What's in a Job? Measuring Skills From Online Job Adverts

Gueorguie Vassilev

(Office for National Statistics)

gueorguie.vassilev@ons.gov.uk

Oleksii Romanko

(King's College London)

oleksii.romanko@kcl.ac.uk

Khloe Evans

(Office for National Statistics)

khloe.evans@ons.gov.uk

Paper prepared for the 36th IARIW Virtual General Conference
August 23-27, 2021
Session 2: The Potential and Challenges of Big Data and other Alternative Data in the Production of Prices, National Accounts, and Measures of Economic Well-Being
Time: Tuesday, August 24, 2021 [14:00-16:00 CEST]

What's in a job? Measuring skills from online job adverts

Authors: Gueorguie Vassilev (Office for National Statistics), Oleksii Romanko (King's College London) & Khloe Evans (Office for National Statistics)

Contents

1. Introduction

The Office for National Statistics (ONS) have long aspired to use more novel data sources, such as online job advert data, to derive an alternative skills-based measure of human capital (ONS, 2018).

Gaining a better understanding of the skills and knowledge, or human capital, of the UK population has become an increasingly important policy question in recent years, which is unsurprising given it is a key determinant of economic success, (Institute for Fiscal Studies, 2020). The established links between improving skills and increased growth, productivity, and earnings are also well documented (OECD, 2017). The specific motivation behind improving the existing measures of human capital is driven by growing user interest, initially highlighted by the Chancellor of the Exchequer in the 2018 Spring Statement who asked ONS to 'develop a more sophisticated measure of human capital so that future investment can be better targeted' (Hammond, 2018).

More recently, as the UK looks ahead to ensuring a strong recovery from the impacts of the Covid-19 pandemic, improving skills is at the forefront of the 'Build Back Better' initiative. There is also renewed focus on 'levelling up' with the 'most important pillar in the approach to levelling up to be supporting individuals across the UK to reach their potential', through investing in people and improving their skills (HM Treasury, 2021).

This online-job-advert driven skill demand approach could allow ONS to measure the 'demand side of human capital' (ONS, 2018), by extracting skills from job descriptions to understand how demand for specific skills is changing over time. As well as providing an aggregate view of employer skill demand, this analysis may then be expanded to consider the relationship between skill requirements and proposed salaries. Using statistical techniques, an equivalent market price could then be generated for each skill, or group of skills. 'These derived relationships between salaries and skills could then be applied to our estimates of human capital stock to work out a current stock of different skills, allowing users to better understand which skills are important for particular sectors of the economy' (ONS, 2018). These analyses could also be used to inform policy on investment in people at a more granular level, 'as well as to inform curriculum design in our education systems' (ONS, 2018).

This alternative measure of human capital would go beyond the current ONS estimates which are derived as a stock measure, using a discounted lifetime earnings approach. Under this approach, human capital is estimated by 'looking at what qualifications people have and what they earn as well as how much longer they will continue to work (ONS, 2019).

However, measuring skills in the first instance is notoriously difficult, with challenges arising from measurement and definition errors and self-reporting bias of non-cognitive skills (Kautz et al, 2014). This has led to a growing appetite for using alternative data sources to traditional surveys, such as online job adverts data, to better understand employer skill demand, and in turn, how to support individuals to learn about the skills

they need. For example, NESTA developed the first data-driven skills-based taxonomy of UK occupations using online job advert data (Djumalieva and Sleeman, 2018) to allow policymakers to better understand how occupations are changing. In addition, OECD have used online job advert data to better understand how the skills individuals require today, differ from those they will need in future, as well as identifying the top transversal skills (OECD, 2021).

This paper elaborates on ONS' aspirations for measuring skill demand using online job adverts, detailing progress to date, key challenges and future organisational plans. In section 2, the current data source used by ONS is discussed, alongside the general strengths and weaknesses of online job adverts. In section 3, appropriate natural language processing techniques are considered and the currently implemented methodology for extracting skills is outlined. Section 4 includes some initial experimental results from the application of the skill extraction algorithm, as well as highlighting validation carried out to date. Section 5 highlights the user requirements identified to date for further work from such analyses as well as broader ONS dissemination plans, and finally, section 6 provides an overall conclusion and outlines next steps for the ONS in this space.

