The Incredibly Differentiated Labor Market Evidence from Job Offers

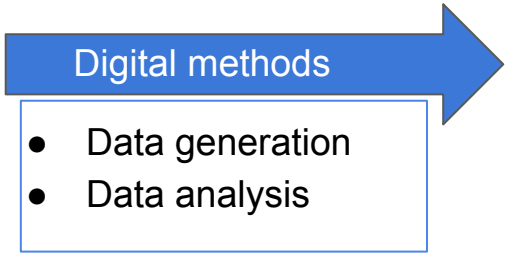
Nenad Pantelic, Vera Maria Charvat, Raven Adams, Saranya Balasubramanian, Jakob Sonnberger, Hanna Misera, Jörn Kleinert, Manfred Füllsack, Georg Vogeler

“What’s Past is Prologue”
NewsEye International Conference, 17th of March, 2021

Why Job Ads?

1) Job advertisements are an indicator for the role of labor in society

2) Job advertisements are sources for the change of perception of labor and job market organisation



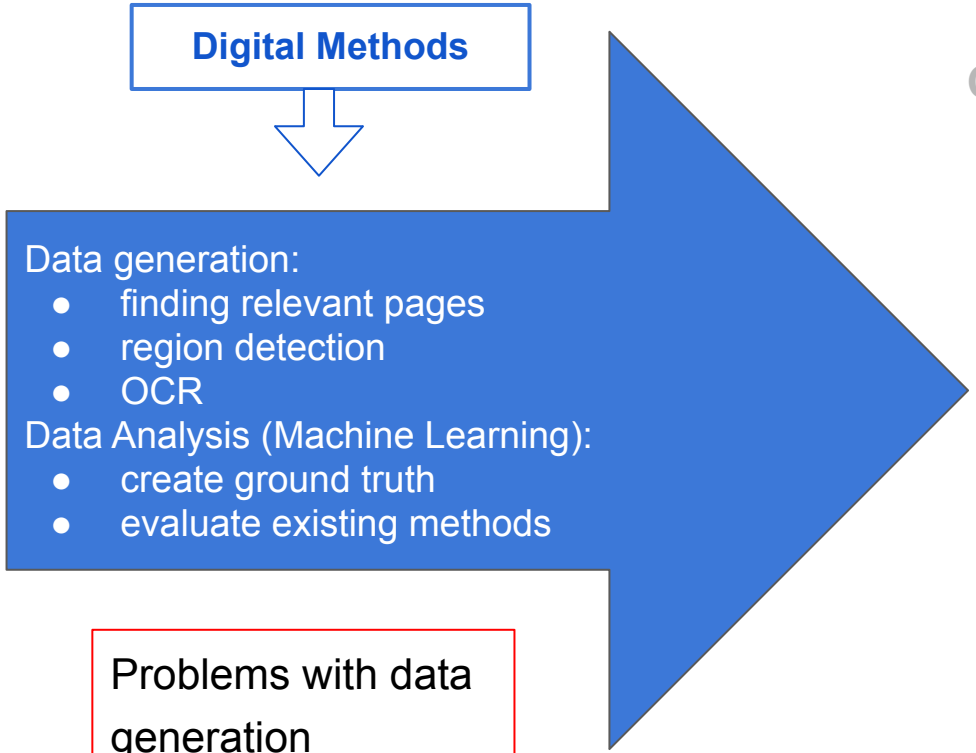
Goals:

- Creating a unique, reusable data set
- Illustrating the strong change in labor relationships in many dimensions
- Develop digital methods for text mining and text analyses

Methodology

1) Job advertisements are an indicator for the role of labor in society

2) Job advertisements are sources for the change of perception of labor and job market organisation



Goals:

- Creating a unique, reusable data set
- Illustrating the strong change in labor relationships in many dimensions
- Develop digital methods for text mining and text analyses

ANNO Repository

Österreichische Nationalbibliothek

ANNO
Historische Zeitungen und Zeitschriften

Über ANNO

- Was ist ANNO?
- 10 Jahre ANNO
- FAQs
- Suchen in ANNO
- Drucken aus ANNO
- Kooperationspartner
- Kontakt
- Impressum

ANNO - AustriaN Newspapers Online

Historische österreichische Zeitungen und Zeitschriften online.

