Real-time Labour Market Information on Skill Requirements

Setting up the infrastructure for EU system

Data pre-processing

Alessandro Vaccarino - Mauro Pelucchi
Expert Workshop
Milan, 20-21 March 2018
Topics

1. Context
2. Data processing pipeline
3. Vacancy detection with machine learning techniques
4. Deduplication issues
5. Date detection and vacancy’s duration
Topics

1. Context
2. Data processing pipeline
3. Vacancy detection with machine learning techniques
4. Deduplication issues
5. Date detection and vacancy’s duration
Context

JUNIOR SOFTWARE DEVELOPER

Location: United Kingdom
Application deadline: Saturday, 30 September 2017
Reference number: 100

As Junior Software Developer, you will develop excellent software for use in field mapping, data collection, sensor

ZARA CENTRAL LONDON TEMPORARY CHRISTMAS VACANCIES SALES ASSISTANT AND CASHIER
LONDON, GREATER LONDON - THE UNITED KINGDOM

GBZA00116121 / Store

<table>
<thead>
<tr>
<th>Job Title</th>
<th>SALES ASSISTANT POST</th>
</tr>
</thead>
<tbody>
<tr>
<td>According to</td>
<td>RESPONSIBLE / DEPUTY MANAGER</td>
</tr>
<tr>
<td>HEAD CASHIER and/or GENERAL MANAGER</td>
<td></td>
</tr>
<tr>
<td>Purpose</td>
<td>To support store objectives by delivering excellent service in line with store needs and to ensure the effective maintenance of designated sections</td>
</tr>
<tr>
<td>Key Responsibilities</td>
<td>- Be aware of what happens day to day in the store</td>
</tr>
<tr>
<td></td>
<td>- Maintain responsibility for one or several sections of the store as instructed by your Manager keeping a basic image of the area (folding and displaying)</td>
</tr>
<tr>
<td></td>
<td>- Process sales, exchanges and refunds efficiently, accurately and quickly in all payment methods</td>
</tr>
<tr>
<td></td>
<td>- Ensure that good housekeeping standards are maintained on the shop floor when required and ensure that the Cash desk areas are replenished and kept tidy throughout the day</td>
</tr>
<tr>
<td></td>
<td>- Ensure that customers are acknowledged and receive a quality service either face to face or over the telephone</td>
</tr>
</tbody>
</table>

Noise
Coverage among all countries and occupations
No standard
Unstructured data
Data driven approach for Real Time Labour Market Intelligence

Sensors receive information from the field after the actuations.

Sensors
- Raw data from the web

Models
- Algorithms analyzing raw data to extract information needed to take better decisions

Policy Makers
- Action that interact with the context
Junior Software Developer

As Junior Software Developer, you will develop excellent software for use in field mapping, data collection, sensor networks, street navigation, and more. You will collaborate with other programmers and developers to autonomously design and implement high-quality web-based applications, restful API’s, and third party integration.

We’re looking for a passionate, committed developer that is able to solve and articulate complex problems with application design, development and user experiences. The position is based in our offices in Harwell, United Kingdom.
Topics

1. Context
2. Data processing pipeline
3. Vacancy detection with machine learning techniques
4. Deduplication issues
5. Date detection and vacancy’s duration
Overall Data Flow - Recap

Data Ingestion → Pre-Processing → Information Extraction → ETL → Presentation Area
Challenges of Data Processing

- How do we measure the quality of collected data?
- How do we increase the quality of our data?
- Which variables affect quality of data?
- How do we keep track of quality of data?
Challenges of Data Processing

Which variables affect quality of data?

- Duplicate job ads
- Availability of metadata
- Timeliness and periodicity of the data
- Complexity of data
Challenges of Data Processing

How do we keep track of quality of data?

Continuous quality check gates

Merge
Cleaning
Information Extraction
Some examples

Ads for services, courses, ...

Type: Housekeeping and Home - Seamstress
Requested for Beatrice in 53100 Siena (Siena)
Service ID: 406157 | Do you want to access Sarta's works?

Description
We are looking for a Sarta service with the following features:

Description of the work to be carried out (indications to be followed)
It's about having to resize a cashmere coat, then shorten it, tighten it, shorten the sleeves
Description of the type of arrangement that the suit will require and whether it is necessary to make or not a custom suit
- The types of garments that you will have to arrange or tailor-made are (shirts, pants, evening dresses, clothes, ...)
Coat
Number and type of animals requested
A garment, cashmere coat

Banner and other forms nested in web pages

Interested?
Become a member for free to remember jobs, apply faster - and get to know your future colleagues!

List of similar jobs, recommendations, you might also be interested in ...

