German wings challenge

Kai Chen

Github repo: https://github.com/ck-unifr/german-wings-airline-challenge

Resume

- Oct 2018 - present, Data Science Manager, Unitymedia, Germany
- May 2017 - Oct 2018, Senior Data Scientist, AGT International, Germany
- Sep 2012 - Jun 2017, PhD in CS, University of Fribourg, Switzerland
- Jun 2015 - Sep 2015, Visiting PhD, Chinese Academy of Sciences, China
- Sep 2009 - Sep 2012, Master in CS, University of Fribourg, Switzerland
- Oct 2005 - Feb 2008, Bachelor in CS, University of Applied Science, Switzerland

Github: github.com/ck-unifr
Google scholar: kai chen unifr
Linkedin: linkedin.com/in/kai-chen-29503288/

Use Cases

- Case 1: Sentiment Analysis
  - Predict a review is positive or negative

- Case 2: Topic Modeling
  - Find topics in the reviews

Methodology

- **Data Loading**
  - Load data from a txt file and save the data into a pandas dataframe
- **EDA (Exploratory Data Analysis)**
  - Plot distribution of variables
  - Show relationship between the variables
  - Text analysis: Plot word frequencies, Topic modelling with LDA
- **Feature Engineering**
  - TFIDF (term frequency–inverse document frequency)
  - Count Features
  - Dimensionality reduction using truncated SVD
  - Word Embedding: GloVe
- **Evaluation Metrics**
  - Cross-Entropy Loss
  - Precision and Recall
  - ROC (Receiver Operating Characteristics)
- **Modelling**
  - Logistic regression, Gradient boost machine, Deep Learning
- **Error Analysis**

Notebook

Word Embedding

The slides are taken from, *Sequence Models, Deep Learning Specialization, Coursera* by Andrew Ng.

Word representation

V = [a, aaron, …, zulu, <UNK>]

$|V| = 10{,}000$

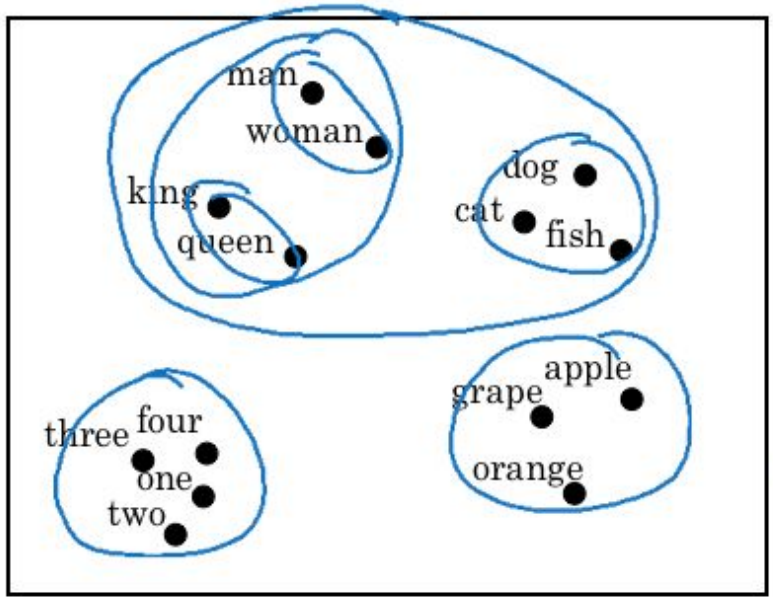
1-hot representation

Man	Woman	King	Queen	Apple	Orange
(5391)	(9853)	(4914)	(7157)	(456)	(6257)

I want a glass of orange _____.

I want a glass of apple _____.

