Identifying Historical Occupations in Text using HISCO: A User-Centric Approach [category: poster presentation]

Jaan Joosep Puusaag[1], Stacey Koolman[2], Guilherme Arashiro[3], Shuai Wang[4], Xander Wilcke[5]

{jaanjoosep23 | staceykoolman}@gmail.com, gui.arashiro@hotmail.com,
{shuai.wang,w.x.wilcke}@vu.nl

Vrije Universiteit Amsterdam

Introduction

To tackle the difficulties in the study of social mobility in history, the Historical International Standard of Classification of Occupations (HISCO) was created in 2002 [1]. As a multilingual thesaurus of historical occupations, HISCO provides a standard 5-digit code for a large number of occupational titles, enabling researchers to lookup and refer to historical occupations with ease. However, detecting and correctly mapping occupations in historical manuscripts to curated thesauri such as HISCO is still challenging and often done largely by hand. The same applies to many other facets that might be interesting to researchers and which, in principle, can be inferred from the texts, including social status, gender, and various relations.

This paper proposes a semi-automated workflow to support researchers in detecting and mapping occupational titles in their manuscripts to the HISCO thesaurus. This workflow employs natural language processing (NLP) for preprocessing and to perform named entity recognition (NER) on the result. We outline our workflow and demonstrate a prototype implementation on a dataset of Dutch biographical texts. Detected matches are proposed to the users via a simple interface. We publish our code as an open source tool for future reuse[6].

Approach

Our workflow starts by preprocessing the historical manuscripts and the choice of language. Entities in HISCO were trimmed accordingly. Then we parse the historical text to find string values that match HISCO entries (e.g. by computing their Levenshtein distance with potential matches). Upon detection, the title and the HISCO code are presented to the user in the format of a yes/no question, together with the context in which the title was detected. The user can then choose to accept the match, in which case the pair is written to a

[1] ORCID: 0000-0002-7013-8668
[2] ORCID: 0000-0002-9869-8405
[3] ORCID: 0000-0002-7508-4363
[4] ORCID: 0000-0002-1261-9930. Twitter username: shuai_wang_ai
[5] ORCID: 0000-0003-2415-8438.
[6] The code, the report, and the data corresponding to the use case are available in the repository https://github.com/DigitalHumanitiesMinorVU/project2023.

file for future use, or continue. These steps are illustrated in Figure 1.

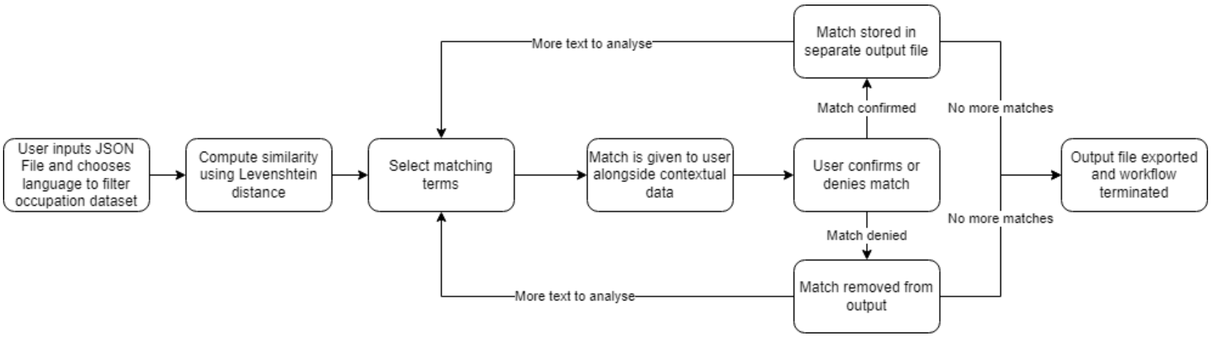
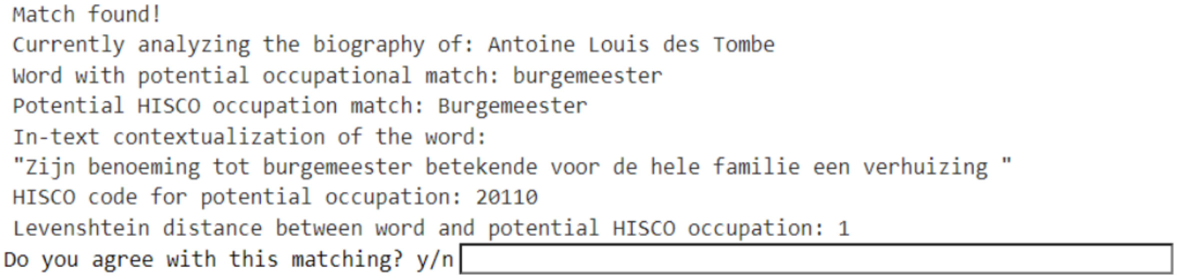