Listen

- Alphabetische Liste der Zeitungen und Zeitschriften
- Jahresübersicht der Zeitungen und Zeitschriften
- Thematischer Einstieg

Suchen

- ANNO-Suche: Volltextsuche in historischen Zeitungen und Zeitschriften

Anno Corpus

1689 - 1947

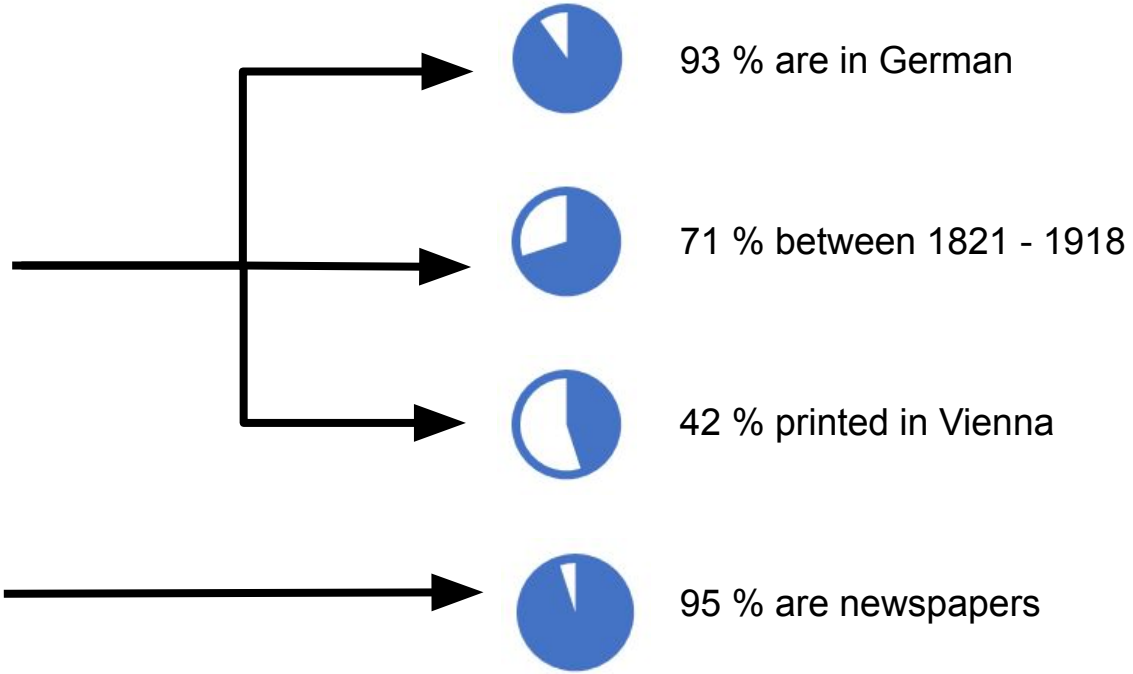
22 M. pages

=

1,5 M. editions

=

1000 titles



Our Corpus

Only **40** journals have more than **10.000** editions.