Visualizing word embeddings

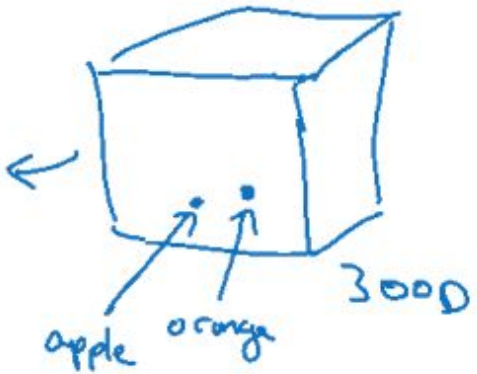


t-SNE

→ 300 D

↓

2D



[van der Maaten and Hinton., 2008. Visualizing data using t-SNE]

Featurized representation: word embedding

	Man (5391)	Woman (9853)	King (4914)	Queen (7157)	Apple (456)	Orange (6257)
Gender	-1	1	-0.95	0.97	0.00	0.01
Royal	0.01	0.02	0.93	0.95	-0.01	0.00
Age	0.03	0.02	0.7	0.69	0.03	-0.02
Food	0.04	0.01	0.02	0.01	0.95	0.97
⋮	⋮	⋮				
size						
cost						
alive						
verb						

300

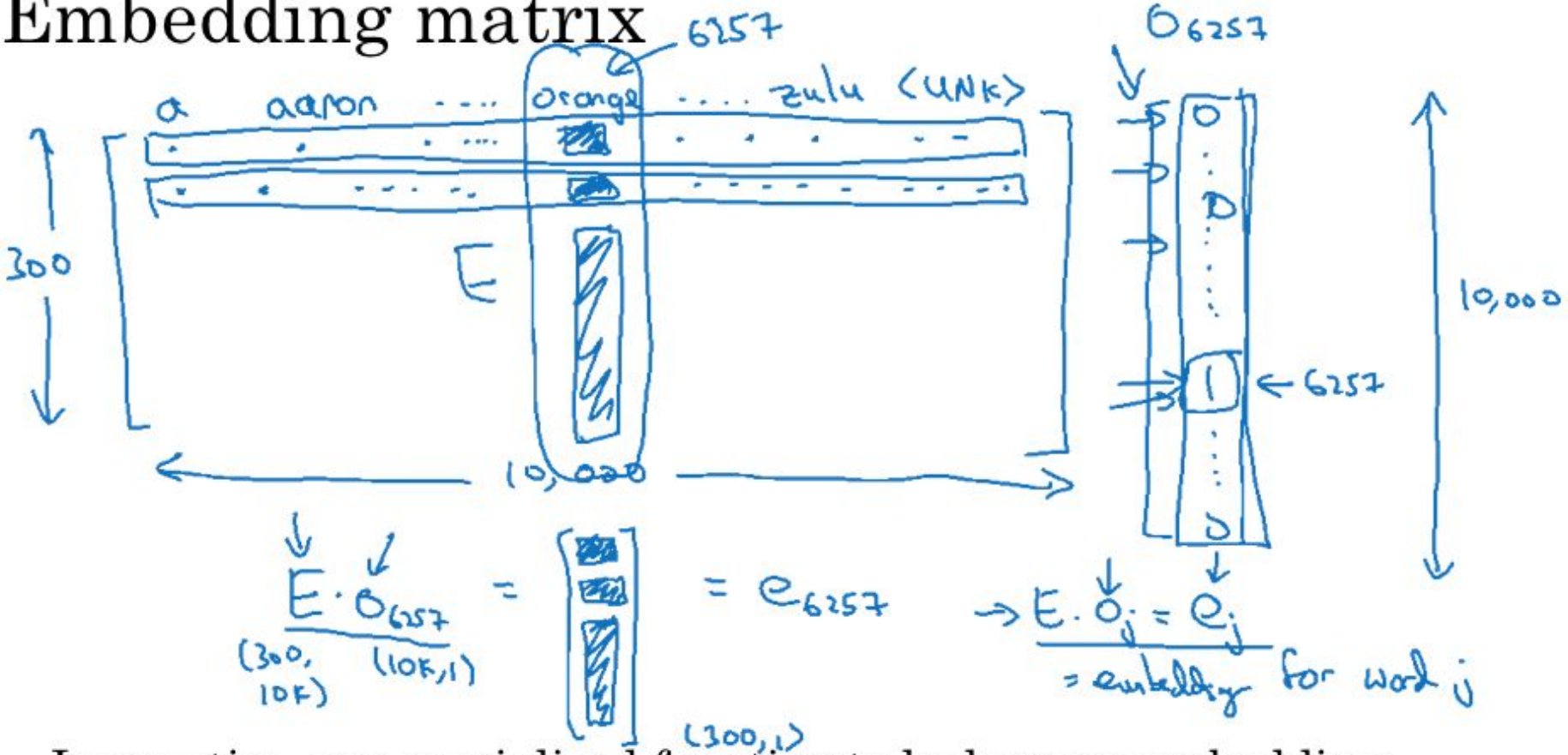
e_{5391} e_{9853}

I want a glass of orange juice.

I want a glass of apple juice.

Andrew Ng

Embedding matrix



In practice, use specialized function to look up an embedding.