2. Data sources

This section describes the online job advert data which ONS currently has access to, as well as the strengths and weaknesses of online job advert data in general, before highlighting the specific structure of ONS' current dataset.

2.1 Adzuna data and its use to date

ONS is currently working in partnership with Adzuna, an online job search engine that collates information from thousands of different sources in the UK (Adzuna, 2021). 'These range from direct employers' websites to recruitment software providers to traditional job boards, providing a comprehensive view of current online job adverts' (ONS, 2021a).

Following the onset of the Covid-19 pandemic, ONS has used the Adzuna data to produce weekly experimental job advert indices, for example see (ONS, 2020), (ONS, 2021b), which can provide a proxy measure of how labour demand has changed over the course of the pandemic. Producing these experimental breakdowns of job advert indices by Adzuna category[1] and by Nomenclature of Territorial Units for Statistics (NUTS 1)[2], has allowed better understanding of the economic impact of the pandemic and post-pandemic recovery to be tracked in a timelier and more granular way than is possible using official statistics from the ONS Vacancy Survey.

[1] 'Adzuna uses a neural network to assign categories to the job adverts. The model uses natural language processing to analyse the text in both the job title and description fields and uses the data to assign the most suitable job category' (ONS, 2021a)

[2] ONS allocates regions to job adverts using a combination of manual review, postcode matching and text matching locations to existing geography look-up files (ONS, 2021a)

The ONS Vacancy Survey is a 'statutory monthly survey of businesses' which asks, 'how many job vacancies did a business have in total (on a specified date)' (ONS, 2012). The data are available by Standard Industry Classification 2007[3] and by employment count but are not available by NUTS 1 regions. In terms of timeliness, the data are published 'within 6 weeks of the reference date of the survey' (ONS, 2012).

The timeliness of the experimental Adzuna data meant that an indicator of post Covid-19 labour demand recovery was first visible on 29th April 2021 when online job adverts 'for the first time exceeded its February 2020 average level' (ONS, 2021c). This recovery was not visible in the official statistics until 15th July when figures showed that 'the number of job vacancies in April to June 2021 was 9.9% above its pre-pandemic level in January to March 2020' (ONS, 2021d). Additionally, the NUTS1 breakdowns available in the Adzuna data have allowed us to understand regional differences whereby London has recovered more slowly than other regions which would not be possible through official statistics.

In addition to its usefulness of providing timely, proxy reporting of changing labour demand, online job adverts provide a rich data source that allows for a wide scope of analyses. With detailed job descriptions often providing granular information such as key responsibilities, the job title, the salary and benefits available, the advertising company as well as details of the skills, knowledge and qualifications required from the applicant, the scope of potential insights that can be extracted is vast. For example, recent extraction of key-word phrases allowed ONS to publish trends of adverts mentioning home-working, or similar flexible working arrangements (ONS, 2021e).

2.2 Strengths and weakness of online job advert data

Despite the many strengths in online job advert data such as timeliness, frequency and low-level detail, there are also accompanying challenges in working with these data. These strengths and weaknesses are outlined in Table 1 before being discussed further below.

Table 1: Strengths and weaknesses of online job advert data

Strengths	Weaknesses
Timeliness: Extracts are taken in real-time and published 6 days later with ONS publication and processing schedules	Duplication: Elaborated on below; it relates to multiple entries in the dataset for the same post. Reverse duplication refers to multiple posts with one advert.
Frequency: Extracts are currently taken weekly, but this is restricted only by ONS resources.	Coverage: Although a high proportion of adverts will be covered, there are gaps such as shop window advertisements, word of mouth and head hunted posts.

[3] 'The current Standard Industrial Classification (SIC) used in classifying business establishments and other statistical units by the type of economic activity in which they are engaged' (ONS, 2016)

Detailed information available: As described above the job descriptions are usually very detailed offering an abundance of information.	Sensitivity: Changes in trends may be influenced by recruiter behaviour instead of genuine changes in the number of job adverts.
	Recruiter dependency: There may be a time delay from when a post is filled, to when an advert is withdrawn by a recruiter

Due to the nature of the data, a job advert may be advertised by multiple different recruiters, which means that duplication is a key challenge. Although ONS have taken initial steps to identify and remove duplicate adverts using document similarity detection methods (ONS, 2021a), some duplicates remain in the data, potentially resulting in an inflated value of job adverts, if considered as a measure of vacancies. Additional weaknesses in the dataset include coverage limitations as not all vacancies will be advertised online and that adverts do not provide a direct measure of labour force demand and may at times be influenced by changes in recruiter behaviour. It is worth noting official sources of vacancies may also suffer from some coverage issues such as very small businesses, or vacancies that open and are filled between collection periods.