Title	years	editions
Wiener Zeitung	146	46.223
Innsbrucker Nachrichten	92	27.536
Neue Freie Presse	76	26.650
(Linzer) Tages-Post	80	23.291
Linzer Volksblatt	73	21.364
Neues Wiener Tagblatt	63	21.074
(Neuigkeits) Welt Blatt	70	20.773
Salzburger Volksblatt	72	19.844

Finding pages with job ads

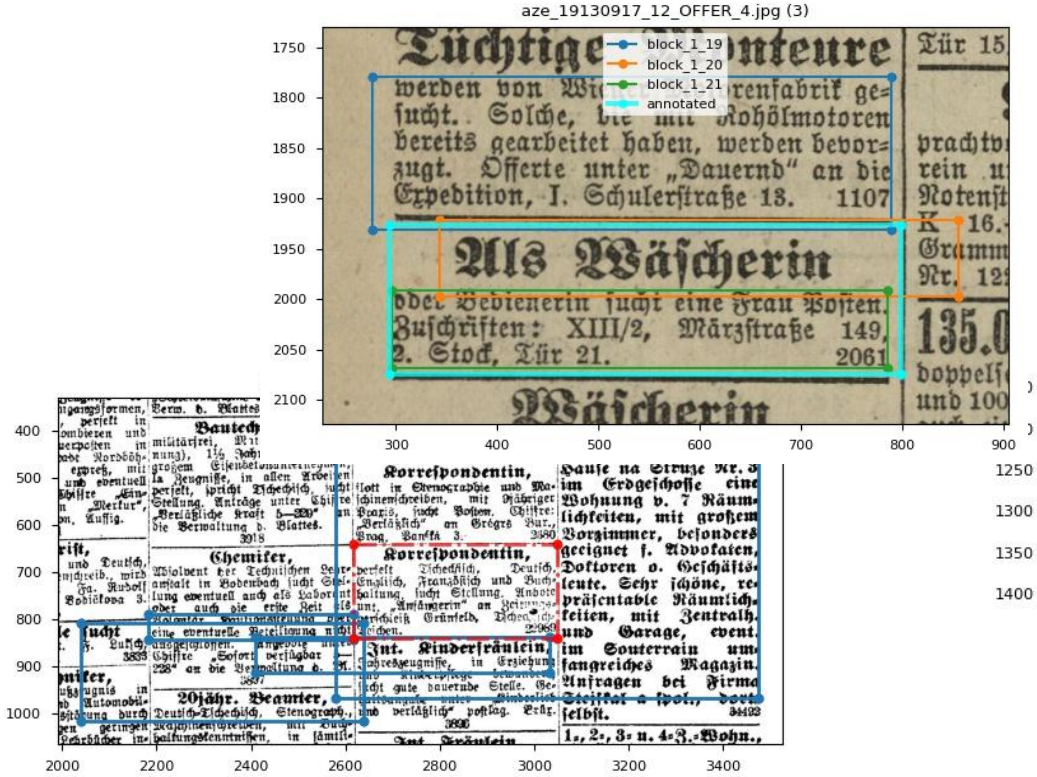
1. Step - Manual Selection
   - Browsing pages
   - Keyword search
   - Publication patterns

2. Step - Algorithmic Detection
   - Keyword search -> 70 % - 90 % accuracy in identifying pages with job ads
   - Problems:
     - No identification of job ads on the page: bad OCR, missing keywords
     - Findings due to coincidence

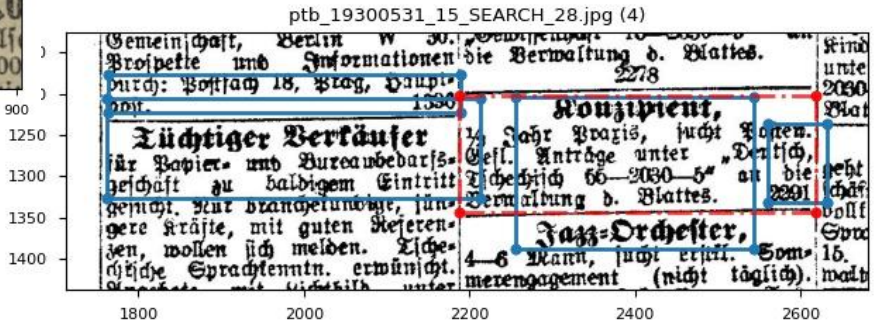
3. Step - Improve Detection
   - Train classifiers combining dirty OCR and images
   - To be done...

