

Semi-automatic Entity Recognition Using Historical Occupational Vocabulary

Xander Wilcke (VU Amsterdam), Shuai Wang (UCDS@VU Amsterdam), Richard Zijdeman (IISG)
Bass on work of Jaan Joosep Puusaag, Stacey Koolman, Guilherme Arashiro, and Louisa Klüter

Part 1: Introduction to HISCO

Part 2: Semi-automatic entity recognition using HISCO

Part 3: Learning Male and Female Representation from Occupational Data

Part 1: Introduction to HISCO

From ISCO to HISCO

Let's start with some history

- The International Standard Classification of Occupations (ISCO)
 - Created by the International Labor Organisation (ILO)
 - Classification structure on labour and jobs
- Editions: ISCO58, ISCO68, ISCO88, and ISCO08
- ISCO68 contains 1,506 distinct occupations with 6,000 titles
- Each occupation has a unique 5-digit code, with multilingual labels
- ISCO68 was used to create HISCO as a complementary of ISCO
- ISCO is about contemporary occupations and jobs while HISCO captures that in the 19th and 20th century

Coding Scheme

HISCOde: adopted from ISCO

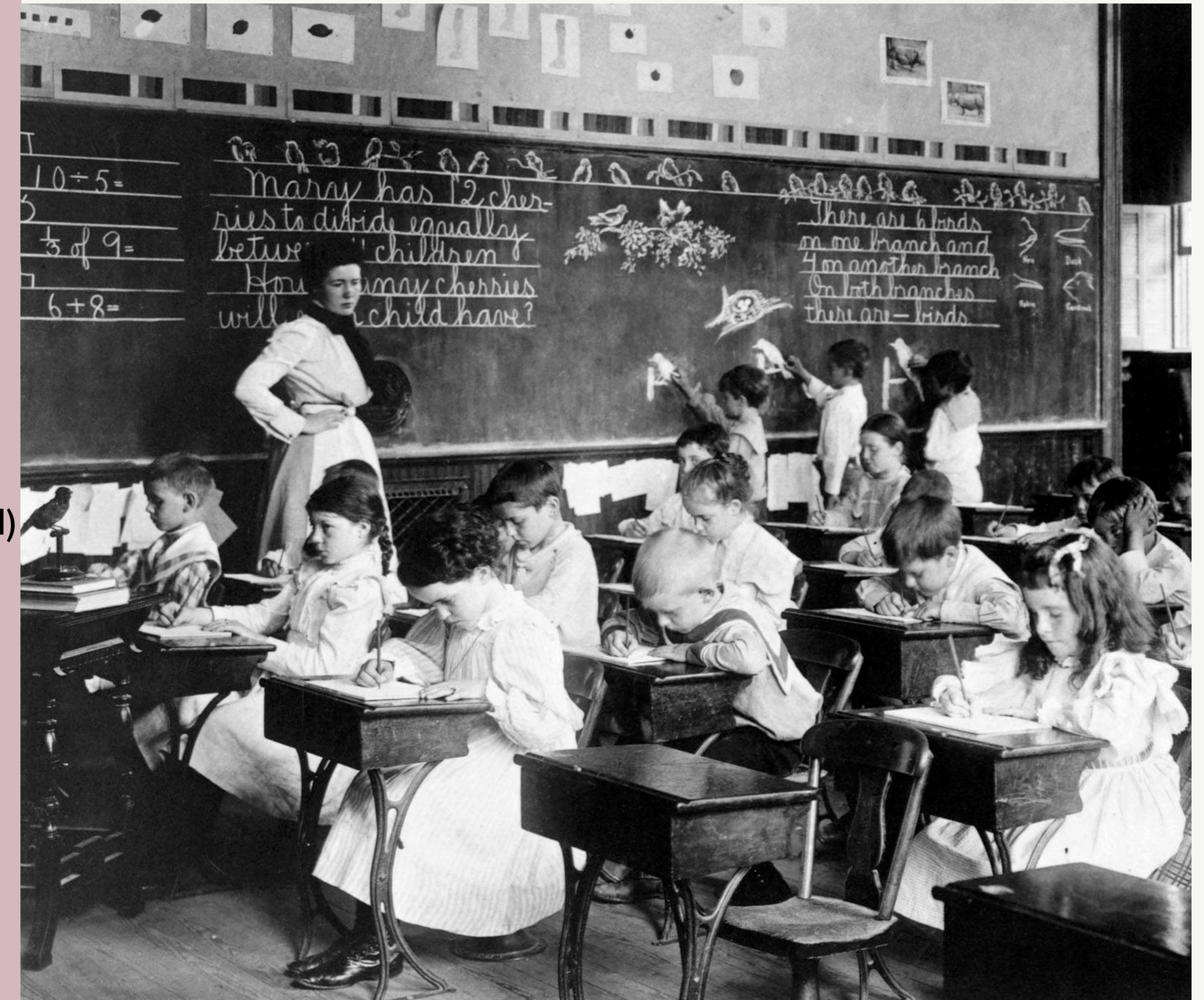
- Why a unique code?
- In the form of **x-xx.xx** (five-digit code)
 1. The first digit stands for **major** group
 2. The first two digits are for **minor** group
 3. The first three digits are for **unit** group
- The last two digits are about each occupation or the general occupation
 - 4: Sales Workers
 - 4-1: Working proprietors (wholesale and retail trade)
 - 4-10: Working proprietors (wholesale and retail trade)
 - 4-10.40: Working proprietors (Hiring Out)
 - 4.10-30: Working proprietors (Retail Trade)