→ Embedding

Analogies

	Man (5391)	Woman (9853)	King (4914)	Queen (7157)	Apple (456)	Orange (6257)
Gender	−1	1	-0.95	0.97	0.00	0.01
Royal	0.01	0.02	0.93	0.95	-0.01	0.00
Age	0.03	0.02	0.70	0.69	0.03	-0.02
Food	0.09	0.01	0.02	0.01	0.95	0.97

e_{5391} e_{man} e_{woman}

$e_{man} - e_{woman} \approx \begin{bmatrix} -2 \\ 0 \\ 0 \\ 0 \end{bmatrix}$

Man → Woman as King → ? Queen

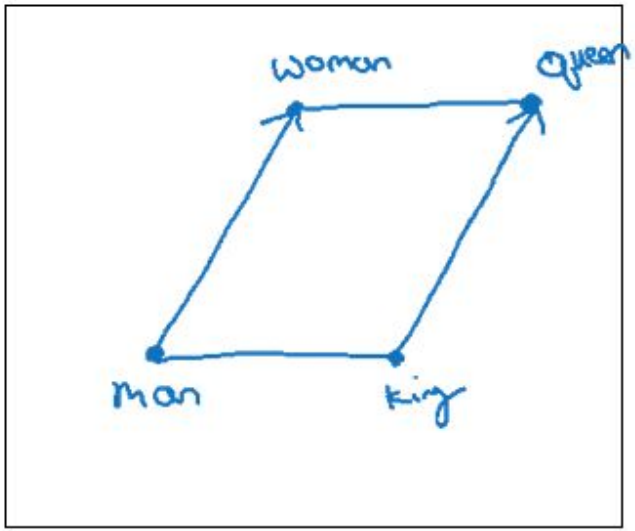
$e_{king} - e_{queen} \approx \begin{bmatrix} -2 \\ 0 \\ 0 \\ 0 \end{bmatrix}$

$e_{man} - e_{woman} \approx e_{king} - e_?$

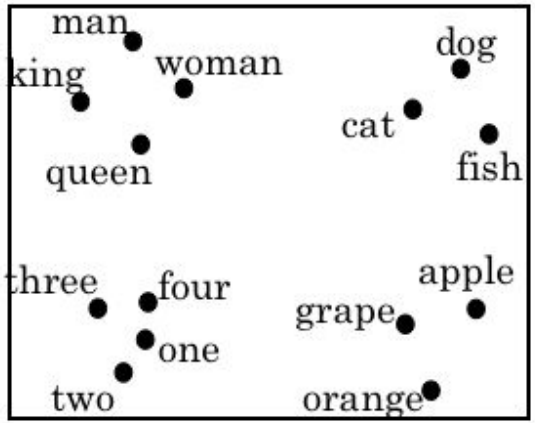
$e_?$ ≈ e_{queen}

[Mikolov et. al., 2013, Linguistic regularities in continuous space word representations]