Layout detection: Tesseract standard model

Segmentation - Too Big



Segmentation - Broken regions



Layout detection: Transkribus

Transkribus experiment looks better, but …

ptb_19300531_15.jpg: text line crossing four columns

Variety of headings

e.g. Neue Freie Presse (1864 - 1939)

Offene Stellen.

Praktikant (Chr.), mit schöner Schrift, wird sofort acceptirt. Absolvirte Handelsschüler bevorzugt. Offerten unter **„Bankgeschäft Nr. 100"** an das Ank.-Bur. d. Bl. 433—2

Reisender für Mühlen- und Maschinenbau. Für eine Maschinen-Fabrik wird ein tüchtiger Reisender gesucht, der schon Verbindungen unter Müllern hat, für Oesterreich, Ungarn, Galizien; muß technisch gebildet sein. Antr. unt. **B. M.** an das Ank.-Bur. d. Bl. 511—2

Ein Bank-Commissions-Geschäft sucht hier und in der Provinz unter äußerst günstigen Bedingungen tüchtige Agenten zum Acquiriren von Commitenten. Ausführliche Offerten mit Referenzen sub **L. T.** an das Ankündigungs-Bur. d. Bl. 544—2

Agent gesucht für den Verkauf eines Geschäftes in Wien gegen gute Provision. Bekanntschaft in wohlhabenden Kreisen erforderlich. Zuschriften unter **„Agent Nr. 1000"** an das Ank.-Bur. d. Bl. 518—2

Vacanzen

(durch directe Aufträge).

Mehrere **Verkäufer** in Mode-Manufactur-Engros, **Reisende, Buchhalter** aus der Waarenbranche, **Commis** in Nürnberger-Kurzwaaren-Engros für **Wiener Großhandlungen u. 1 Reisender**, junger Mann (das Ausland zu bereisen, auch zu Einkassirungen) werden gesucht. Vorstellung sogleich; Offerten mit Referenzen vorläufig per Post mit Retourmarke an **Leo Binder**, Wien, Lazenburgerstraße **228**. (Gegründet im Jahre 1854.) Ferner ein **Vertrauensposten** mit weit über 10.000 fl. inclusive Tantième **pro anno bei** persönlicher Betheiligung **vorhanden. Ich intervenire** jetzt persönlich. 16631

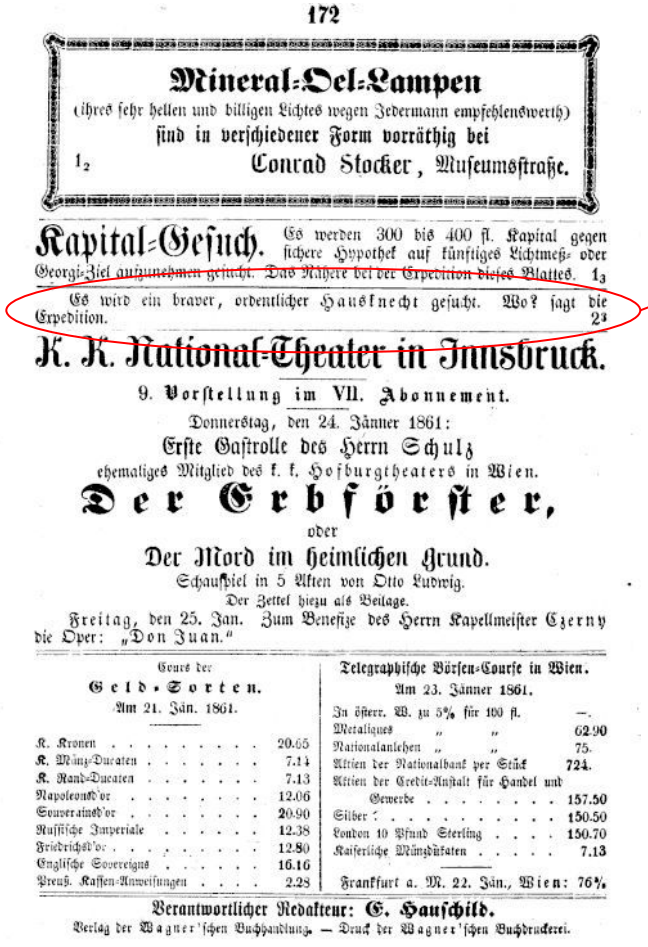
Seite 16 — Wien, Freitag — Neue Freie Presse. — 26. November 1880 — Nr. 5836

Kleine Anzeigen.