<table>
<thead>
<tr>
<th>Similar jobs</th>
</tr>
</thead>
<tbody>
<tr>
<td>NET Developer</td>
</tr>
<tr>
<td>Software Developer</td>
</tr>
<tr>
<td>NET Developer, ASPNET MVC, C# - Swansea</td>
</tr>
<tr>
<td>NET Developer, ASPNET MVC, C# - Swansea</td>
</tr>
<tr>
<td>NET Developer, ASPNET MVC, C# - Swansea</td>
</tr>
</tbody>
</table>

CEDEFOPEuropean Centre for the Development of Vocational Training
What are the effects of Low-quality?

- Decisions are slowing down
- Loss of opportunity
- Credibility
- Increased costs
- Inefficiency in operating processes
Experimental case study
Measure of quality

Completeness measures the availability of metadata

<table>
<thead>
<tr>
<th></th>
<th>Primary Job Portals</th>
<th>Combination of primary and secondary portals</th>
<th>Secondary Job Portals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Completeness</td>
<td>90%</td>
<td>77%</td>
<td>68%</td>
</tr>
</tbody>
</table>
Pre-processing is an important and critical step. During this phase, through a series of activities, the data will be cleaned from noise, summarized and prepared for the information extractions step.
Merging

Structured and unstructured data are merged with the purpose to create a complete dataset. The same job ads can be crawled from several sources: using "real page URL" we can merge this information.
Cleaning

The cleaning task is a sequence of activities that remove the noise and improve quality of the imported data.

Check imported text
Drop Duplicates
Remove Noise with Machine Learning techniques
Text processing and summarizing

The text processing and summarizing phase aims at reducing the text to improve the process of classifications of job vacancies according to the European standards.

Job posting text

Denoising and processing

Vector Space Model representation

JUNIOR SOFTWARE DEVELOPER
Location: United Kingdom
Application deadline: Saturday, 30 September 2017
Reference number: 100

As Junior Software Developer, you will develop excellent software for use in field mapping, data collection, sensor networks, street navigation, and more. You will collaborate with other programmers and developers to autonomously design and implement high-quality web-based applications, restful APIs, and third party integration.

We're looking for a passionate, committed developer that is able to solve and articulate complex problems with application design, development and user experiences. The position is based in our offices in Harwell, United Kingdom.

As Junior Software Developer, you will develop excellent software for use in field mapping, data collection, sensor networks, street navigation, and more. You will collaborate with other programmers and developers to autonomously design and implement high-quality web-based applications, restful APIs, and third party integration.

We're looking for a passionate, committed developer that is able to solve and articulate complex problems with application design, development and user experiences. The position is based in our offices in Harwell, United Kingdom.

We’re looking for a passionate, committed developer that is able to solve and articulate complex problems with application design, development and user experiences. The position is based in our offices in Harwell, United Kingdom.
Text processing and summarizing

All the components are developed to work in multi-language environment

For example, stop-words and our ontologies cover about 60 languages

Language Detector (Naïve Bayes classifier trained on Wikipedia corpus) recognizes 60 languages with an accuracy of ~99%
Other techniques

• NLP off-the-shelf framework
  • Difficult to find resources for all languages
  • Low performances
Topics

1. Context
2. Data processing pipeline
3. Vacancy detection with machine learning techniques
4. Deduplication issues
5. Date detection and vacancy’s duration
Vacancy detection

1. ~1 million of job ads (last year) (WollyBI - Italy)

3. ~3 millions of job ads collected from 2013 (WollyBI - Italy)

7. ~7 millions of job ads collected in about 1 month

A “bunch” of unstructured data.

How identify job ads?

How remove the noise?

How extract information from the data?
Vacancy detection

Data Analysis Engineer

£35k - £40k pa | Cambridge | Permanent

Our client, a global leader in their field, are seeking a Data Analyst Engineer to join their global headquarters in Cambridge.

The company work in a very niche field and have successfully built a world-class product. They now have many clients under their belt and are looking to further the development and reach of the product.

Noise detection

Detect Noise with Machine Learning techniques
Which kind of noise?

1. Pages that are no job ads
2. Ads of training courses
3. Privacy and cookie law pages
4. Blog articles, news pages, contact forms, ...
5. List of job ads
Experimental case study
Measure of quality

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Completeness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary Job Portals</td>
<td>80%</td>
<td>90%</td>
</tr>
<tr>
<td>Combination of primary and secondary portals</td>
<td>76%</td>
<td>77%</td>
</tr>
<tr>
<td>Secondary Job Portals</td>
<td>67%</td>
<td>68%</td>
</tr>
</tbody>
</table>

Accuracy is the degree to which the information correctly describes the phenomena.
Noise filter pipeline

Header filter
Looks at page headers (title and other metadata) to judge if forged or not

Content filter
Scans text content of pages with fuzzy logic

Machine Learning filter
It's particularly effective to filter out noise written in several languages.

