ONLINE JOB ADVERTISEMENTS DEDUPLICATION USING LARGE LANGUAGE MODEL

JAKUB ŻEREBECKI, MIKOŁAJ TYM

Web Intelligence Deduplication Challenge

- Challenge was announced by European Statistics Awards
- The Deduplication Challenge was focused on identifying potential duplicates of job postings published on the web
- Companies often publish job advertisements on different web portals
- Posting advertising the same jobs must be identified and removed using automatic and robust solutions to avoid double counting

Dataset

- The competition dataset contain 112,000 online job advertisements, retrieved from around 400 websites active in the European Union
- The competition organizers have taken authentic job advertisements and created full, semantic, temporal, partial duplicates across different languages
  - Thus, organizers created a **synthetic** dataset for the competition
- 12.5B possible combinations

Considered duplicates

- Full
- Semantic
- Temporal
- Partial
- Non-duplicate

Full duplicates

- Two job advertisements are both exactly the same, i.e. they have the same job title and job description
- They may have differing sources and retrieval dates

Semantic duplicates

- Two job advertisements advertise the same job position and include the same content in terms of the job characteristics
  - The same occupation, education or qualification requirements
- They may be expressed differently in natural language or in different languages

Temporal duplicates

- Temporal duplicates are semantic duplicates with varying advertisement retrieval dates

Partial duplicates

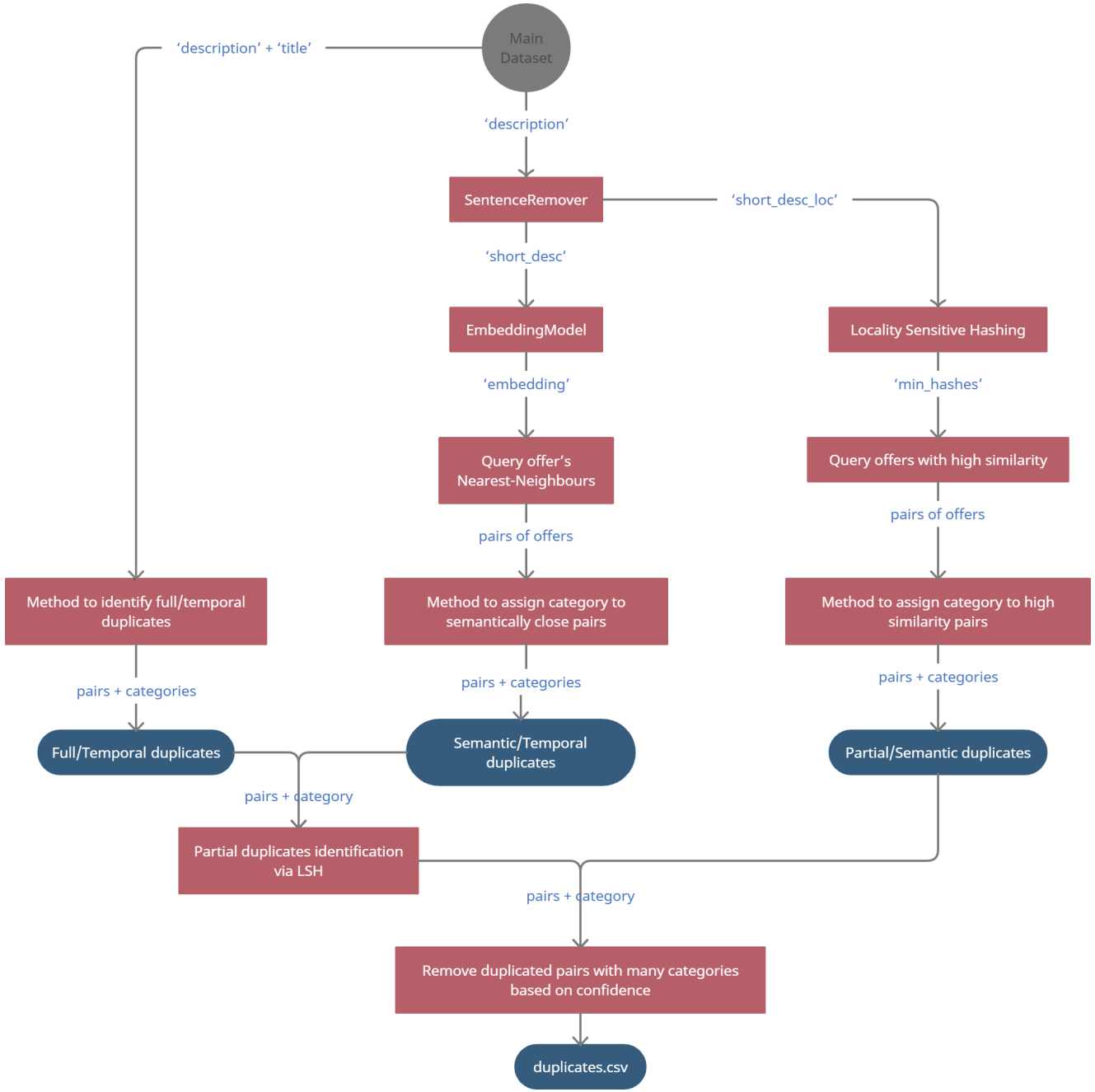
- Two job advertisements describe the same job position but do not necessarily contain the same characteristics
  - One job advertisement contains characteristics that the other does not
- Partial duplicates can be identified by searching the parent offer
  - It is common that one job advertisement (parent) contains all the information, while another advertisement (child) with missing words from the parent offer’s text is placed on another website