Analogies using word vectors



300D



300 → 2D

t-SNE

$$e_{man} - e_{woman} \approx e_{king} - e_?$$

e_w

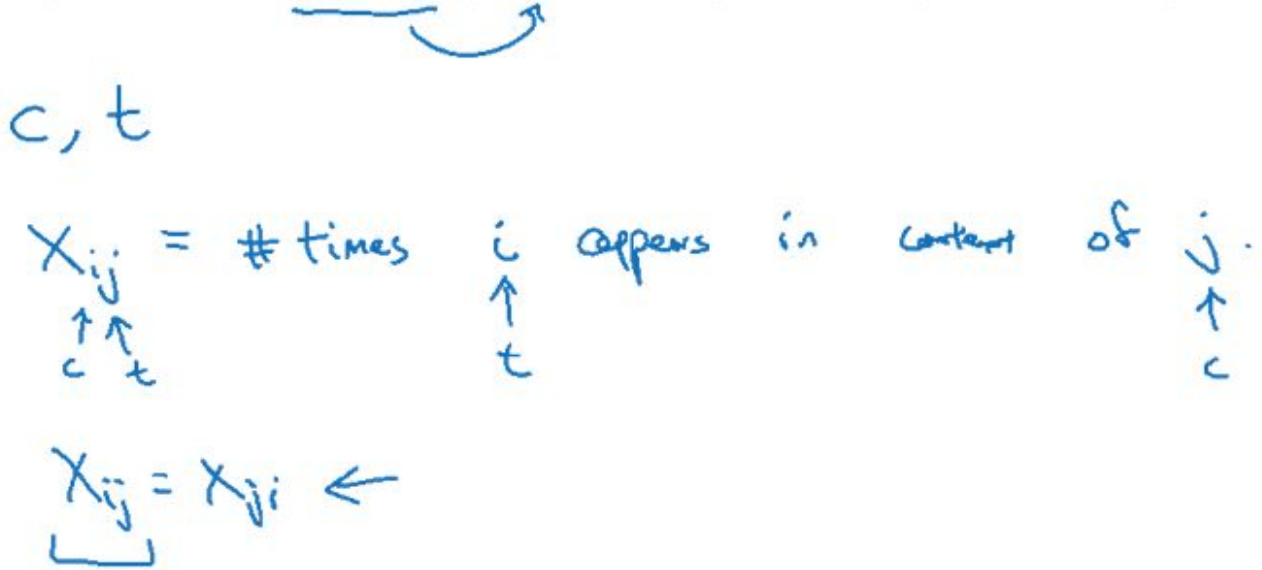
Find word w: $\arg\max_w$ $\text{sim}(e_w, e_{king} - e_{man} + e_{woman})$