Despite these challenges, the potential of using online job adverts to measure employer skill demand warrants further investigation and the potential methods for analysing these data are discussed in the next section.

2.3 Data structure

Overall, ONS has access to ~140 million job adverts covering a time period from February 2018 to August 2021 and counting. Fields available in the data include Adzuna category, job description, location, salary, and job title. Excluding the Adzuna category, each of these fields are unstructured free text, thus requiring natural language processing.

The variable used for extracting skills is the job description. On average, the job description is 324 words long, although it ranges from just two, to thousands of words across the dataset, and often contain a lot of extra noise that is not necessary for such extraction, such as company information or generic text used by the recruitment agency.

The dataset is also subject to some missingness across the variables. For job descriptions, this rate is tiny with only 0.0001% of adverts in the entire dataset with a missing job description. This low missing rate is also true of other variables that ONS plan to make use of, except for the salary field which has a higher missing rate of 34%.

Due to the size and unstructured nature of this dataset, ONS are taking advantage of distributed computing systems to process the data. As such, any methodology defined in future is required to be compatible with these distributed systems, and thus this is a required part of future development work, which is fully discussed in section 6.

3. Methodology

The analysis of job advertisements is not a new concept, with a review undertaken in 2012 highlighting that using 'job adverts as a data source for research' was a considered methodology used as early as the 1970s (Harper, 2012). In the review, it was highlighted that the number of studies using these data was increasing and although many of the studies relied on manual intervention with researchers reading adverts and extracting themes themselves, the deployment of 'text analysis software' was also mentioned briefly.

Text analysis has only increased in popularity since then as more and more electronic text has become available. In this section of the paper, the text analysis methods considered by ONS are discussed, and a detailed overview of how the current skill extraction methodology was implemented is provided.

3.1 Text analysis methods

Many different methods exist across different software packages to analyze text data, although the core principles of text analysis remain the same. The focus of this paper will be on a specific subset of text analysis, natural language processing (NLP).

'Natural language processing is the computerized approach to analyzing text that is based on both a set of theories and a set of technologies' (Liddy, 2001). The field itself has been and continues to evolve over time, but overall, the aim of NLP is to 'understand the linguistic use and context behind the text' (Pure Speech Technology, 2021) which is particularly helpful when working with online job adverts, as context is important to identify useful information from the lengthy job descriptions.

Some common NLP techniques and their application using online job adverts are discussed forthwith. These techniques are discussed as previous research using online job adverts data specifically exists, but there are other techniques outside of natural language processing, which could also be applied to extract further insights from the data such as key word matching or generative adversarial networks. Although these alternative methods have not yet been considered fully, this will form part of the longer-term development work for ONS in future.

3.1.1 Topic modelling

Topic modelling is an unsupervised learning method which tries to learn underlying structures in text by clustering different words by assigning them to a 'topic' (Sergey, 2017). The topics are not defined, and the algorithm goes through each word iteratively, reassigning them by considering the probability that a word belongs to a topic. A common form of topic modelling is Latent Dirichlet Allocation (LDA) which can assign each text to a mixture of topics.

Topic modelling can be extended to topic classification, which is a supervised version of the algorithm which requires the topics of a set of texts before analysing them (Pascual, 2019). Using these topics, data is tagged manually so that a topic classifier can learn and later make predictions by itself.

Recent analysis (Arthur, 2021) has applied supervised topic modelling to extracts of Reed data in order to identify job classes such as ‘delivery’, ‘nursing’ and ‘hospitality’. The approach taken was to combine job titles and descriptions and manually classify a proportion to their classes. This manually labelled data was then used to train a classifier which could be applied to the entire dataset. A similar approach could be taken to create a skills topic classification.

In 2018 the Bank of England also used Reed data and applied an LDA topic model (Bank of England, 2018) to assign job descriptions to different topics based on their similarity.

