An Open and Data-driven Taxonomy of Skills Extracted from Online Job Adverts

16 April 2019

Jyldyz Djumalieva and Dr Cath Sleeman
About our project

• ESCoE project 3.2
• Leveraging Big Data to improve understanding of the labour market
Building a data-driven taxonomy of skills

**Why?**
- Skill shortages are costly
- And skill needs are changing
- But we don’t measure skill demand or supply in a detailed or timely way
- First step to fixing this: create an open classification of skills

**How?**
- Take the skills mentioned in online job adverts
- Cluster the skills based on co-occurrence in the same advert
- Pros:
  - Objective
  - Easy to update
- Cons:
  - Work not advertised
  - Skills not mentioned
### Online job advert dataset

- 41 million adverts collected by Burning Glass Technologies (BG) in 2012 - 2017
- Over 11,000 unique ‘skills’
- Variables on position, geographic location, offered salary and requirements

<table>
<thead>
<tr>
<th>JobDate</th>
<th>CleanTitle</th>
<th>LAD</th>
<th>MinSalary</th>
<th>MaxSalary</th>
<th>Skills</th>
</tr>
</thead>
<tbody>
<tr>
<td>02/12/2012</td>
<td>Freelance Front End Developer - Reading</td>
<td>Reading</td>
<td>52,000</td>
<td>78,000</td>
<td>&quot;Advertising Design&quot;, &quot;Front-end Development&quot;</td>
</tr>
</tbody>
</table>
Building a data-driven taxonomy of skills

1. Count co-occurrences of skills in all job adverts
2. Train word2vec on skills across all adverts and calculate pairwise cosine similarities
3. Build skills graph
4. Identify and remove transversal skills
5. Group skills into hierarchical clusters

nesta
## Word2vec

<table>
<thead>
<tr>
<th>Skill</th>
<th>Mechanical engineering</th>
<th>Machine learning</th>
<th>German</th>
<th>Google adwords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 5 most similar skills in the model</td>
<td>Engineering support</td>
<td>Data science</td>
<td>Arabic</td>
<td>Keyword research</td>
</tr>
<tr>
<td></td>
<td>Process engineering</td>
<td>Artificial intelligence</td>
<td>Spanish</td>
<td>Search marketing</td>
</tr>
<tr>
<td></td>
<td>Manufacturing processes</td>
<td>Text mining</td>
<td>Dutch</td>
<td>Display campaigns</td>
</tr>
<tr>
<td></td>
<td>Equipment design</td>
<td>Natural language processing</td>
<td>Swedish</td>
<td>Link building</td>
</tr>
<tr>
<td></td>
<td>Technical drawings</td>
<td>Pattern recognition</td>
<td>Portuguese</td>
<td>Pay-per-click</td>
</tr>
</tbody>
</table>
Strength of relationships in the skills network

radiology $f = 1,585$

c = 0.66
diagnostic imaging
Detecting communities of skills

- Use Louvain community detection algorithm with cosine similarity as weights
- Progressively split clusters to detect hierarchical relationships
- Use bootstrapping for ensure robustness

For 500 iterations:
- Sample edges from original graph
- Build sub-graph
- Find clusters in the sub-graph with best modularity

Generate consensus cluster assignments from 500 iterations using HGPA and MCLA
data engineering

key skills
- Python, optimisation, big data, Ruby, NoSQL

job titles
- Developer, Java developer, Devops engineer

median salary: £55,000 (£43k - £73k)
The value of skills

annual salary offered
lower quartile | median | upper quartile

demand for skill cluster
low | high

LOW SALARY

information technology £45k

business administration £35k

HIGH SALARY

elle & data warehousing
software development
system administration
networks
web development

financial asset management
procurement
accounting & financial management
business management

retail management
office administration
Additional considerations/constraints

- Identify and set aside highly transferable skills to detect communities → taxonomy focuses on more technical skills
- Ignore temporal aspect for now → might not capture changes over time
- Clean and normalise job titles to enable linking to skills and offered salary
- Stop at level three, but could continue splitting skill clusters
- Manually labelling skill clusters
- Do not explore qualifications and experience requirements
Contributions of this work

• Deliver the data-driven taxonomy of skills that doesn’t rely on expert judgement
• Produce first ever estimates for the value of skills
• Link official occupations to skills (using SOC to skill cluster crosswalk)
What’s next

• Extend the taxonomy with estimates for qualification and experience requirements
• Measure skills demand and supply at regional level
• Study evolution of skills over time to identify transferable and new skills
Thank you

Jyldyz Djumalieva @d_jyldyz
jyldyz.djumalieva@nesta.org.uk

Dr Cath Sleeman @CathSleeman

November 2018

nesta