# Geographic visualization of a Swedish encyclopedia with a classifier based approach

Students: Axel Ahlin and Alfred Myrne Blåder
Supervisor: Pierre Nugues


LUNDS UNIVERSITET

# Problem

- Encyclopedia (Nordisk Familjebok, 2a uppl. 1904-1926)
- Some of the entries are locations
- How do we know which ones?
- Can we map these locations?
- What can the resulting map tell us about its time and historical context?

# Process

- Classify locations
- For every location, retrieve a corresponding Wikidata entry (QID)
- For every QID, retrieve coordinates
- Map coordinates

- We used two methods:
  - first, pretrained model
  - second, our model

# Initial classification - KBBERT NER

- KBBERT = Kungliga Biblioteket’s Bidirectional Encoder Representations from Transformers
- Great understanding of context
- Specifically the NER (Named Entity Recognition) version
- Model pretrained on 200 million sentences, sourced from diverse channels (books, news articles, government publications, Swedish Wikipedia and internet forums.)
- Used directly for annotation in this phase
- Performance was underwhelming

# Our classifier

- Another KBBERT model but not fine-tuned for NER
- Trained on manually annotated dataset with ~300 entries consisting of “sentence” and “is_location”
- The resulting hidden states are used as features and fitted with logistic regression
- The classifier is then used to predict and annotate the whole encyclopedia

# Background: QID - Wikidata entry ID

- Wikidata is a separate service from Wikipedia, containing data points used in Wikipedia, Wikimedia, et.c.
- People, places, animals, and more.
- Location entries have a coordinate property (P625) that can be queried

# QID retriever

- Every entry that is annotated as a location is mapped to a QID
- Each such entry has a corresponding first word
- This word is used in an HTTP request to wikidata, returning (up to) five candidate wikidata entries
- The candidate are then ranked by cosine similarity between their description and the dictionary entry text
- The QID of the best candidate is saved and mapped to the dictionary entry
- QID is used in a SPARQL query to retrieve coordinates (P625)