30 - 75%

Andrew Ng

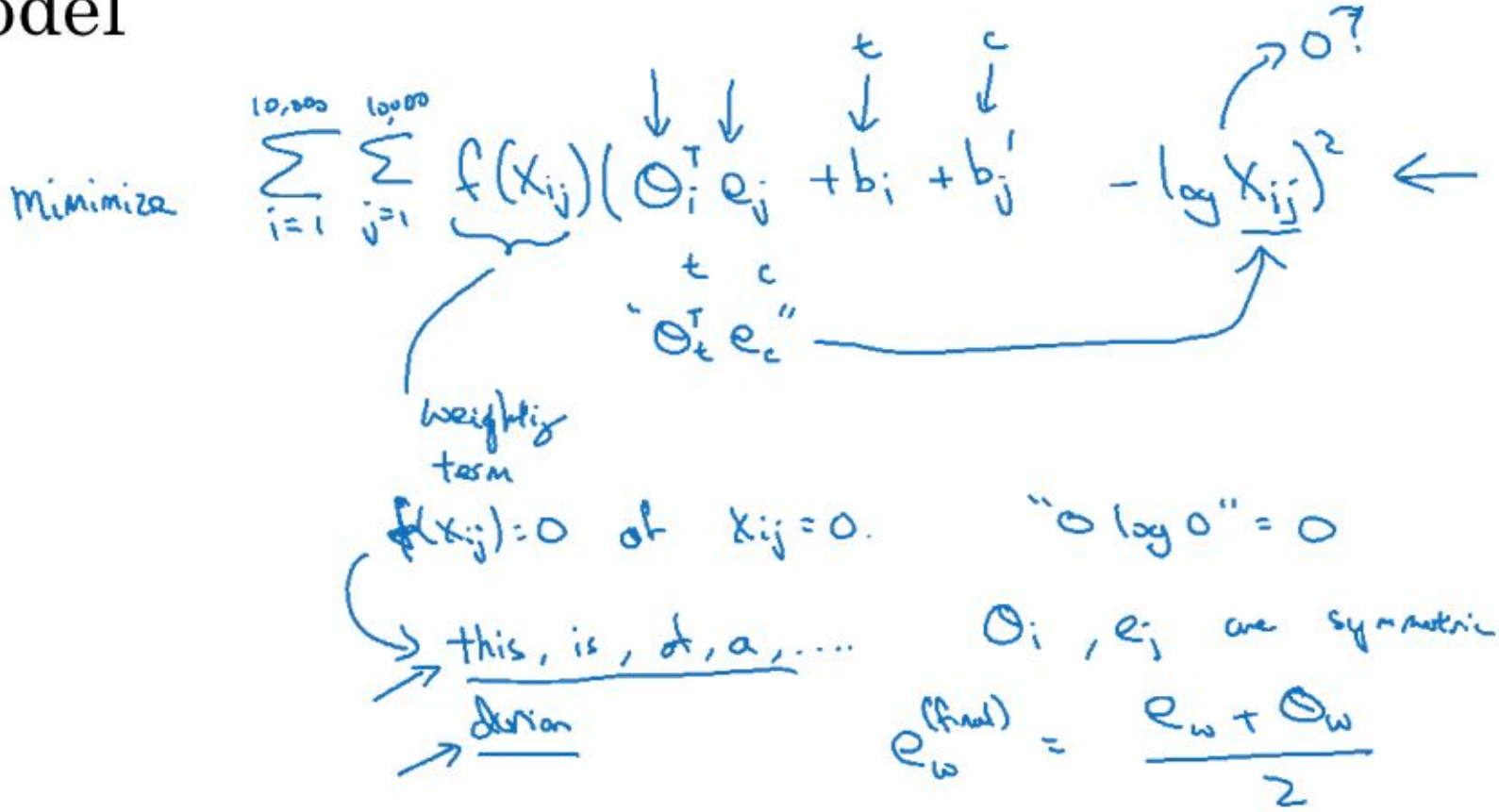
GloVe (global vectors for word representation)

I want a glass of orange juice to go along with my cereal.



[Pennington et. al., 2014. GloVe: Global vectors for word representation]

Model



Andrew Ng

Work Experience

Work at Unitymedia

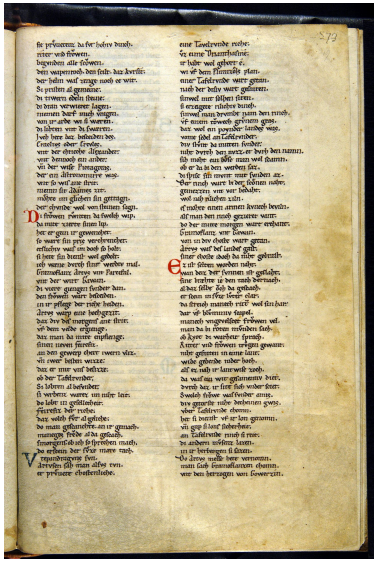
- Recommendation System
- Churn Prevention
- ETL in Hadoop

