Session 2: Let the Data speak. Labour market information in transformation – Big Data analytics in application: Tunisia and Ukraine. Main conclusions. Visualisation of the results in interactive Dashboards.

Alessandro Vaccarino, Burning Glass Europe
Labour Market Information in Transformation | 10 December 2020
CONTENTS

1. Context and Goals
2. Methodological approach
3. Ukrainian and Tunisian experiences
4. Interactive Dashboard
1. Context and Goals
CONTEXT

**Continuously evolving** Labour Market:
- Digitalization of professions
- Relevance of Soft skills
- Internationalisation
- New professions and skills emerging
- Smart and Remote working
- Impact of Covid-19 pandemic
- ...

We need *something* that can help us monitor and analyze how LM is evolving, to support Decision Makers taking *the right decisions at the right time*
WHAT WE HAVE / WHAT WE NEED

We already have **official statistics**, that are:

- *Representative*
- *Strong in terms of value*

But we can benefit of additional, complementary information that could be:

- *Fast*, to track what’s happening now (e.g. Covid-19 Impact analysis)
- *Granular* and *adherent* to real and current market terms, to capture emerging trends analyzing what companies are actually looking for

How to find a similar, complementary source of information?

**Using Web Labour Market**
WHY WEB LABOUR MARKET

It’s the **exact representation** of what companies are looking in a given period:

- Up to date: companies publish an announcement *when* they actually need to hire
- Detailed: an announcement describes **as well as possible** the specific need, in terms of:
  - Profession needed
  - Requirements (skills, experience, educational level,…)
  - Working context (place, contract, sector, working hours,…)
- Adherent to reality: **market terms** are used, both for occupation and skills. This helps identify emerging terminology adopted by Market

It would be great to use those information in addition to better and deeper understand how Labour Market is evolving in a given country, even compared to other countries
OUR GOAL

Transform those...

...to this
2. Methodological approach
METHODOLOGICAL BACKGROUND

KDD – Fayyad, 1997
OUR APPROACH

Let’s take a deeper look on this framework
The process of obtaining and importing data from web portals and storing them in a Database

Focus on volume

Coverage augmentation

Balance between quality and effort
DATA INGESTION - GOALS

- Robustness of process
- Quality of data collected
- Scalability and governance
DATA INGESTION – ROBUSTNESS

**Issue:** potential technical problems when gathering data from a source (unavailability, block, changes in data structure)

**Risk:** loss of data

**Solution:** redundancy

- Have the most important sites (by volume and/or coverage) ingested from two or more sources
- Avoid loss of data in case of troubles with a source
- Collect data from both primary and secondary sources
DATA INGESTION – QUALITY

**Issue**: need to obtain data as clean as possible, detecting structured data when available

**Risk**: loss of quality

**Solution**: tailored ingestion. We collect data using a specific approach based on the single source:

- API
- Scraping
- Crawling
DATA INGESTION – QUALITY FRAMEWORK
DATA INGESTION – SCALABILITY AND GOVERNANCE

**Issue**: need to handle a real and complex Big Data environment, simultaneously connecting to thousands of websites

**Risk**: Loss of Process control and loss of OJVs due to slowness of the process

**Solution**:
- A **scalable** infrastructure
- A monitoring and governance **custom tool**
DATA INGESTION – RECAP

After this phase, we have web pages, most likely Online Job Advertisements. But they are:

- Noisy
- Duplicated
- Unstructured

As discussed, a proper source selection is strategic: it’s mandatory to identify the most relevant web portals, in terms of numbers, quality and informational value. How to ensure a proper selection?

With a Landscaping phase
The process of **cleaning** ingested data and **deduplicating** OJVs, to guarantee that analytical phase’ll work on data at the **highest quality possible**

- **Language detection**
- **Noise reduction**
- **OJVs Deduplication**
DATA PRE-PROCESSING – LANGUAGE DETECTION

Why:
• Each language has different keywords, stopwords,…
• It can reflect different cultures and Labour Market scenarios…

How:
• We trained 60+ specific classifiers based on Wikipedia corpus
• Models are accurate (~99% of precision) and fast to adopt

What we obtain:
• A fast and strong classification of the language used in each OJV
• A way to archive OJVs for which we don’t have language support
DATA PRE-PROCESSING – NOISE DETECTION

Why:
• In a Big Data environment, we must deal with noise
• Information gathered from the web, one of the most noisy place available

How:
• AI based models, similar to mail spam filters

What we obtain:
• Identification of:
  • Web pages explicitly not related to OJVs
  • Web pages disguised as OJVs
DATA PRE-PROCESSING – DEDUPLICATION

Why:
• Companies post several advertisements for each vacancy → Visibility
• It cannot affect analysis: no over-estimation due to multiple postings

How:
• Statistical-based approach: identification of the standard duration of an OJV
• Text-analysis to detect similar/identical advertisements

What we obtain:
• Unique OJVs, to ensure coherent analyses
Extract and structure information from data, with respect to the most proper taxonomy.