Non-duplicates

- If specific job advertisements cannot be described as full duplicates, partial duplicates, semantic duplicates or temporal duplicates they are considered non-duplicates

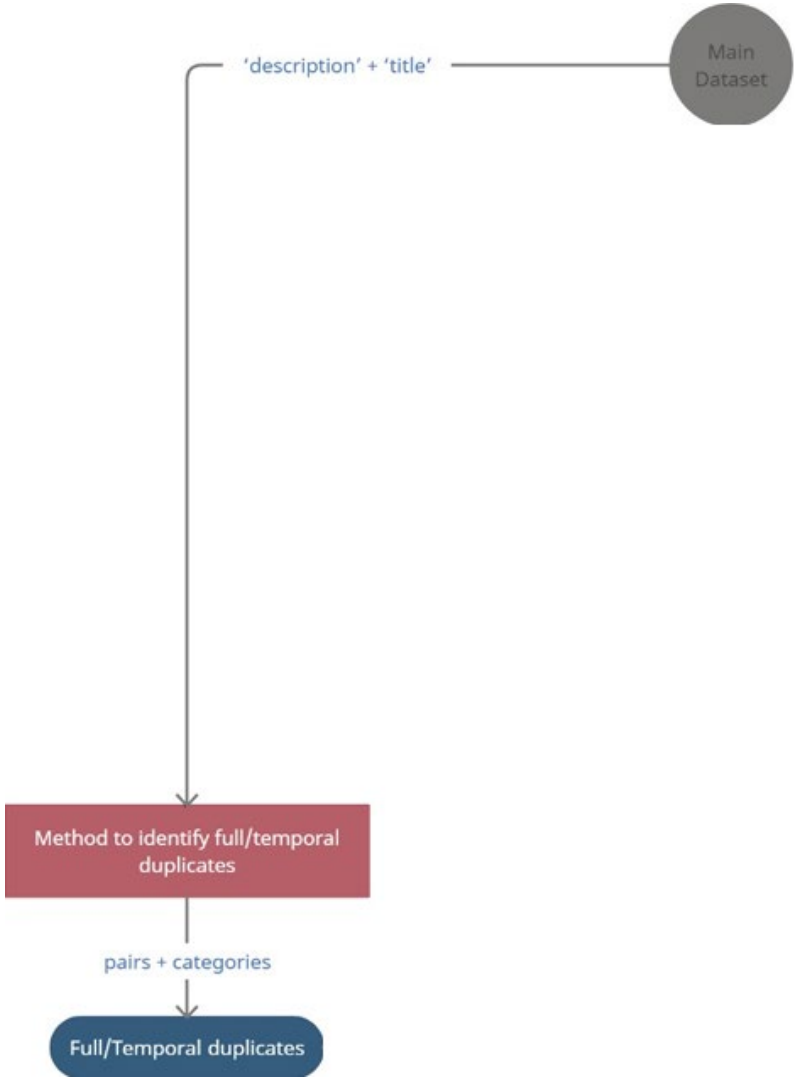
APPROACH

- Three different types of methods
  - Full duplicates identification
  - Comparison of encoded information
  - Words similarity



