



**University of
Zurich^{UZH}**



**Institute of Forrester
Landscape and Snow
Research^{WSL}**

SHAX: A Semantic Historical Archive eXplorer

Michael Feldman, Shen Gao, Marc Novel, Katerina
Papaioannou and Abraham Bernstein



Dynamic and Distributed
Information Systems



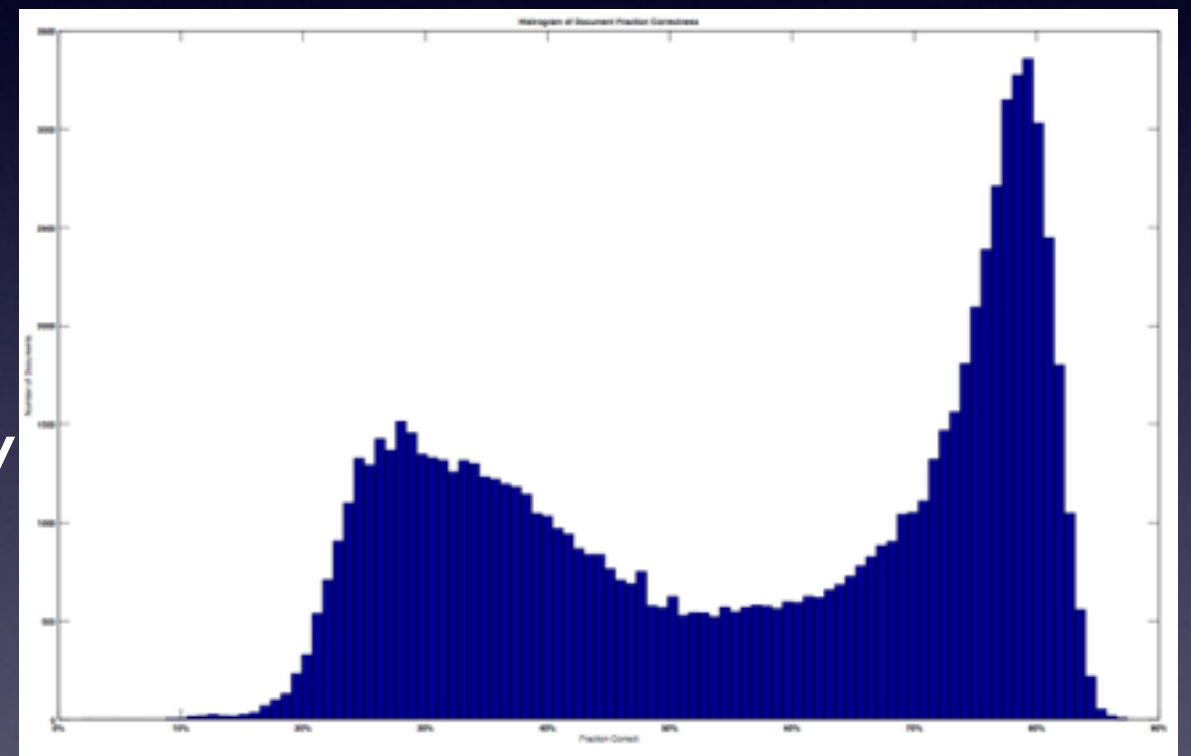
University of
Zurich^{UZH}

Background

- Dataset: Neue Zürcher Zeitung
- Years: 1910 - 1920
- Number of Files Processed: 111906 (OCR from German)
- Size:
 - Scanned PDF: 203 GB
 - OCR XML: 1.74 GB