<https://druid.datalegend.net/HistoryOfWork/historyOfWork-all-latest>

- 1-3 Teachers
- 1-30 Teachers, Level and subject Unknown
- 1-31 University and Higher Education Teachers
- 1-31.45 Teacher in Economics and Commerce (Third Level)
- 1-32.40 Social Studies Teacher (Second Level)
- 1-34.20 Kindergarten Teacher
- 1-35.20 Teacher of the Blind
- 1-39.40 Head Teacher

Example



Part2: Semi-automatic entity recognition using HISCO

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

Jaan Joosep Puusaag¹, Stacey Koolman², Guilherme Arashiro³, Shuai Wang⁴, Xander Wilcke⁵

{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 look up and refer to

DH Benelux'23

A project at the Digital Humanities Minor, VU Amsterdam

Entity Recognition in Historical Text

- Goal: associates occupational description with the corresponding HISCO code
- A semi-automatic approach enacts the validation by experts
- Data from pages of Het Biografisch Woordenboek Gelderland
- 283 historical biographies from the second half of the 19th century.

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.

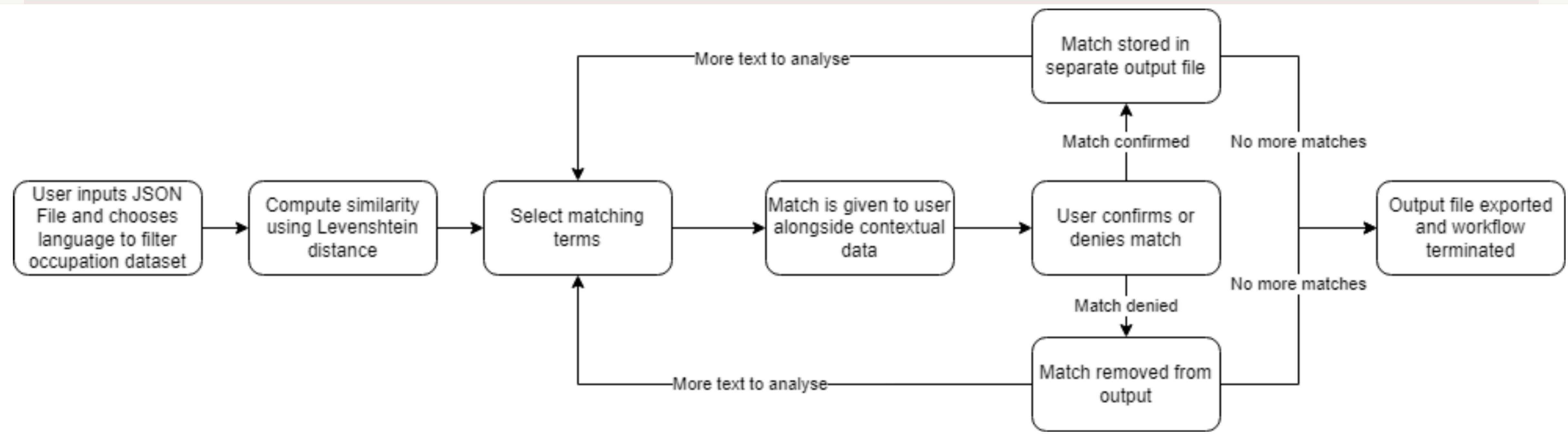
Antoine Louis des Tombe stamde uit een familie die van oorsprong afkomstig was uit de Noord-Franse textielstad Tourcoing. Tijdens de Hugenotoorlog vluchten zijn voorouders uit Frankrijk en vestigden zich eind 16de eeuw in Leiden waar ze via de lakenindustrie tot aanzien kwamen. Een van de nazaten, Jacob 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