Full duplicates identification

- The easiest to classify
- The method uses exact comparison of text
  - MD5 (maps any length job offer to fixed-size values)
  - Character-level
- Positive matches were full duplicates
- Time difference between offers changes pair classification

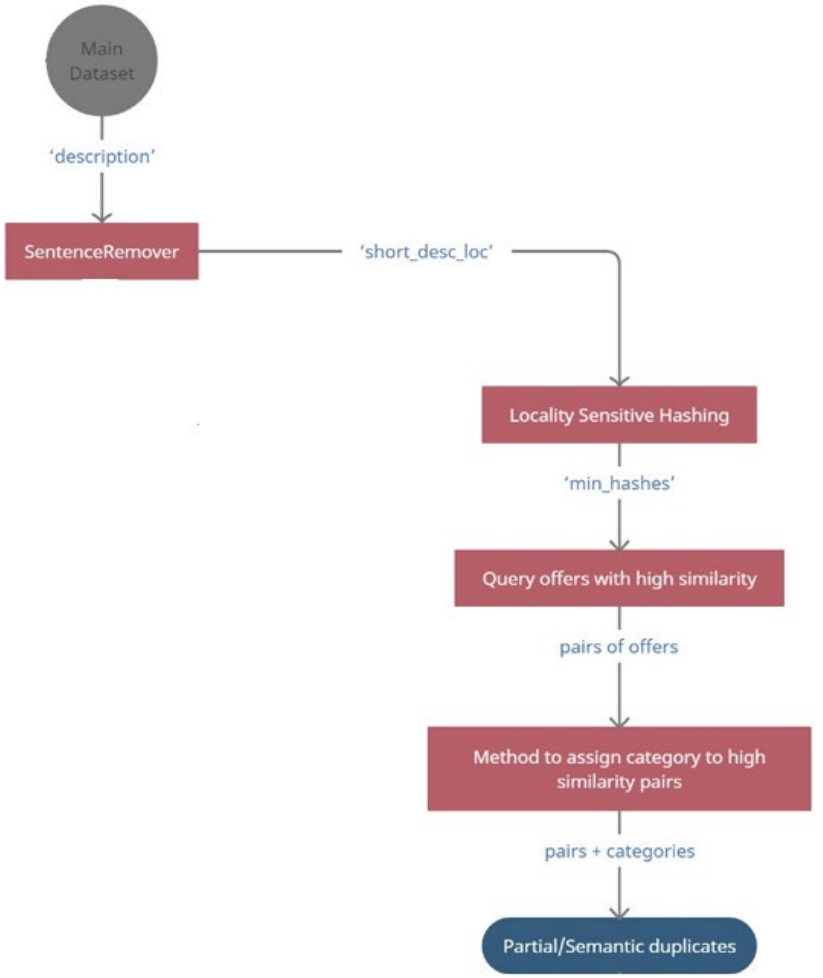


Semantic duplicates identification

- Use embeddings to compare texts expressed differently in natural language or in different languages
  - Encoder-like model to transform text
  - Similar offers have close distance
- Challenges:
  - A lot of jobs had the same informations like GDPR clauses, registration forms
    - Removal of common phrases at the 1st step
  - To classify 100 000 offers we need to compare above 10 billion pairs
- High similarity metric means semantic duplicates

Partial duplicates identification

- The hardest to identify
- To find partial duplicates in the same language we focus on comparing text
    - Use hash function different from MD5
    - Word level comparison
- Using this method, we could find a pairs of offers that are similar then we measured if any words are missing



Parent offer

Qualifications: Proficient in at least one of programming and query languages like Python, PySpark, SQL etc Understand various Data Science, Machine Learning concepts & algorithms such as clustering, regression, classification, forecasting, neural networks, hyperparameters optimization, NLP, LLM. Very good command of English language. What we offer: Permanent employment contract Employee Assistance Program - legal, financial, and psychological consultations. Paid employee referral program.