User inputs JSON File and chooses language to filter occupation dataset
Compute similarity using Levenshtein distance
Select matching terms
Match is given to user alongside contextual data
User confirms or denies match
Match confirmed
Match stored in separate output file
More text to analyse
Match denied
Match removed from output
More text to analyse
No more matches
No more matches
Output file exported and workflow terminated

Figure 1: An illustration of the workflow.

A common measure of string similarity is the Levenshtein distance, which computes the lexicographic distance between a pair of strings, as the number of steps necessary to alter one string to match another [3]. A low number has a higher chance of being a match. This number is rarely zero as there are other factors such as plurality and, in some languages, gender, that can produce false positives. Note that a distance of zero can be a mismatch, as words may have multiple meanings.

Another important factor is the context. The workflow prompts the user with not only the string match between the historical occupation and the matching HISCO concept, but also with contextual information, including the sentence that the word was taken from (Figure 2). This gives the user more information to help decide whether the match is valid.



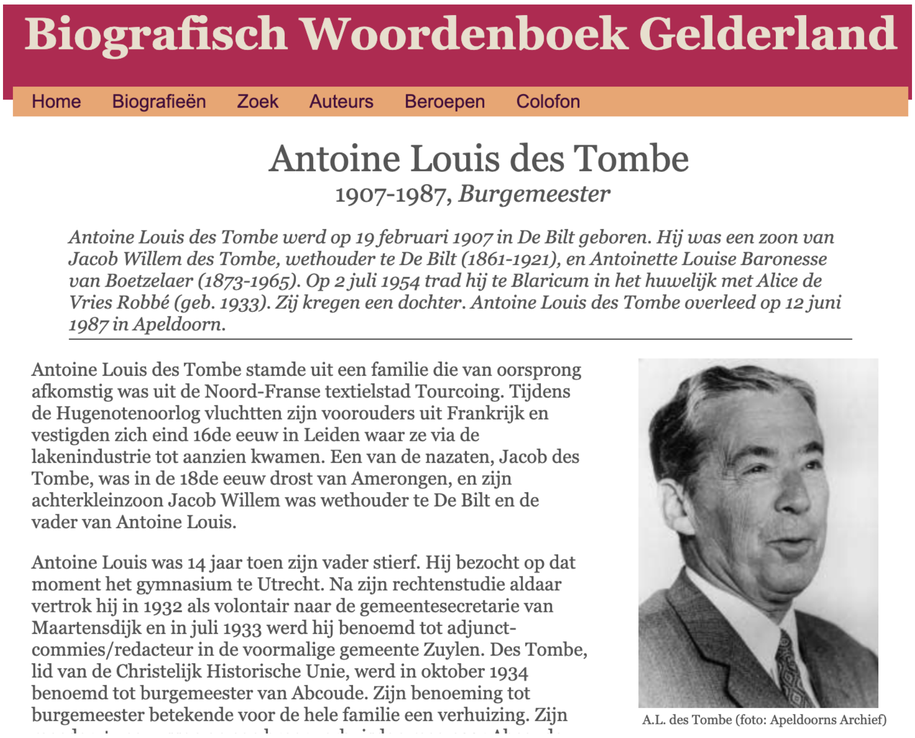
```
Match found!
Currently analyzing the biography of: Antoine Louis des Tombe
Word with potential occupational match: burgemeester
Potential HISCO occupation match: Burgemeester
In-text contextualization of the word:
"Zijn benoeming tot burgemeester betekende voor de hele familie een verhuizing "
HISCO code for potential occupation: 20110
Levenshtein distance between word and potential HISCO occupation: 1
Do you agree with this matching? y/n
```

Figure 2: Current workflow's output from any flagged potential match.