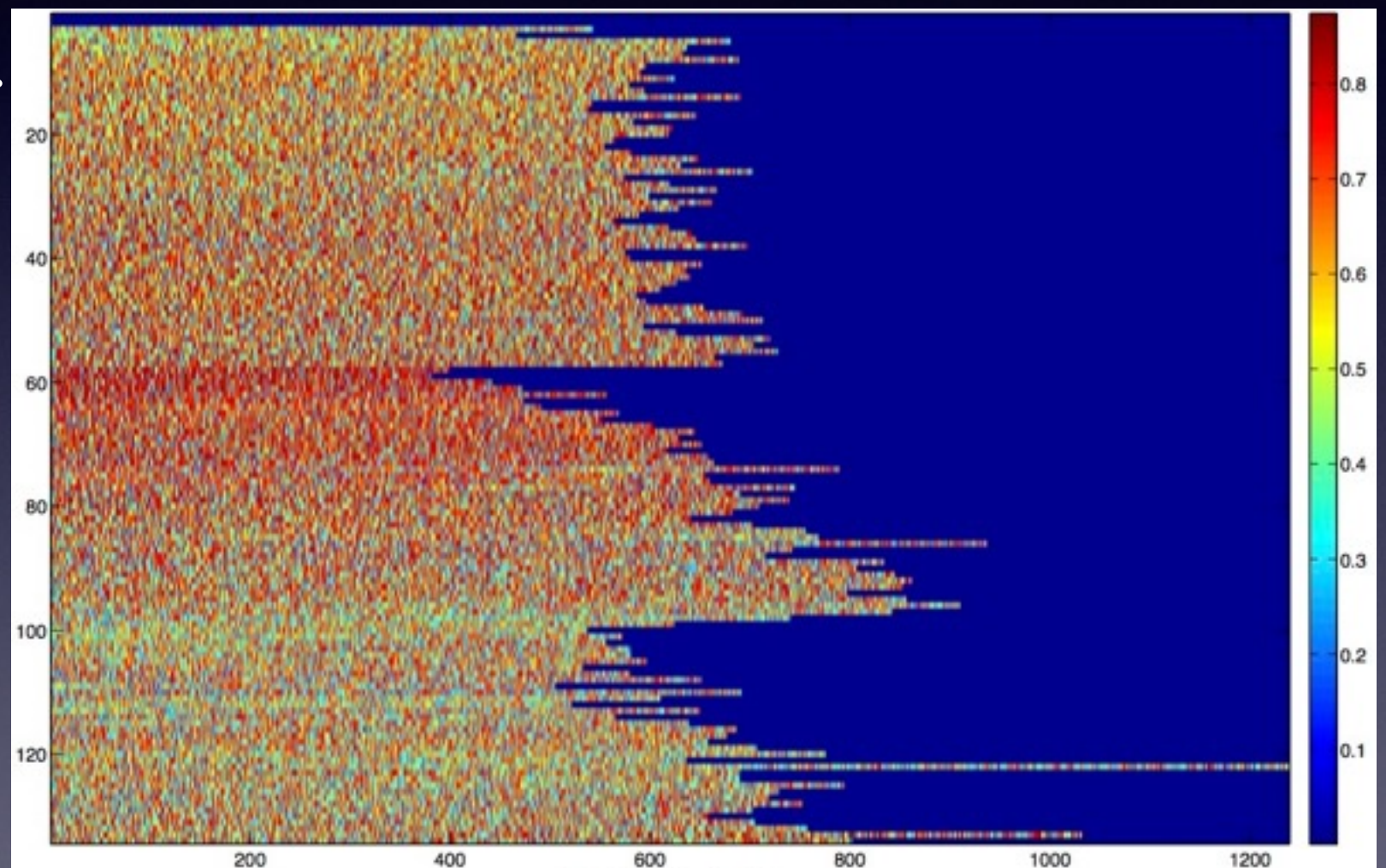
Histogram

- A histogram of all the scanned documents ranging from 1910-1920
- Bi-modal distribution:
 - first peak at 80% correctly detected words.
 - second peak at 38% correctly detected words.

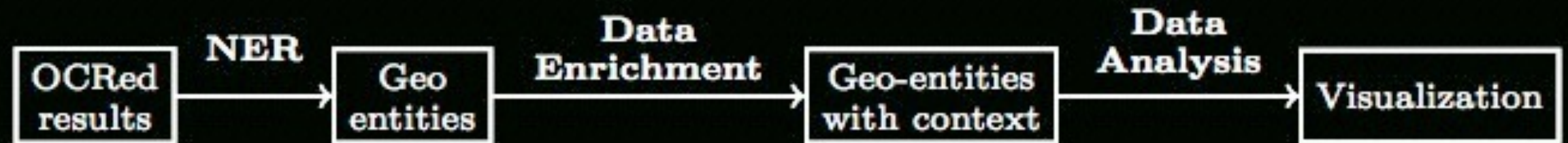


Correctness by Month

- Heatmap for detected words by the spell checker.
- Recognition rate graph
 - One row = one month (pages count between 300 and 1200)
 - One page = one point on the graph.
 - Blue Points = no page
 - Correlates with the bi-modal distribution
 - red: 80%
 - green: 30%



Pipeline



- Dbpedia-Spotlight as NER
- Linking to DBpedia
- Get the right Confidence and Support Values
- Geocoding
- Reverse Geocoder
 - Geonames
 - Google Map
- Aggregating by
 - Country
 - Month
- Normalized by
 - Population
 - Relative nearness

Interactive Visualization

- Function 1: Country Mentions over time
- Function 2: Linking countries, issues and historical event
- Function 3: Entity Mentions over time