Although a useful method to consider, and conceptually applicable as the dataset ONS are working with has both many documents to apply the method to and relatively long job descriptions which are favourable (Tang et al, 2014), more investigation is required before applying this method for the identification of skills. In particular, the number of topics needs to be identified prior to applying topic modelling, and iterative work would be required to arrive at an optimal number of topics. Additionally, there is the consideration of whether skills could be identified accurately as one ‘topic’ given the vast range of possible skills included i.e., technical, digital, soft, language etc. Hence, further work would need to be undertaken to test the feasibility and relative advantage of this method against others.

3.1.2 Word embedding algorithms

Word embedding algorithms are a set of various methods, techniques, and approaches for creating natural language processing models that associate words, word forms or phrases with number vectors, i.e., ‘numerical representations of semantic units’ (Alvarez, 2017). The principles of word embedding are that words that appear in the same context have similar meanings.

Examples include Word2Vec models which predict the probability of a word by its context. The model will train by vectors of words in such a way that the probability assigned by the model to a word will be close to the probability of its matching in each context. Word2Vec uses neural networks to calculate word embedding based on words’ context.

Word embeddings can be used in conjunction with other techniques such as Long Short-Term Memory deep learning networks to extract skills from job adverts (Sharma, 2019). In this approach, the relatively small training dataset is built up using a subset of noun phrases taken from job adverts, with some promising results produced.

In general, use of word embedding algorithms seems a useful approach to take, and hence this is the initial algorithm that has been applied to ONS’ online job advert data. However, currently, specific labelled noun phrases have not been used to build the algorithm, due to resource and time constraints. This is something that will be considered in future, but initially, a pre-defined list of skills has been used instead. The full implementation of this word embedding algorithm is outlined in section 3.3.

3.2 Data preprocessing

Prior to applying a word embedding algorithm, data pre-processing is carried out. It is widely documented that data cleaning is a vital step to carry out prior to any application of text analysis (Towards data science, 2020). ONS have implemented a number of core, widely cited data cleaning techniques to Adzuna job descriptions, to improve the accuracy of analysis. These techniques include:

- *Standardising text to ensure all words are lower case*

- *Removing punctuation*
  Regular expressions have been used to remove any punctuation as well as to recognize any URLs which are commonly found in online job adverts.

- *Tokenization*
  Sentence tokenization has been applied as a way of separating each job description into a list of individual sentences, represented as 'tokens'.

- *Removing stop words*
  Currently the list of stop words removed are consistent with the pre-defined list provided in the Natural Language Toolkit python package (NLTK, 2021). However, further improvements of the data pre-processing methods could be to expand this list to also include common words or phrases specific to online job adverts such as 'company', 'opportunity' etc. to further reduce noise in the data.

Applying these techniques produces a job description which has some noise removed and is in a format that can be efficiently used as an input into a word embedding algorithm. A future pre-processing step considered but not yet implemented is lemmatization, which is 'the process of converting a word to its base form' (Prabhakaran, 2018) which would allow each word to be analysed as a single term.

3.3 Application of a word embedding algorithm

The method that will be the focus of this paper is using a word embedding algorithm in conjunction with a pre-defined list of skills, which has been developed in partnership with colleagues in King's College London (Romanko, forthcoming with permission). The detail of the method and its implementation is covered here, and the next section reviews some initial results of the algorithm once it has been applied to a small subset of online job adverts.

The algorithm structure and methodology were based on previous work which also aimed to extract skills from job descriptions (Van-Duyet et al, 2017) as well as several other

open-source python repositories.[4] The main differences between the approach outlined in this paper and the previous work is that a more advanced Word2Vec model 'fastTEXT' has been applied (fasttext, 2020) with a continuous bag of words set up, which is found to train faster than the alternative Skip-Gram approach (Riva, 2021). To produce the word embedding algorithm, the following steps were taken:

- *Identify an initial list of skills*

To build ONS' initial word embedding algorithm, an initial skill list was used to produce a training dataset. For the purposes of this analysis, skills included on the website Dice were used. Dice are an American company providing a database for technology professionals (Dice, 2021) and the skills were created using a data-driven (using job adverts and labour market information) and expert-driven approach (using existing taxonomies). As well as including around 11,000 skills on their website, Dice also regularly publishes insights into the job market with a specific focus on the technology sectors. To provide additional context of what the skill entailed, descriptions were taken from Wikipedia (following recommendations from existing research (Zhao et al, 2015)) which provided extra semantic information of each skill to create a skills dictionary. This process involved dropping any description for skills unassigned to articles and prioritizing the best description for the skill based on key-word based weighting if there are multiple descriptions assigned.

