

ONLINE JOB
ADVERTISEMENTS
CLASSIFICATION USING
ENCODER-LIKE LARGE
LANGAUGE MODEL

MIKOŁAJ TYM, JAKUB ŻEREBECKI

WEB INTELLIGENCE CHALLENGE

EUROPEAN STATISTICS AWARDS

Web intelligence classification challenge

- Challenge was announced by European statistics awards
- Each team could submit 10 submissions which contain classification of online job advertisements occupations
- Predicted classes were evaluated by Lowest Common Ancestor metric
- Competitors must provide fully documented scripts in R or Python
- Approaches were evaluated not only for accuracy but also reusability, so they should be scalable and open



The International Standard Classification of Occupations

- Four-level classification of occupation groups managed by the International Labour Organisation
- There are 436 occupation classes
 - Despite of some of the classes are strongly semantically related, they occur in different ISCO tree branches
 - Accountants
 - Professionals
 - Accounting and bookkeeping clerks
 - Clerical support workers
 - LCA metric heavily penalizes such mistakes

| ISCO code | 3214 |
|----------------|---|
| Label 1 | Technicians and associate professionals |
| Label 2 | Health associate professionals |
| Label 3 | Medical and pharmaceutical technicians |
| Label 4 | Medical and dental prosthetic technicians |

Dataset

- The competition dataset contains 26,000 multilingual online job advertisements
- They were retrieved from around 400 websites active in the European Union
- These advertisements were scrapped from the web, so they contain many irrelevant data
 - GDPR clause
 - HTML tags
 - Job benefits
 - Company policies

DUTY CLASSIFIER

DATA PREPROCESSING STEP TO CLEAN JOB ADVERTISEMENTS

Job offer example

- Advertisement contains many sections
- Not all of them are relevant in case of classification
 - Employer overview is misleading
- The key part of job offer are requirements but the text which describes them is often shorter compared to other details

1. Employer overview

Accenture is a leading global professional services company that helps the world's leading businesses, governments, and other organizations build their digital core, optimize their operations, accelerate revenue growth, and enhance citizen services. We offer solutions and assets across Strategy & Consulting, Technology, Operations, Industry X, and Accenture Song

2. Requirements

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 All employment decisions shall be made without regard to age, race, creed, color, religion, sex, national origin, ancestry, disability status, veteran status, sexual orientation, gender identity or expression, genetic information, marital status, citizenship status or any other basis as protected by federal, state, or local law.

3. Benefits

4. Equal employment opportunity statement

Filtering non-meaningful informations

- We have trained a model to classify sentence whether it contains text regarding job requirements or not
- How to split offer without dots?
 - Solution: high-quality collection of words to separate sentences
 - Not best split points: SQL, Python, Data Science
 - Risk of merging informations from multiple sections

Accenture is a leading global professional services company that helps the world's leading businesses, governments, and other organizations build their digital core, optimize their operations, accelerate revenue growth, and enhance citizen services. We offer solutions and assets across Strategy & Consulting, Technology, Operations, Industry X, and Accenture Song

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

All employment decisions shall be made without regard to age, race, creed, color, religion, sex, national origin, ancestry, disability status, veteran status, sexual orientation, gender identity or expression, genetic information, marital status, citizenship status or any other basis as protected by federal, state, or local law.

Filtering example

- Sentences from requirements section are classified with high confidence as important
- Verification step not to remove too much information
- Precisely filtered out employer overview and equal employment opportunity statement

--- Sentences Above Duty Threshold (0.5) ---

Selected: Qualifications Proficient in at least one of programming and query languages like Python, PySpark, SQL etc, Probability: 0.9000 (Above threshold)

Selected: Understand various Data Science, Machine Learning concepts & algorithms such as clustering, regression ... NLP, LLM, Probability: 0.8000 (Above threshold)

Selected: Very good command of English language, Probability: 0.9900 (Above threshold)

Additional sentences needed. Current: 3 out of 4.

Added: What we offer: Permanent employment contract Employee Assistance Program ... Paid employee referral program, Probability: 0.4500 (To meet minimum count)

--- NOT Selected Sentences ---

Sentence: Accenture is a leading global professional services company that helps the world's leading businesses, governments, ... and enhance citizen services., Probability: 0.3000

Sentence: We offer solutions and assets across Strategy & Consulting, Technology, Operations, Industry X, and Accenture Song, Probability: 0.3500

Sentence: All employment decisions shall be made without regard to age, race, creed, color, religion, sex, national origin, gender identity ..., Probability: 0.0700

--- Final Selected Sentences ---

Sentence: Qualifications Proficient in at least one of programming and query languages like Python, PySpark, SQL etc, Probability: 0.9000

Sentence: Understand various Data Science, Machine Learning concepts & algorithms such as clustering, regression ... NLP, LLM, Probability: 0.8000

Sentence: Very good command of English language, Probability: 0.9900

SYNTHETIC TRAINING SET

FROM DATA CLEANING TO ASSIGNMENT OF LABEL FOR EACH OFFER

Training set for ISCO classification

- New model and training examples are needed
 - Competition dataset (lack of labels & ambiguous informations)
- Ready to use dataset - European Skills, Competences, Qualifications and Occupations dataset was used to create training examples
 - Content of ESCO
 - Doesn't contains any details that are irrelevant for classification

Software integration engineer

Integration engineers develop and implement solutions which coordinate applications across the enterprise or its units and departments. They evaluate existing components or systems to determine integration requirements and ensure that the final solutions meet organisational needs. They reuse components when possible and assist management in taking decisions. They perform ICT system integration troubleshooting.