# Pitfalls - QID retriever

## Iowa (Q99670857)

the federated state of Iowa in the USA

# Pitfalls - QID retriever

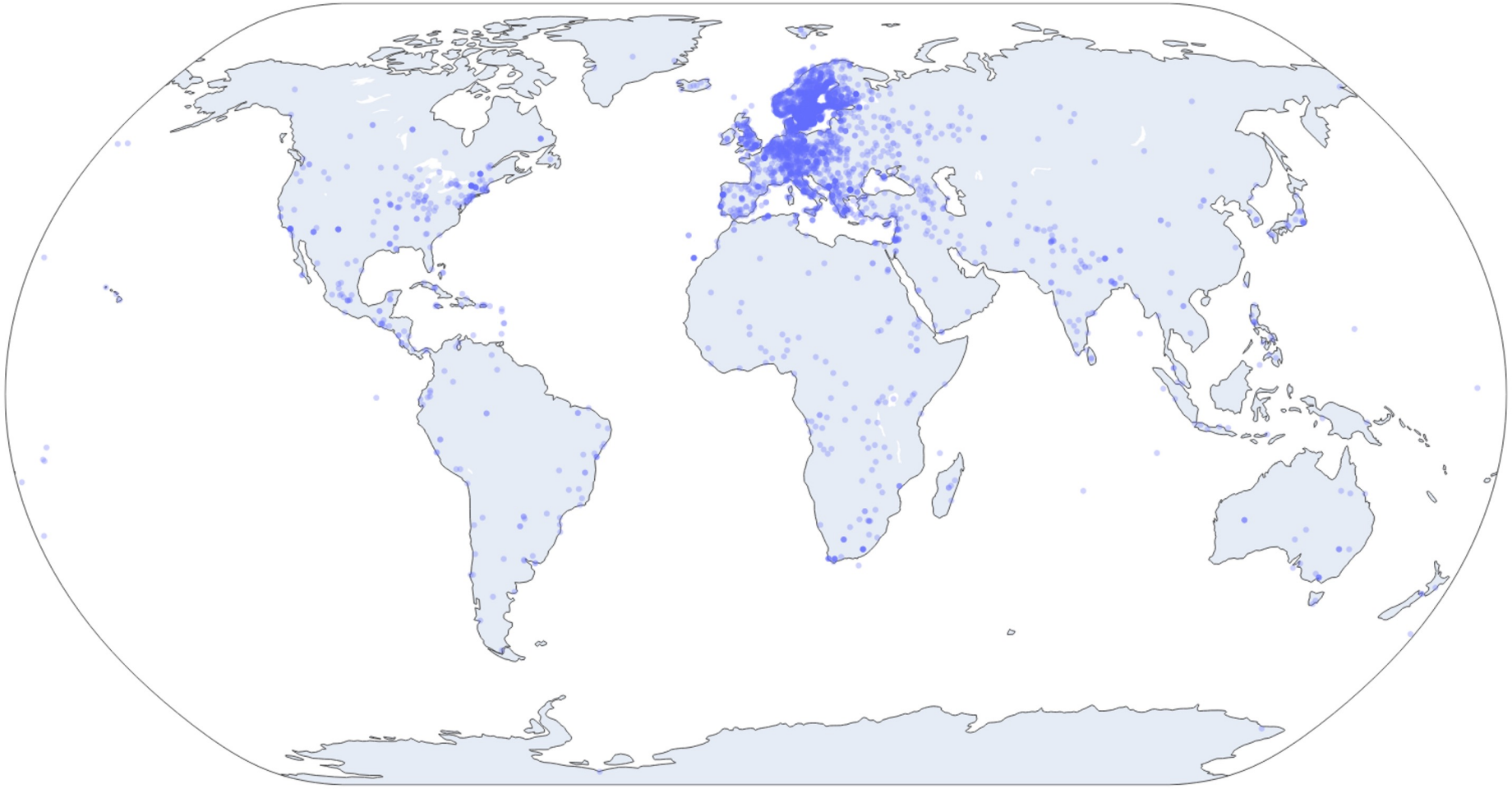
## Iowa (Q99670857)

the federated state of Iowa in the USA *as depicted in Star Trek*

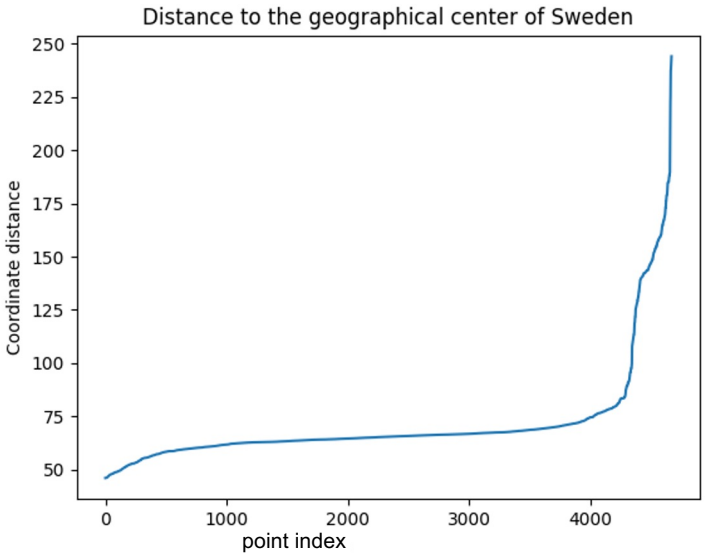
# Results

- 52310 entries
- 11325 locations  => ~22 percent of the entries are locations
- 8224 valid coordinates extracted

Difference between locations and valid coordinates may be attributed to errors in the SPARQL query, as well as false positive location annotations (= no coordinate feature) and QID retrieval errors.