ONS recognize that the results of the skill extraction algorithm are heavily dependent on the original list of skills used to train the algorithm and acknowledge that the currently chosen list from Dice may be subject to bias towards technological skills. Considering this, colleagues in King's College London working in partnership with ONS (Romanko, forthcoming with permission) are undertaking a wider review of available skills taxonomies to determine their strengths and weaknesses. The current methodology has been designed to be fully adaptable in future to potentially utilize an alternative skill list. More details on the review of skills taxonomies are included in section 5.2.

- *Develop a neural network algorithm*

A neural network algorithm was trained to learn word associations from the list of skills and their descriptions (skill dictionary). The model chosen was 'Word2Vec' which produces a vector space where each unique word in the original list of skills and descriptions is assigned a corresponding vector in the space (Gilyadov, 2021). 'The model can extract both acquired skills (e.g., tutoring) and technologies required from prospective employees (e.g., Microsoft Word)' (Romanko, forthcoming with permission).

The model works by producing combinations of words from job descriptions (n-grams) which are then used to search for skill-related phrases in the skill dictionary produced in

[4] https://github.com/Msq-9/Extraction-of-Skills
https://github.com/duyet/skill2vec
https://github.com/workforce-data-initiative/skills-ml/blob/master/examples/TrainEmbedding.py

step 1. The model uses a metric called cosine similarity distance to determine words which have a similar semantic context with those in the skill dictionary, and these words are located in close proximity to one another in the vector space.

The result of the algorithm is such that it returns a list of skills identified in the job description. The text values of the skills correspond with the values which exist in the pre-defined list of skills sources from Dice.

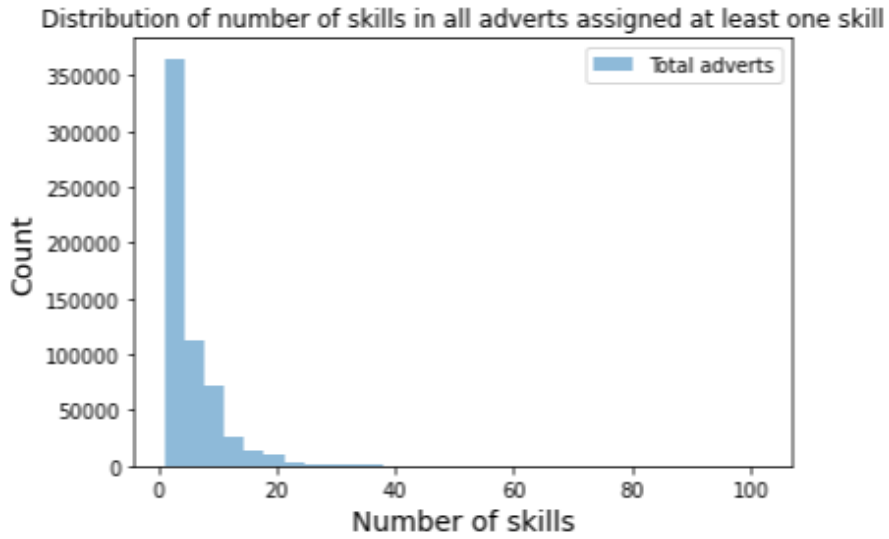
4. Initial results

The word embedding algorithm described in section 3 was applied to one extract of Adzuna data taken on 23rd July 2021, accounting for ~1.2 million job adverts. The results of this application are discussed in this section.

4.1 High-level summary statistics

The algorithm successfully extracted at least one skill for 51% of the job adverts in the weekly extract. For those job adverts which had skills identified, the number of skills identified ranged from 1 to 102, with the average number per advert being 3 skills. The full distribution of the number of skills identified per advert is shown in Figure 1 below.