Job ads
Machine learning filter

• Noise Filtering can be seen as a specific text categorization: It’s a classification problem (supervised learning)!
  • Like Spam detection
• Particular words have particular probabilities of occurring in spam pages and in job ads
• The filter doesn't know these probabilities in advance, and must first be trained so it can build them up
• To train the filter, we can use thousands of pages crawled in web and thousands of job ads from API and legitimate sources.
Algorithms Used in Noise Detection
Which Algorithm is Better?

- Naïve Bayes: 95%
- Linear Support Vector Machine: 90%
- Decision tree: 85%
- Random forest: 92%
- Gradient-boosted tree: 90%
- Logistic regression: 88%
Which Algorithm is Better?

- Train Data: ~14,000 pages; ~50% noise (balanced dataset) for each language
  - Training time: about 30 minutes
- Evaluation Data: ~6,000 new pages for each language
- Naïve Bayes reported to do very well
  - Results: Job Ads precision 99%, Job Ads recall 97%
  - More complex algorithms performs similarly, but need more time (!!!) to train
Preliminary results - Experimental Case Study
Multi country, Single source, Multi language

Job Ads - Language

<table>
<thead>
<tr>
<th>Language</th>
<th>Others</th>
<th>en</th>
<th>fr</th>
<th>es</th>
<th>it</th>
<th>de</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>34</td>
<td>9443</td>
<td>210</td>
<td>60</td>
<td>1004714</td>
<td>246</td>
</tr>
</tbody>
</table>

Primary Job Portals

<table>
<thead>
<tr>
<th>Job Ads</th>
<th>Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>N. 1196435</td>
<td>12628</td>
</tr>
</tbody>
</table>

~2%
Topics

1. Context
2. Data processing pipeline
3. Vacancy detection with machine learning techniques
4. Deduplication issues
5. Date detection and vacancy’s duration
Problem
Examples of job ads

Fabasoft International Services GmbH
Application Developer (w / m)
Linz, Vienna  Solid employment  full-time  Erschienen: vor 12 Tagen

You are the right one, if

- you are interested in designing, implementing and integrating specialized applications and you like to work with high-end development tools and infrastructures
- you like using state-of-the-art web technologies to realize customized software solutions while being able to expertly use standard products / SDKs as well as libraries
- You have a solid computer science education (HTL, FH or University) and solid programming skills with Java and Eclipse
- Ideally, you have already gained experience with SOAP, WSDL and Web Services, as well as object-oriented modeling (UML) skills

We offer you

- working in a dedicated, highly motivated team in a professional environment
- Consistent training and further education as well as broadening horizons by attending international conferences
- interesting development opportunities for Certified Project Manager, Certified Scrum Master etc.
- numerous benefits, benefits and cool events as well
- above-average payment (at least EUR 35,000 gross annual salary, depending on your experience and qualifications also more) and a performance-based bonus system

Contact Person: Mag. Dipl.-Ing. Christian Distelberger
E-mail: job@fabasoft.com
Tel: +43 732 606162-0

We offer you

- working in a dedicated, highly motivated team in a professional environment
- Consistent training and further education as well as broadening horizons by attending international conferences
- interesting development opportunities for Certified Project Manager, Certified Scrum Master etc.
- numerous benefits, benefits and cool events as well
- above-average payment (at least EUR 35,000 gross annual salary, depending on your experience and qualifications also more) and a performance-based bonus system

Contact Person: Mag. Dipl.-Ing. Christian Distelberger
E-mail: job@fabasoft.com
Tel: +43 732 606162-0

CEDEFOP
European Centre for the Development of Vocational Training
Goals

Avoid that the same job ad, published on various sources, appears in the database several times.
Problems

- When a job vacancy could be considered a duplicate of another one?
- When a job vacancy could be considered “consistent” and do not be subject to the deduplication process?
- Is it necessary to make both an intra-source and an extra-source process to recognize the job vacancies duplicates?
Deduplication

Physical deduplication or fuzzy matching

Made on the description (or content) part of the job vacancy.

Metadata matching

Using metadata coming from job portals to remove job vacancies duplicates on the aggregators websites (e.g. reference id, page url)
Due to vacancy redundancy, duplication values with complete data source will be higher.
Consistency

The same job ad can be published several times by the same source: copy/paste of the description, re-post of an old job ad.