# Results


Distance to the geographical center of Sweden
Coordinate distance
250
225
200
175
150
125
100
75
50
0
1000
2000
3000
4000
point index

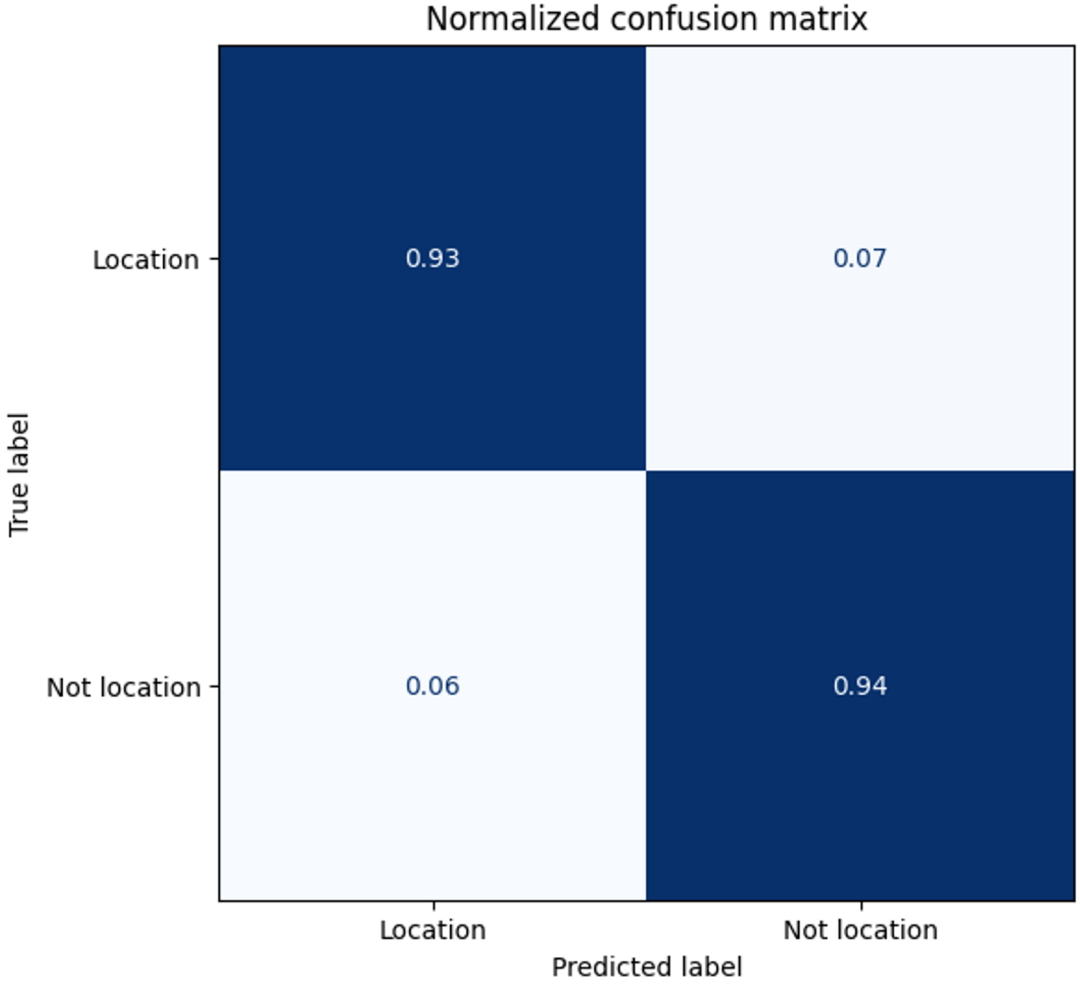
# Classifier evaluation

Test set consisting of ~200 manually annotated entries, previously unseen to the classifier.

Precision: 0.939

Recall: 0.93

F1-score: 0.9347



# Areas of improvement

- Preparing and reprocessing: scraping, splitting, clean up dirty entries, increase context size for training
- Evaluate and improve QID retrieval further
- Improve classifier further (update hyperparameters of the pretrained model)
- Crowdsourcing for more annotated data and/or verification

# Conclusions and further research

- Heavy focus on Europe and especially Sweden, which is to be expected

- Further research:
  - compare with later encyclopedias
  - compare with encyclopedias from other countries
  - include an interactive time axis