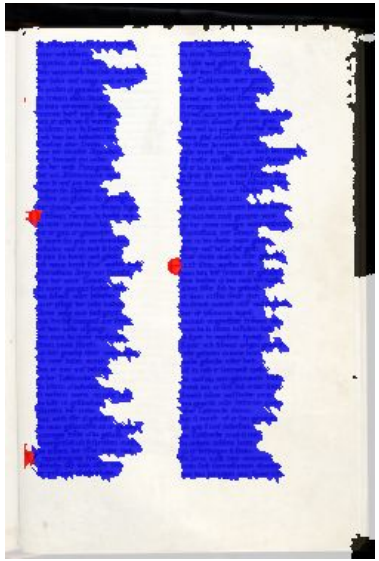
Work at AGT International

- Punch Recognition with Deep Learning
- A/B Testing for Punch Recognition Evaluation
- Anomaly Detection

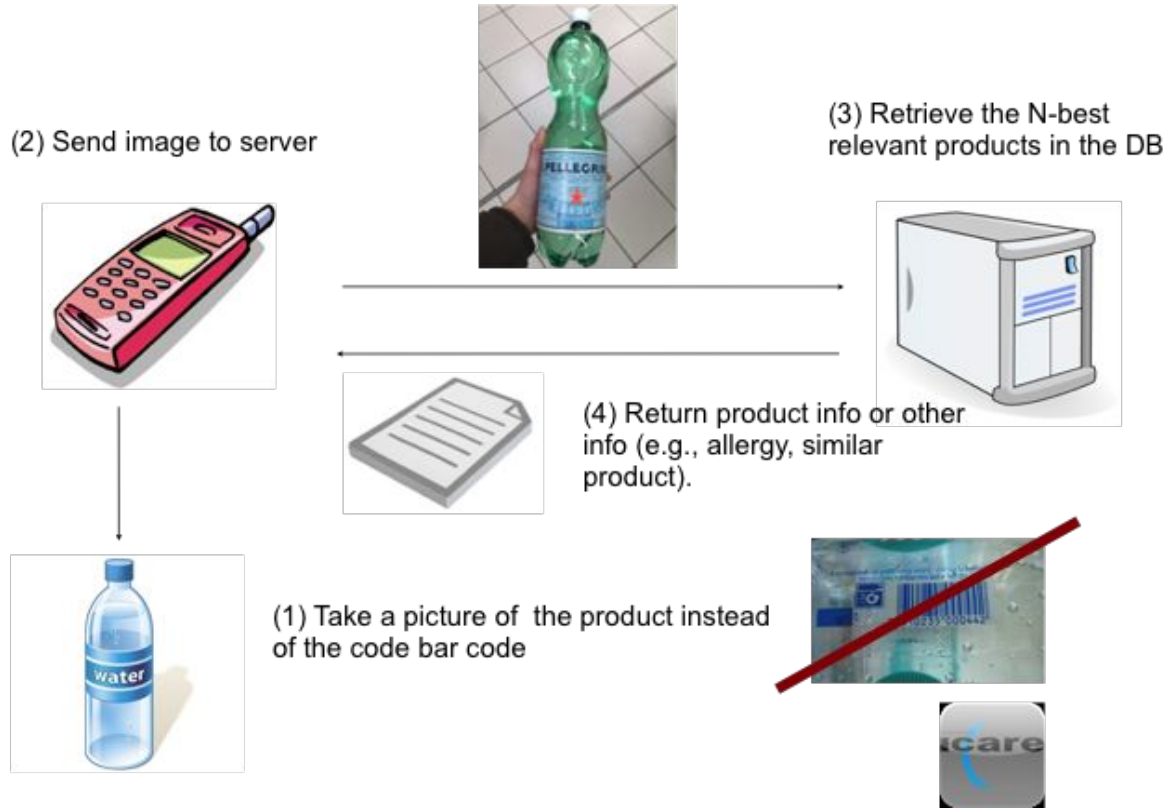
PhD thesis: Historical Document Layout Analysis with Machine Learning

- Goal
  - Developing a general page segmentation method with minimal prior knowledge.
- Basic Idea
  - Page Segmentation ⟶ Pixel Labeling