The release date is an important feature to be considered during deduplication steps.
Topics

1. Context
2. Data processing pipeline
3. Vacancy detection with machine learning techniques
4. Deduplication issues
5. Date detection and vacancy’s duration
Challenging issues

Not in all job ads are explicitly report the release date and the expired date

• How detect, if reported, the release date?
• How cover all languages and formats of a date?
• How disambiguate between release date and expired date?
## Experimental case study

### Measure of quality

<table>
<thead>
<tr>
<th>Measure</th>
<th>Primary Job Portals</th>
<th>Combination of primary and secondary portals</th>
<th>Secondary Job Portals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Completeness</td>
<td>90%</td>
<td>77%</td>
<td>68%</td>
</tr>
<tr>
<td>Accuracy</td>
<td>80%</td>
<td>76%</td>
<td>67%</td>
</tr>
<tr>
<td>Complexity</td>
<td>82%</td>
<td>80%</td>
<td>64%</td>
</tr>
</tbody>
</table>

Complexity measures the availability of data structures and data formats.
Examples

Junior Java Developer (m/w)
idealo internet GmbH
5 days ago, Berlin
erste Programmierfähigkeiten in Java und eine hohe Motivation, neue Technologien zu erlernen - idealo ist eine Berliner Start-up-Erfolgsgeschichte.

Заведение в кв. Враждебна търси човек за работа на бара за дневна смяна

Враждебна, гр. София, Област София-град
Добавена от телефон в 16:17, 24 февруари 2018,
Номер на обявата: 108261351

Make up artist (m/ž)
15.02.2018 | Zagreb

Trenkwalder d.o.o. je agencija specijalizirana za zapošljavanje, odnosno ustupanje zaposlenika drugim tvrtkama. Dio smo međunarodnog koncerna koji posluje u 17 europskih zemalja te smo jedna od vodećih agencija na području Republike Hrvatske. Za Klijenta vršimo potragu za otvoreno radno mjesto:
Pattern matching

Find one of a specified set of strings in text.

A regular expression is a notation to specify a set of strings.

Reg-ex repository

Release date
Expire date

Raw Content (e.g. HTML source)

Content Meaning Filter
Preliminary results - Experimental Case Study
Multi-language, Multi-country, Single-source

<table>
<thead>
<tr>
<th>Language</th>
<th>NL</th>
<th>DE</th>
<th>FR</th>
<th>EN</th>
</tr>
</thead>
<tbody>
<tr>
<td>N.</td>
<td>3748</td>
<td>27158</td>
<td>5951</td>
<td>45135</td>
</tr>
</tbody>
</table>

**Job Ads - Language**

- **82%**
  - Release date from structured data
- **12%**
  - Expire date from structured data
Preliminary results - Experimental Case Study
Multi-language, Multi-country, Single-source

<table>
<thead>
<tr>
<th>Language</th>
<th>en</th>
<th>fr</th>
<th>de</th>
<th>nl</th>
</tr>
</thead>
<tbody>
<tr>
<td>N.</td>
<td>3748</td>
<td>27158</td>
<td>5951</td>
<td>45135</td>
</tr>
</tbody>
</table>

- **Job Ads - Language**
  - English (en): 82%
  - French (fr): 12%
  - Dutch (nl): 12%
  - German (de): 82%

- **Release date from structured data**: 100%
- **Expire date from structured data**: 54%

- **Date Detection**: 100%
Challenge

Ok, finding the release date is relatively simple. But the expired date of job ads is often absent. In many case, job portals leave close job ads when they expire.
Experimental case study
Measure of quality

<table>
<thead>
<tr>
<th>Measure of Quality</th>
<th>Primary Job Portals</th>
<th>Combination of primary and secondary portals</th>
<th>Secondary Job Portals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Timeliness and periodicity</td>
<td>72%</td>
<td>70%</td>
<td>80%</td>
</tr>
<tr>
<td>Complexity</td>
<td>82%</td>
<td>80%</td>
<td>64%</td>
</tr>
<tr>
<td>Accuracy</td>
<td>97%</td>
<td>74%</td>
<td>72%</td>
</tr>
<tr>
<td>Completeness</td>
<td>90%</td>
<td>77%</td>
<td>68%</td>
</tr>
</tbody>
</table>

Timeliness and periodicity of the data
Challenge: expire date auto-detection

Ok, finding the release date is relatively simple. But the expired date of job ads is often absent. In many case, job portals leave close job ads when they expire.

Data with outliers and asymmetry

**Inter-Quartile Range** is a measure of spread:

$$IQR = Q3 - Q1$$

A job ad is an outlier if:

duration is greater than \(Q3 + 1.5 \times IQR\)
Job ads duration cleaning
Level of application

• Multilevel approach
• IQRs area apply for
  • All job ads
  • NUT0 (e.g. country code)
  • NUT3 (e.g. regional area)
  • NUT3 / Esco Occupation Level 3
Preliminary results - Experimental Case Study
Multi-source, Single-country, Single-language

Q3 + 1.5 x IQR

111 days

18% Of jobs ads
Preliminary results - Experimental Case Study
Multi-source, Single-country, Multi-language

Q3 + 1.5 x IQR

50 days

2% Of jobs ads
Thanks for your attention