Figure 1: Frequency counts for the number of skills identified per advert



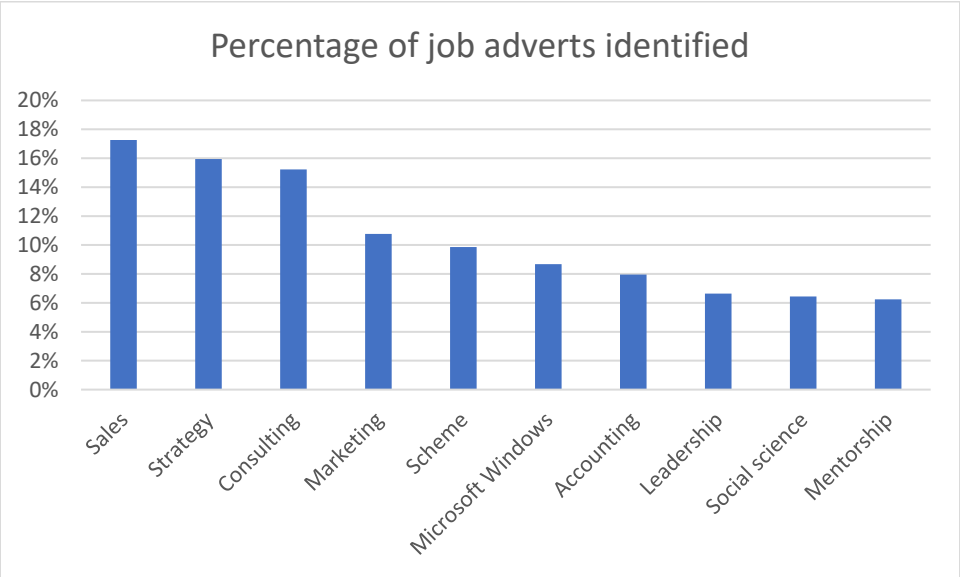
Notes:

1. *Source is Adzuna*

The top five most common skills identified in job adverts by the algorithm were ‘Reporting’, ‘Recruitment’, ‘Sales’, ‘Strategy’ and ‘Consulting’, however, as discussed fully in section 4.2, the skills ‘Reporting’ and ‘Recruitment’ are likely to in part represent non-skill related parts of the job adverts. They are therefore excluded from analysis from this point forward.

The most common skills identified by the algorithm, shown as the number of adverts the skill is identified in, as a percentage of the total adverts is shown in Figure 2.

Figure 2: The 10 most common skills identified, shown as the number of adverts the skill is identified in, as a percentage of the total number of adverts



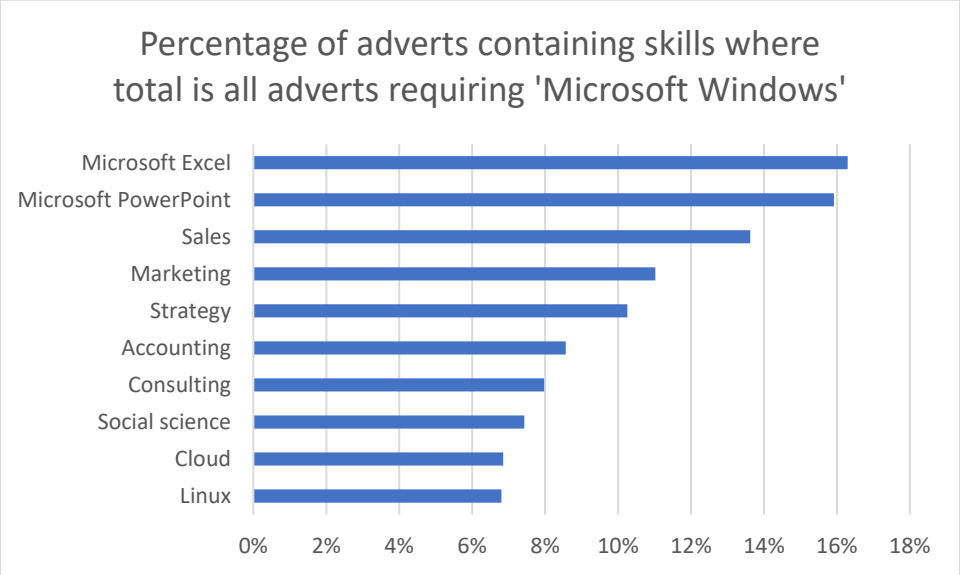
Notes:

1. *Source is Adzuna*
2. *'Reporting' and 'Recruitment' skills are excluded*

When considering the most commonly identified skills by the algorithm, the most frequent skills which appeared alongside said skill were analyzed with some interesting results. For example, of all adverts which the algorithm identified as requiring 'Sales' skills, 43% were also identified as requiring 'Marketing' skills. Other skills which commonly occurred in adverts looking for 'Sales' skills were 'Pricing', 'Finance' and 'Presentation'.