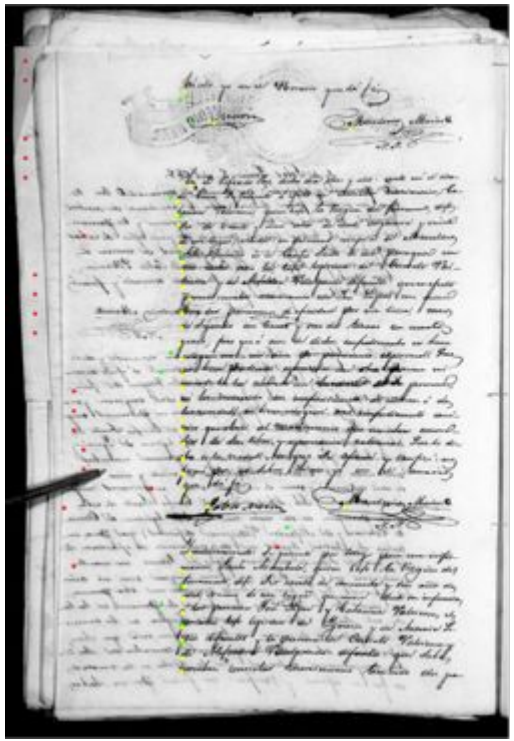




Camera-based image retrieval (master project)



Text line detection in historical document

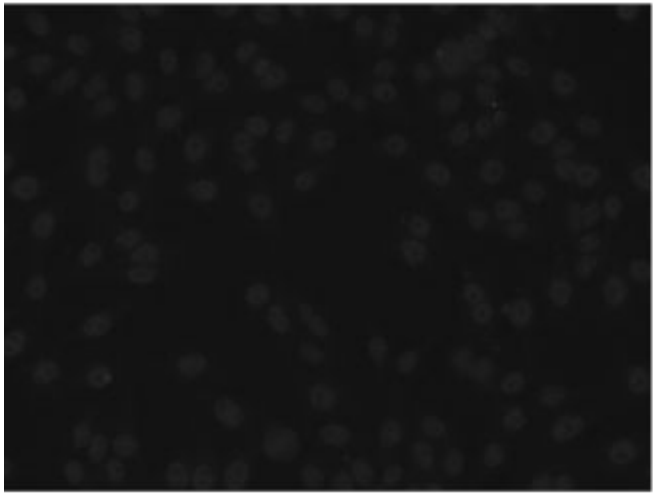


Ground truth

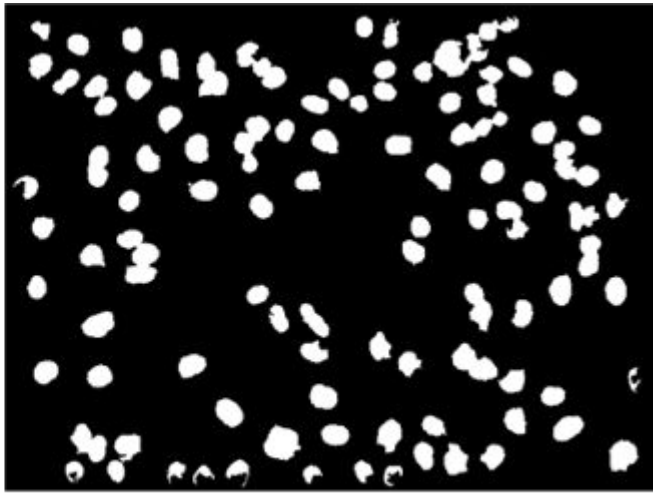
Correct detection

Error

Cell image segmentation



input



output

Other Projects

- Structured Data
  - Booking prediction
  - Revenue per click prediction
  - Consumer shopping prediction
  - Car price prediction
- Computer Vision (CV)
  - Product category classification
- Natural Language Processing (NLP)
  - Toxic Comment Classification (Kaggle challenge, top 16%)
- CV + NLP
  - Product description generation

Github: github.com/ck-unifr

Why I want to join zeroG?