A.L. des Tombe (foto: Apeldoorns Archief)

Workflow



Demo

One of the words detected was “burgemeester”

It has Levenshtein distance of 1 when evaluating against the HISCO entry “Burgemeester”

Printed on the screen for manual annotation

Record in a dictionary for future use

Burgemeester

HISCO code

2-01.10

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 Hugenootsoorlog 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



A.L. des Tombe (foto: Apeldoorns Archief)

Demo: user interface

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

Vrije Universiteit Amsterdam



Bachelor Thesis

**Learning Male and Female Embeddings from
Occupational Data and Computing
Male-to-Female Translation Vectors**

Author: Louisa Klüter (2710420)

1st supervisor: Xander Wilcke
2nd reader: Frank van Harmelen

*A thesis submitted in fulfillment of the requirements for
the VU Bachelor of Science degree in Computer Science*

July 17, 2023

Part3: Learning Male and Female Representation from Occupational Data

BSc Thesis Artificial Intelligence by Louisa Klüter

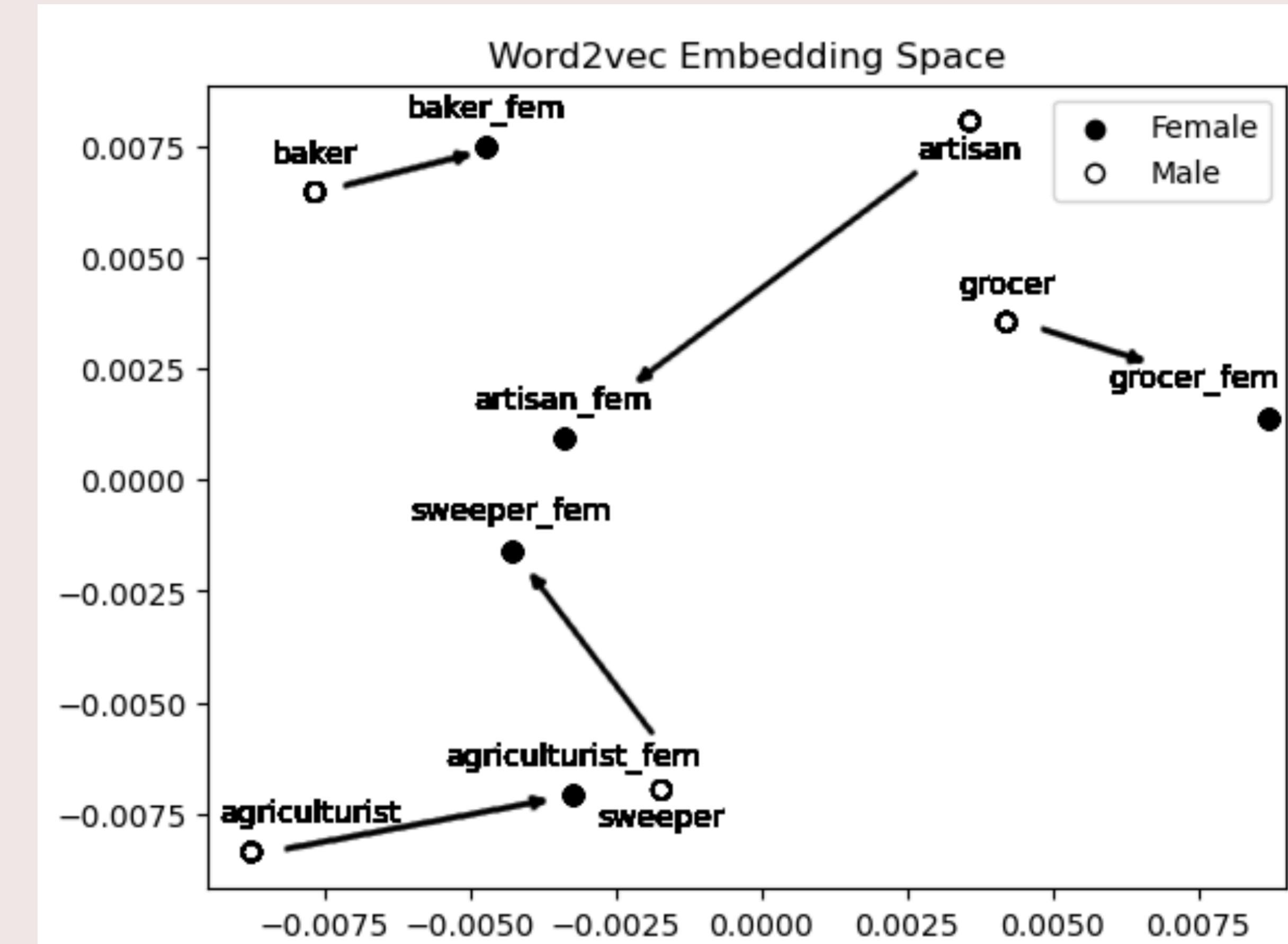
Supervisor: Xander Wilcke

HISCO lacks Female Occupations

- The labels of HISCO occupation are mostly male occupations
- Female is under represented (missing data points)
- Little examination about female historical occupations
- Extending HISCO with female historical occupation could enable new research
- The representation (as embedding) could be the input of future projects

Embedding