Child offer

Qualifications: Proficient in at least one of programming and query languages like Python, PySpark, SQL etc Understand various Data Science, Machine Learning concepts & algorithms such as clustering, regression, classification, forecasting, neural networks, hyperparameters optimization, NLP, LLM. What we offer: Permanent employment contract Employee Assistance Program - legal, financial, and psychological consultations. Paid employee referral program.

Partial duplicates cross-lingual identification

- Use embeddings to find the most similar offers to child offer

Parent offer

Qualifications: Proficient in at least one of programming and query languages like Python, PySpark, SQL etc Understand various Data Science, Machine Learning concepts & algorithms such as clustering, regression, classification, forecasting, neural networks, hyperparameters optimization, NLP, LLM. Very good command of English language. What we offer: Permanent employment contract Employee Assistance Program - legal, financial, and psychological consultations. Paid employee referral program.

Child offer

Qualifications: Proficient in at least one of programming and query languages like Python, PySpark, SQL etc Understand various Data Science, Machine Learning concepts & algorithms such as clustering, regression, classification, forecasting, neural networks, hyperparameters optimization, NLP, LLM. What we offer: Permanent employment contract Employee Assistance Program - legal, financial, and psychological consultations. Paid employee referral program.

Partial duplicate

High similarity

Sibling offer (Polish language)

Kwalifikacje: Znajomość przynajmniej jednego z języków programowania i zapytań, takich jak Python, PySpark, SQL itp. Zrozumienie różnych koncepcji i algorytmów Data Science, Machine Learning, takich jak klasteryzacja, regresja, klasyfikacja, prognozowanie, sieci neuronowe, optymalizacja hiperparametrów, NLP, LLM. Co oferujemy: Stała umowa o pracę Program pomocy pracownikom - konsultacje prawne, finansowe i psychologiczne. Płatny program poleceń pracowników.

COMPETITION RESULTS

Results

- Robust methods
- We have achieved the 3rd place in Accuracy category
  - Macro F1 metric -> unweighted mean of per class F1 scores
- The 2nd highest score in partial duplicates identification

Results

User	Submissions	Date of Last Entry	Full F1 ▲	Semantic F1 ▲	Temporal F1 ▲	Partial F1 ▲	Non-Duplicate F1 ▲	Macro F1 ▲
TwoTired	10	03/31/23	0.99 (1)	0.85 (3)	0.87 (4)	0.77 (1)	1.00 (1)	0.90 (1)
TheDeDuplicators	10	03/31/23	0.99 (1)	0.89 (1)	0.92 (1)	0.30 (3)	1.00 (1)	0.82 (2)
IDA	10	03/31/23	0.99 (1)	0.84 (4)	0.88 (3)	0.37 (2)	1.00 (1)	0.82 (2)
Nins	10	03/31/23	0.99 (1)	0.86 (2)	0.87 (4)	0.17 (4)	1.00 (1)	0.78 (3)
SPub.Fr	10	03/31/23	0.99 (1)	0.83 (5)	0.86 (5)	0.17 (4)	1.00 (1)	0.77 (4)
Hyeny	11	03/31/23	0.99 (1)	0.80 (7)	0.91 (2)	0.10 (8)	1.00 (1)	0.76 (5)
Flouss	10	03/29/23	0.99 (1)	0.80 (7)	0.82 (7)	0.15 (5)	1.00 (1)	0.75 (6)
smrek	10	03/31/23	0.99 (1)	0.72 (10)	0.87 (4)	0.11 (7)	1.00 (1)	0.74 (7)
clemanev	8	03/31/23	0.99 (1)	0.70 (11)	0.80 (8)	0.13 (6)	1.00 (1)	0.73 (8)
OptimalEffort	10	03/31/23	0.99 (1)	0.78 (9)	0.80 (8)	0.07 (9)	1.00 (1)	0.73 (8)

Contact

- For more technical details, feel free to contact us

JAKUB ŻEREBECKI

MIKOŁAJ TYM