Demo

Future Work

- Reference Corpus via Mechanical Turk
- Extracting learning the pictorial vector feature of the different words
- Extracting: Advertising, Stockmarket data
- Generalise the work to 250 years



Future Work

- Reference Corpus via Mechanical Turk
- Extracting learning the pictorial vector feature of the different words
- Extracting: Advertising, Stockmarket data
- Generalise the work to 250 years



Future Work

- Reference Corpus via Mechanical Turk
- Extracting learning the pictorial vector feature of the different words
- Extracting: Advertising, Stockmarket data
- Generalise the work to 250 years

	Samstag 27. Januar	29. Januar	30. Januar	31. Januar	1. Februar	2. Februar
Schweiz. Bankverein	805 $\frac{1}{2}$	805 c.	808 $\frac{1}{2}$	806 c.	807 $\frac{1}{2}$	808 $\frac{1}{2}$
Elektrobank	1955 $\frac{1}{2}$	1958 c.	1965 $\frac{1}{2}$	1958 c.	1960 $\frac{1}{2}$	1958 $\frac{1}{2}$
Motor (alt)	711 $\frac{1}{2}$	711 c.	711 $\frac{1}{2}$	710 $\frac{1}{2}$	710 $\frac{1}{2}$	705 $\frac{1}{2}$
Brown Boveri & Co.	1815 $\frac{1}{2}$	1820 $\frac{1}{2}$	1819 $\frac{1}{2}$	—	—	—
Banca Commere. Italiana	875 $\frac{1}{2}$	878 c.	878 $\frac{1}{2}$	869 $\frac{1}{2}$	871 $\frac{1}{2}$	873 $\frac{1}{2}$
Credito Italiano	—	—	580 $\frac{1}{2}$	580 $\frac{1}{2}$	579 $\frac{1}{2}$	—
Società Bancaria Italiana	108 $\frac{1}{2}$	107 $\frac{1}{2}$	107 $\frac{1}{2}$	107 $\frac{1}{2}$	—	107 $\frac{1}{2}$
Goerz	23 $\frac{1}{2}$	23 $\frac{1}{2}$	—	23 $\frac{1}{2}$	23 $\frac{1}{2}$	23 $\frac{1}{2}$
General Mining	27 $\frac{1}{2}$	—	26 $\frac{3}{4}$	—	26 $\frac{3}{4}$	26 $\frac{3}{4}$
Franco-Suisse Electr.	503 $\frac{1}{2}$	—	—	505 $\frac{1}{2}$	506 $\frac{1}{2}$	511 $\frac{1}{2}$
Nestlé-Chamer Milch	1747 $\frac{1}{2}$	1752 $\frac{1}{2}$	1750 $\frac{1}{2}$	—	1745 $\frac{1}{2}$	174 $\frac{1}{2}$
Aluminium Neuhäusen	2282 $\frac{1}{2}$	2300 c.	2314 $\frac{1}{2}$	2320 $\frac{1}{2}$	2323 $\frac{1}{2}$	2330 $\frac{1}{2}$
Officine Elettr. Genovesi	—	499 $\frac{1}{2}$	500 $\frac{1}{2}$	499 c.	—	498 c.
D.-Nebersee. Electr. Gesellsch.	2303 $\frac{1}{2}$	2315 $\frac{1}{2}$	2310 $\frac{1}{2}$	2302 $\frac{1}{2}$	2308 $\frac{1}{2}$	2310 $\frac{1}{2}$
G. f. elektr. Beleucht.-Priorität	2330 $\frac{1}{2}$	2340 $\frac{1}{2}$	2354 $\frac{1}{2}$	2360 $\frac{1}{2}$	236 $\frac{1}{2}$	2360 $\frac{1}{2}$
Bank Petersburg (Stamm)	1843 $\frac{1}{2}$	1869 $\frac{1}{2}$	1875 $\frac{1}{2}$	1873 $\frac{1}{2}$	1875 $\frac{1}{2}$	1878 $\frac{1}{2}$
Baltimore & Ohio	545 $\frac{1}{2}$	547 $\frac{1}{2}$	545 c.	549 $\frac{1}{2}$	546 $\frac{1}{2}$	546 $\frac{1}{2}$
National Railways of Mex. II	178 $\frac{1}{2}$	181 $\frac{1}{2}$	184 $\frac{1}{2}$	183 $\frac{1}{2}$	181 $\frac{1}{2}$	177 $\frac{1}{2}$
Schweiz. Anstalt. f. d. Ind.-Bef.	—	520 $\frac{1}{2}$	520 $\frac{1}{2}$	520 $\frac{1}{2}$	—	520 $\frac{1}{2}$