CLASSIFIER RESULTS

ISCO CODE CLASSIFIER

Software engineer example

We are looking for a Python Developer to join our team and work on scalable web applications and data processing solutions. Required skills include Python, Django or Flask, and experience with REST APIs and databases (SQL/NoSQL). Knowledge of Docker and cloud platforms is a plus. We offer a competitive salary, flexible working hours, and remote options. Join us and build innovative software solutions!

| Code | Probability | Label 4 | Label 3 | Label 2 | Label 1 |
|------|-------------|---------------------------------------|---|---|---------------|
| 2512 | 0.82 | Software developers | Software and applications developers and analysts | Information and communications technology professionals | Professionals |
| 2514 | 0.04 | Applications programmers | Software and applications developers and analysts | Information and communications technology professionals | Professionals |
| 2521 | 0.03 | Database designers and administrators | Database and network professionals | Information and communications technology professionals | Professionals |

Sales representative example

We need a Sales Representative to build relationships with clients and drive revenue growth. Strong communication skills and previous sales experience are required. Knowledge of CRM tools and negotiation techniques is a plus. We offer a base salary plus commission and career development opportunities. Join us and grow with our company!

| Code | Probability | Label 4 | Label 3 | Label 2 | Label 1 |
|------|-------------|---|---|---|---|
| 3322 | 0.66 | Commercial sales representatives | Sales and purchasing agents and brokers | Business and administration associate professionals | Technicians and associate professionals |
| 2434 | 0.10 | Information and communications technology sales professionals | Sales, marketing and public relations professionals | Business and administration professionals | Professionals |
| 2433 | 0.08 | Technical and medical sales professionals (excluding ICT) | Sales, marketing and public relations professionals | Business and administration professionals | Professionals |
| 5244 | 0.04 | Contact centre salespersons | Other sales workers | Sales workers | Service and sales workers |

Hotel receptionist example

We are looking for an employee to provide excellent guest service and manage reservations.\ Strong communication skills and experience in hospitality are required. Familiarity with booking systems is a plus. We offer competitive pay, training, and career growth opportunities. Join our team and create a welcoming experience for guests!

| Code | Probability | Label 4 | Label 3 | Label 2 | Label 1 |
|------|-------------|-------------------------------|-------------------------------|---|--------------------------|
| 4224 | 0.75 | Hotel receptionists | Client information workers | Customer services clerks | Clerical support workers |
| 1411 | 0.08 | Hotel managers | Hotel and restaurant managers | Hospitality, retail and other services managers | Managers |
| 4221 | 0.03 | Travel consultants and clerks | Client information workers | Customer services clerks | Clerical support workers |

Model fine-tuning

- We fine-tuned a lightweight LLM (400m parameters) for classification task
- The model was trained on clean data instead of real data with biases
 - Model classifies job offer into one of 436 classes
- Best results (competition): 0.58 LCA score
- Human benchmark we developed (non-expert): 0.58 LCA score
- Our lightweight model's result: 0.52 LCA score
 - Top-5 accuracy instead of LCA: 0.8 accuracy score

CONCLUSIONS & RECOMMENDATIONS

OJA CLASSIFICATION RECOMMENDATIONS

How to Mitigate Challenges?

- Challenges in OJA Classification:
 - Ambiguous ISCO labels.
 - Irrelevant texts in job adverts.
- Solutions:
 - Use top-5 accuracy for ISCO labels instead of LCA metric.
 - Extract HTML content from the web page instead of plain text.
 - This helps data engineers filter out irrelevant sections without needing an extra classifier.

Production Solution

- Web Intelligence Hub contains around 200 million job adverts
 - The approach for classifying this data must be highly efficient.
 - Keyword-based classification is unreliable – NLP models offer better accuracy.
- Why Not Cosine Similarity?
 - Requires encoding all job adverts (identical complexity as fine-tuned classifier).
 - Needs a vector database for embeddings – extra complexity for similarity search.
- Encoder-based classifier as a better alternative
 - Same or even lower complexity than cosine similarity.
 - Model trained specifically for job adverts may encode them better than pre-trained model for embeddings.

Contact

- For more technical details, feel free to contact us



JAKUB ŻEREBECKI



MIKOŁAJ TYM