1880-11-26, p. 16

… or no headings at all

e.g. Innsbrucker Nachrichten (1854 - 1945)



Mineral-Oel-Lampen

(ihres sehr hellen und billigen Lichtes wegen Jedermann empfehlenswerth)
sind in verschiedener Form vorräthig bei
1/2 Conrad Stocker, Museumsstraße.

Kapital-Gesuch. Es werden 300 bis 400 fl. Kapital gegen sichere Hypothek auf künftiges Lichtmeß- oder Georgi-Ziel aufzunehmen gesucht. Das Nähere bei der Expedition dieses Blattes. 1/3

Es wird ein braver, ordentlicher Hausknecht gesucht. Wo? sagt die Expedition. 2/3

K. K. National-Theater in Innsbruck.

9. Vorstellung im VII. Abonnement.

Donnerstag, den 24. Jänner 1861:

Erste Gastrolle des Herrn Schulz
ehemaliges Mitglied des k. k. Hofburgtheaters in Wien.

Der Erbförster,

oder

Der Mord im heimlichen Grund.

Schauspiel in 5 Akten von Otto Ludwig.

Der Zettel hiezu als Beilage.

Freitag, den 25. Jan. Zum Benefize des Herrn Kapellmeister Czerny die Oper: „Don Juan."

Cours der
Geld-Sorten.
Am 21. Jän. 1861.

K. Kronen	20.65
K. Münz-Ducaten	7.14
K. Rand-Ducaten	7.13
Napoleonsd'or	12.06
Souverainsd'or	20.90
Russische Imperiale	12.38
Friedrichsd'or	12.80
Englische Sovereigns	16.16
Preuß. Kassen-Anweisungen	2.28

Telegraphische Börsen-Course in Wien.
Am 23. Jänner 1861.

In österr. W. zu 5% für 100 fl.	—.
Metaliques „ „	62.90
Nationalanlehen „ „	75.
Aktien der Nationalbank per Stück	724.
Aktien der Credit-Anstalt für Handel und Gewerbe	157.50
Silber	150.50
London 10 Pfund Sterling	150.70
Kaiserliche Münzdukaten	7.13

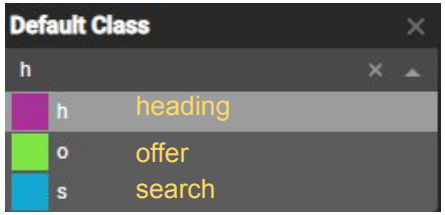
Frankfurt a. M. 22. Jän., Wien: 76¾

Verantwortlicher Redakteur: E. Hauschild.
Verlag der Wagner'schen Buchhandlung. — Druck der Wagner'schen Buchdruckerei.

Es wird ein braver, ordentlicher Hausknecht gesucht. Wo? sagt die Expedition. 2/3

1861-01-24, p. 8

Working with SuperAnnotate




```
{
    "instances": [
        {
            "type": "bbox",
            "classId": 3,
            "probability": 100,
            "points": {
                "x1": 192.6,
                "x2": 1069.7,
                "y1": 1891.7,
                "y2": 2152.3
            },
            "groupId": 0,
            "pointLabels": {},
            "locked": false,
            "visible": true,
            "attributes": []
        }
```

