is able to extract these variations and provides the end-user with a list of all the alternative locations that a place may be referring to. The user can then specify which location a place refers to before it is rendered on the map. Additionally, the user can also view the weight of each place on the map, where the weight of a place is the frequency by which it appeared in the text. That is, the most frequent places mentioned in the text will have bigger markers on the map. An example map is shown in Figure 3.

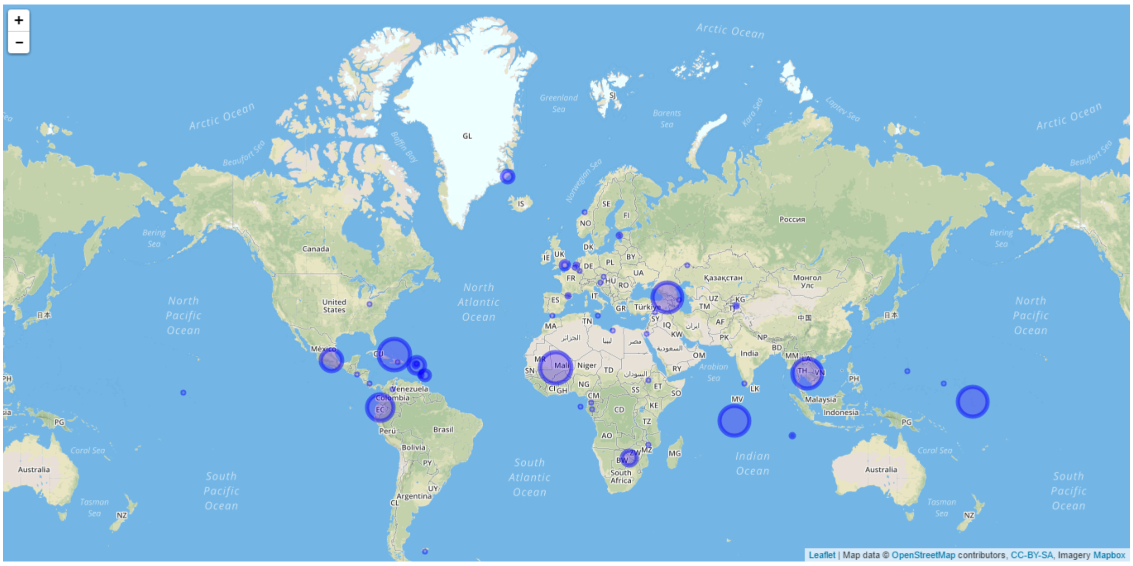


Figure 3: An example map generated by TopoText.

- The user can generate a Word Cloud for a given place. This word cloud can be a local word cloud from n words around the selected place or a global word cloud from all the words around the place in the text. In both cases, a web service provided by Wordle (wor, 2016) is used to generate the word cloud. Figure 4 shows a snapshot of an example word cloud. The user has also the option to select certain part of speech (POS) tags to include in the word cloud. The POS tags relevant in TopoText are: adjective, adverb, verb and noun. For instance, the user can exclude the adverbs from appearing in the word cloud, or select only the adjectives to have a better idea about the author's feelings about a certain place. To be able to do this, the text is POS tagged using the Stanford POS tagger (Toutanova and Manning, 2000).

- Finally, the user can also export all extracted information to a CSV file which contains every place extracted, their corresponding geo-coordinates, the country and an empty column where the user can provide annotations for places (see Import Component next). An example exported file is shown in Figure 5. This exported CSV file can be edited and reloaded by other users and can be re-used on other mapping platforms as well.

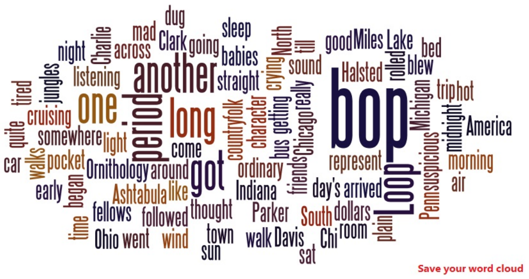
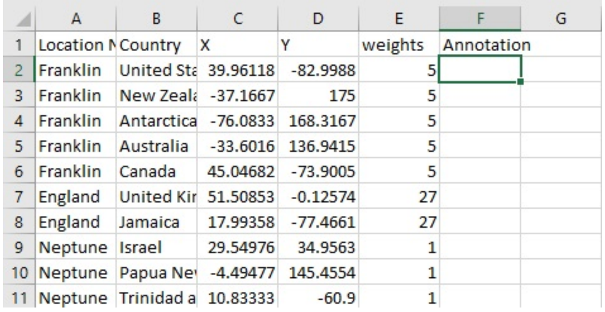


Figure 4: An example word cloud.



	A	B	C	D	E	F	G
1	Location N	Country	X	Y	weights	Annotation	
2	Franklin	United Sta	39.96118	-82.9988	5		
3	Franklin	New Zeala	-37.1667	175	5		
4	Franklin	Antarctica	-76.0833	168.3167	5		
5	Franklin	Australia	-33.6016	136.9415	5		
6	Franklin	Canada	45.04682	-73.9005	5		
7	England	United Kir	51.50853	-0.12574	27		
8	England	Jamaica	17.99358	-77.4661	27		
9	Neptune	Israel	29.54976	34.9563	1		
10	Neptune	Papua Nev	-4.49477	145.4554	1		
11	Neptune	Trinidad a	10.83333	-60.9	1		

Figure 5: An example exported CSV file.

2.2 Import Component

Using this component, the user can interact with a perviously exported CSV file. The user can, for instance, change the location a given place was mapped to by going through the list of alternative locations provided for that place, which are also included in the CSV file when it was exported. In addition, the user can provide any annotations for one or more places. Finally, the user can import this annotated CSV file, which will generate a map that displays all the places in the file along with their annotations.

3 Conclusion

We presented TopoText, an interactive digital mapping tool for literary texts in various language such as English, German and Spanish. TopoText utilizes various NLP tools and resources such as the Stanford NER, Leaflet, GeoNames and Wordle to provide users with a comprehensive location-centered summary of a given text. TopoText keeps the human in the loop by allowing the end-user to disambiguate places and to provide specific place annotations. TopoText is already available as an open source tool and has stirred wide interest in the digital humanities circuits.

In future work, we plan to extend TopoText to support other languages such as Arabic. This would involve finding or constructing appropriate gazetteers and NLP tools for such languages. We also plan to augment TopoText with an image retrieval component where place images are automatically retrieved from the Web and used to visually summarize the text. The images would be retrieved taken into consideration the context in which the place was mentioned, thus providing an accurate visual description of the place in context. Finally, we plan to develop a Web-based version of TopoText to increase its usability.

Acknowledgements

We would like to thank the American University of Beirut Research Board (URB) for their support in funding this project.

References

[Finkel et al.2005] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA. Association for Computational Linguistics.

[geo2016] 2016. The geonames geographical database. http://www.geonames.org/. Accessed: 2016-08-25.

[lea2016] 2016. Leaflet: an open-source javascript library for mobile-friendly interactive maps. http://leafletjs.com/. Accessed: 2016-08-25.

[Toutanova and Manning2000] Kristina Toutanova and Christopher D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13*, EMNLP '00, pages 63–70, Stroudsburg, PA, USA. Association for Computational Linguistics.

[wor2016] 2016. Wordle: Beautiful word clouds. http://www.wordle.net/. Accessed: 2016-08-25.