For those adverts identified as requiring 'Microsoft Windows' skills, other software skills were frequently included alongside such as 'Microsoft Excel' and 'Microsoft PowerPoint' as shown in Figure 3, though there was less dominance of a particular other skill.

Figure 3 The most frequently occurring skills identified alongside 'Microsoft Windows'



Notes

1. Source is Adzuna.
2. 'Reporting' and 'Recruitment' skills are excluded.

The most uncommon skills identified by the algorithm were also considered, and mainly related to specific IT skills such as the programming language 'MATHLAB' or 'Wolfram Mathematica'. However, softer skills such as 'communication skills' were also not commonly identified, with the results showing only one job advert including communication skills as a requirement. This is likely to reflect the performance of the algorithm, rather than a lack of job adverts mentioning communication skills and is discussed further in section 4.3.

When considering the 49% of job adverts for which no skills were identified, some manual investigation was undertaken to understand if the advert contained skills that should have been identified by the algorithm and help spur potential future improvements. There were several reasons for skills not being identified:

- The adverts had not listed specific skills required for the post, and instead focused on wider related requirements. For example, experience e.g., '10 years' experience in industry X required' or specific qualifications. There were also many adverts which listed personal attributes instead of specific skills, for example, 'reliability', 'friendly', 'hard working' or 'having a positive attitude'. Longer term, ONS would aspire to extract these kinds of information as well as the core skills, and this will be considered further in the next steps outlined in section 6.

- There were many adverts which listed 'soft skills' that were not identified by the algorithm, for example 'good communication skills', 'team working', 'time

management' and 'attention to detail'. It is important to note that this could be partially driven by the chosen skill list from Dice, which is biased towards technical skills. Identifying 'soft skills' is a key requirement for ONS; hence this will be a focus of improvement as outlined in section 6.

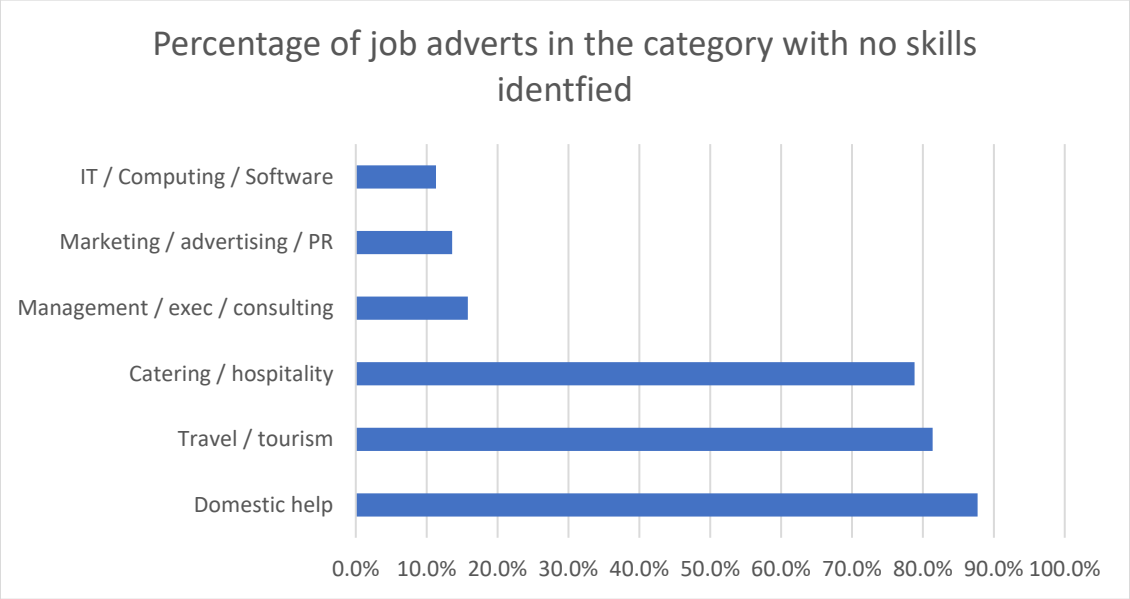
The 49% of adverts for which no skills were identified are further considered in section 4.2 where category and regional breakdowns are used to inform if there is any bias in the adverts most likely to be assigned skills by the algorithm.