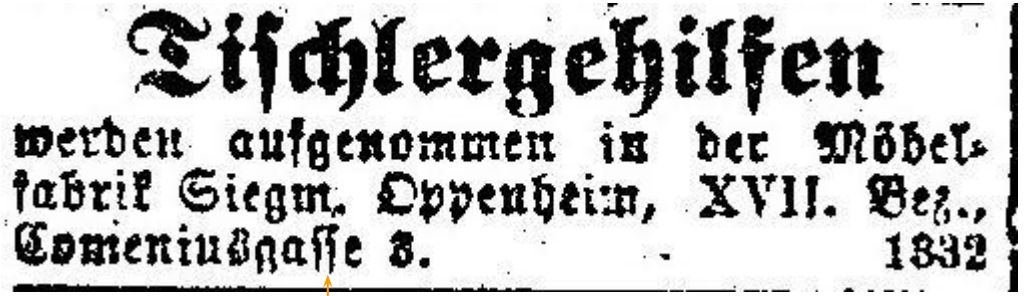

e.g. Innsbrucker Nachrichten

1896-05-30, p. 8

Further OCR attempts: Tesseract

Basic German and ONB trained models in Tesseract → works best

Fraktur!



Long s - ſ

Basic German model

Tiichlergebilfen
werben. aufgenommen im der Möbel:
tabrit Siegm, Oppenheim, XVII. Bez,
Eomeninsnafle 2. . 1832 |

ONB trained model

Tiſchlergehilfen
werden aufgenommen in der Möbel⸗
fabrik Siegm. Oppenheim, XVII. Bez.,
Comeniugaffe 3. 1832

Text Mining as OCR support

Cleaning messy OCR results with FASTtext + Jaro-Winkler similarity (JW):

Assumptions:

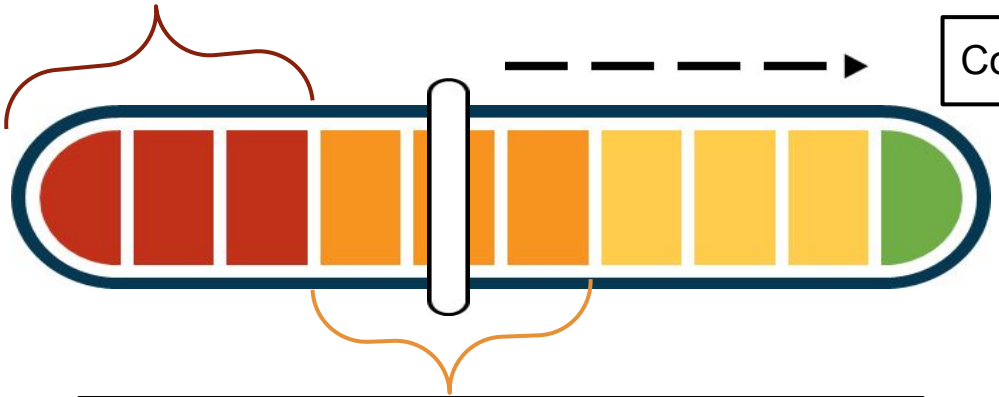
a. mis-recognized words appear in the same context as the correct words
b. correctly spelled words are dominant in the word embeddings
c. FASTtexts character based embeddings further aid in finding correct words

Process:

1. crosscheck OCR results with german dictionary → get misspelled words
2. correct single-letter mistakes with high JW threshold (0.98)
3. generate word embeddings with FASTtext
4. calculate vector similarity for misspelled words → apply JW to suggestions
5. insert corrections into corpus → use cleaned corpus for classification task

Conclusion

Creation of a new, unique text corpus that is useful for digital research on the labor market and the perception of work ***is*** possible:
e.g. from “Mädchen mit besten Zeugnissen” to “praktischer Ökonom / Praktikant mit schöner Schrift”



Complete data set

- Go beyond the standard Tesseract OCR/layout detection
- Word embeddings + edit distance is an interesting quality enhancement method

Thank you for your attention!

UNI GRAZ

HFDT

ÖAW

AUSTRIAN ACADEMY OF SCIENCES