- Artificial Intelligence
- Taxonomy selection
- Information Linkage
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.
DATA CLASSIFICATION – TAXONOMY

Why:
• We need to formalize all our information, to make it consistent and enable analyses
• Occupations/Skills/Places/… must be related to a proper taxonomy
• A unique taxonomy for each dimension enables analyses across countries and projects

How:
• Selection on international and custom taxonomies, that fit Labour Market terms and enable
Most relevant taxonomies adopted:

- Occupation: ESCO/ISCO
- Skills: ESCO
- Places: NUTS and ISO
- Educational Level: ISCED
- Sector: NACE
- Seniority/Working hours/Contract type/…: custom taxonomies
DATA CLASSIFICATION – APPLICATION

Language detection
Data Pre processing
Ontology based models
Machine based classification
Ontology
Machine learning model
Classified item
3. Ukrainian and Tunisian experiences
In January 2020, we started a project to collect, classify and analyze data regarding Web Labour Market in both Tunisia and Ukraine.

The project followed the same methodology presented in the previous section.

- A Source Selection was performed and validated by our Country Experts.
- Data were collected, cleaned and classified on their own languages.
  - Specific classifiers were developed for both Ukrainian and Russian languages.
- Information collected was shared with Country Experts, to identify possible issues in the process and validate it.
UKRAINIAN AND TUNISIAN EXPERIENCES – WORKFLOW

Source ranking

Landscaping validation

Ontology selection

Model development

Country expert

Data collection

Data Pre processing

Data validation

Country expert

Data release
4. Interactive Dashboard
INTERACTIVE DASHBOARDS

You can find dashboards at:

- **Tunisia:**
  https://public.tableau.com/profile/tabulaex#!/vizhome/ETF-BigDataLMI-Tunisia/Time

- **Ukraine:**
  https://public.tableau.com/profile/tabulaex#!/vizhome/ETF-BigDataLMI-Ukraine/Time
INTERACTIVE DASHBOARDS – SOME INSIGHTS

Number of job vacancies collected

262,754

Number of job vacancies deduplicated

45,858

Number of unique Vacancies by Web Source

<table>
<thead>
<tr>
<th>Source</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>TN_TANITJOBS</td>
<td>23,250</td>
</tr>
<tr>
<td>TN_JOERA</td>
<td>8,008</td>
</tr>
<tr>
<td>TN_KEEJOB</td>
<td>5,439</td>
</tr>
<tr>
<td>TANQEEB</td>
<td>2,815</td>
</tr>
<tr>
<td>TNEMPLDINAT</td>
<td>2,543</td>
</tr>
<tr>
<td>TN_OFFREEMPLDOI</td>
<td>1,723</td>
</tr>
<tr>
<td>TN_JOB</td>
<td>820</td>
</tr>
<tr>
<td>TN_TUNISIEANNONCES</td>
<td>531</td>
</tr>
</tbody>
</table>

Number of job vacancies collected

385,207

Number of job vacancies deduplicated

238,974

Number of unique Vacancies by Web Source

<table>
<thead>
<tr>
<th>Source</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>UA_WORK</td>
<td>62,642</td>
</tr>
<tr>
<td>UA_RABOTA</td>
<td>54,613</td>
</tr>
<tr>
<td>UA_CAREERJET</td>
<td>31,992</td>
</tr>
<tr>
<td>UA_NEUVOO</td>
<td>29,775</td>
</tr>
<tr>
<td>UA_OXL</td>
<td>18,902</td>
</tr>
<tr>
<td>UA_HH</td>
<td>11,711</td>
</tr>
<tr>
<td>UA_INDEED</td>
<td>7,797</td>
</tr>
<tr>
<td>UA_JOBRAPIDO</td>
<td>7,470</td>
</tr>
</tbody>
</table>
INTERACTIVE DASHBOARDS – SOME INSIGHTS

**Working hours**

- Full time: 32.26%
- Part time: 7.88%
- Not defined: 59.86%

**Contract**

- Not Defined: 28.39%
- Permanent: 24.16%
- Temporary: 37.59%
- Self Employment: 4.07%

---

**Working hours**

- Full time: 59.95%
- Part time: 8.38%
- Not defined: 31.67%

**Contract**

- Permanent: 22.24%
- Self Employment: 0.28%
- Not Defined: 66.92%
INTERACTIVE DASHBOARDS – SOME INSIGHTS

### Occupation (level 1)

<table>
<thead>
<tr>
<th>Professionals</th>
<th>34.46%</th>
<th>Technicians and associate professionals</th>
<th>26.54%</th>
<th>Clerical support workers</th>
<th>11.02%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Service and sales workers</td>
<td>11.69%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Professionals</th>
<th>30.86%</th>
<th>Technicians and associate professionals</th>
<th>28.29%</th>
<th>Clerical support workers</th>
<th>9.51%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Service and sales workers</td>
<td>12.06%</td>
<td>Managers</td>
<td>7.78%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Craft and related trades</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>workers</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Plant and</td>
<td></td>
</tr>
</tbody>
</table>
Thank you very much

Alessandro Vaccarino, Burning Glass Europe
Labour Market Information in Transformation | 10 December 2020