- Compute the embedding of occupation using HISCO's hierarchy (as a graph)
- The embedding of the female counter part
- The translation vector that converts the (male) occupation to its female counter part
- Word2Vec, Doc2Vec, GCN
- Focus on the English occupations



Summary

- Introduction to HISCO
- Semi-automatic entity recognition using HISCO
 - Demonstrating workflow of the use of HISCO for historical biographical data
 - An open source project that can ease your work or be integrated to future projects
- Extending HISCO for female occupations
 - Examination of female historical occupations that could enrich HISCO
 - Three different embeddings could be used for future projects
 - Basis for the future work on 1) gender equality in history 2) trends in jobs

Jacob Willem des Tombe, wethouder te De Bilt (1907-1987) in De Bilt geboren. Hij was een zoon van 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 Hugenotoornoorlog vluchten 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.



Thank you very much!

Shuai Wang (shuai.wang@vu.nl)
ORCID: 0000-0002-1261-9930

METHOD
Produce a workflow to establish entity
recognition between HISCO occupational
and user chosen biographical text

- 2 : Administrative and managerial workers
 - Major Group
 - Minor Group
- 2-1: Managers
- 2-0 : Legislative Officials And Government Administrators
 - Minor Group
- 2-01: Legislative Officials
- 2-12 : Production Managers
- 2-10 : Legislative Official (Burgemeester)

CORRECT MATCHES
STORED IN NEW FILE

USER CONFIRMATION
FOR MATCHES