Use Case

To utilize the workflow described above, a python tool was coded, where the user inputs their data and chooses the language in which their data is written. The use case described in this paper involved the page of Antoine Louis de Tombe[7] in *Het Biografisch Woordenboek Gelderland* [2], which contains over 283 historical biographies from the second half of the 19th century. One of the words detected was "burgemeester", which has Levenshtein distance of 1 when evaluating against the HISCO entry "Burgemeester". This pair as well as the context of the term appears in were then returned to the user.

[7] https://www.biografischwoordenboekgelderland.nl/bio/4_Antoine_Louis_des_Tombe



Biografisch Woordenboek Gelderland

Antoine Louis des Tombe

1907-1987, *Burgemeester*

Antoine Louis des Tombe werd op 19 februari 1907 in De Bilt geboren. Hij was een zoon van Jacob Willem des Tombe, wethouder te De Bilt (1861-1921), en Antoinette Louise Baronesse van Boetzelaer (1873-1965). Op 2 juli 1954 trad hij te Blaricum in het huwelijk met Alice de Vries Robbé (geb. 1933). Zij kregen een dochter. Antoine Louis des Tombe overleed op 12 juni 1987 in Apeldoorn.

Antoine Louis des Tombe stamde uit een familie die van oorsprong afkomstig was uit de Noord-Franse textielstad Tourcoing. Tijdens de Hugenotenoorlog vluchtten zijn voorouders uit Frankrijk en vestigden zich eind 16de eeuw in Leiden waar ze via de lakenindustrie tot aanzien kwamen. Een van de nazaten, Jacob des Tombe, was in de 18de eeuw drost van Amerongen, en zijn achterkleinzoon Jacob Willem was wethouder te De Bilt en de vader van Antoine Louis.

Antoine Louis was 14 jaar toen zijn vader stierf. Hij bezocht op dat moment het gymnasium te Utrecht. Na zijn rechtenstudie aldaar vertrok hij in 1932 als volontair naar de gemeentesecretarie van Maartensdijk en in juli 1933 werd hij benoemd tot adjunct-commies/redacteur in de voormalige gemeente Zuylen. Des Tombe, lid van de Christelijk Historische Unie, werd in oktober 1934 benoemd tot burgemeester van Abcoude. Zijn benoeming tot burgemeester betekende voor de hele familie een verhuizing. Zijn

Figure 3: Biography of Antoine Louis des Tombe

Discussion

Various challenges were identified during this research. Firstly, it takes a considerable time to detect all matching occupations. This is likely caused by the large size of the vocabulary, and may be alleviated by implementing a smart lookup table. We noted repeated occupations throughout the text, which are presented to the user on a per-match basis. This burdens the researcher unnecessarily. The used metric of Levenshtein distances presents another challenge, as it will always flag tokens that are similar even if it is known that such a word means something else.

Conclusion and Future Work

In this paper, we designed a workflow that can match historical occupations in the text with HISCO codes, which allows users to make decisions based on the context of the sentence. While some issues remain regarding the specifics of the workflow, these problems can be partially resolved considering the program's customisable nature. For example, future users can choose to either use the Levenshtein distance, change its specifics, or use a different NLP method. Finally, other languages could be included in the evaluation.

Reference

1. van Leeuwen, M.H.D., Maas, I. and Miles, A.G. (2004) “Creating a historical international standard classification of occupations an exercise in multinational interdisciplinary cooperation” Historical Methods: A Journal of Quantitative and Interdisciplinary History, 37(4), pp. 186–197. Available at: https://doi.org/10.3200/hmts.37.4.186-197.
2. Huysman, I., Kloek, E. (no date). Biografisch Woordenboek Gelderland. Het Biografisch Woordenboek Gelderland. Available at: https://www.biografischwoordenboekgelderland.nl/. Last Accessed: February 1, 2023.
3. Yujian, L. and Bo, L. (2007) “A normalized Levenshtein distance metric” IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(6), pp. 1091–1095. Available at: https://doi.org/10.1109/tpami.2007.1078.