4.2 Category and regional breakdowns

4.2.1 Breakdowns of adverts with no skills identified by algorithm

Figure 4 shows that the proportion of adverts for which no skills were identified varied substantially by Adzuna category, with the Domestic Help and Travel/Tourism categories most likely to be affected, at 88% and 81% of respective adverts having no skills identified by the algorithm.

Figure 4: The percentage of online job adverts for which no skills were identified, by Adzuna category



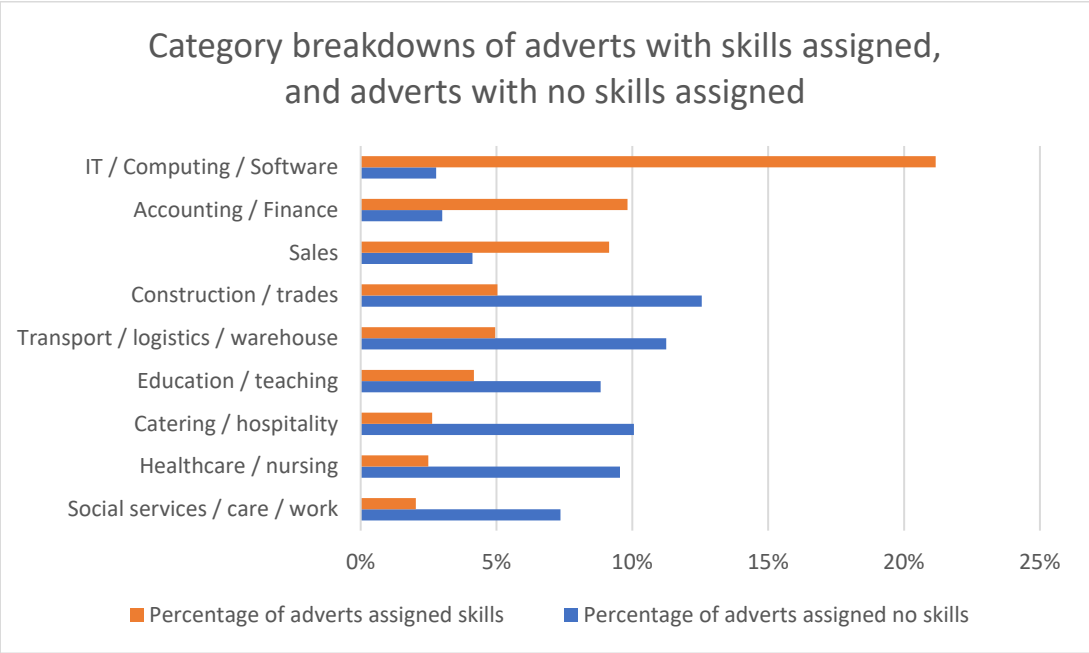
Notes

1. *Source is Adzuna*
2. *Does not include all Adzuna categories*

Considering the overall category distribution of job adverts with at least one skill identified and job adverts with no skills identified simultaneously can inform potential biases. Figure 5 suggests the algorithm is concentrated towards the IT/Computing/Software category as it accounts for 21% of adverts for which the algorithm identified skills and only 3% of adverts for which the algorithm did not identify any skills. On the other hand, potentially

skills in less IT-based roles in the care and hospitality categories are less well-identified, as they are under-represented compared with all adverts.

Figure 5: Category breakdowns of adverts with skills assigned, and adverts with no skills assigned



Notes

1. *Source is Adzuna*
2. *Does not include all Adzuna categories*

Considering potential regional bias, the distributions of regions for adverts with skills assigned and adverts with no skill signs are all broadly similar except for London. London accounts for 23% of all adverts for which skills are identified, but only 15% of adverts for which no skills are identified, suggesting a slight bias towards adverts based in London. However, this is likely driven by the nature of the London job market as a substantial share of job adverts in London were in the IT/Computing/Software category.

4.2.1 Breakdowns of average number of skills

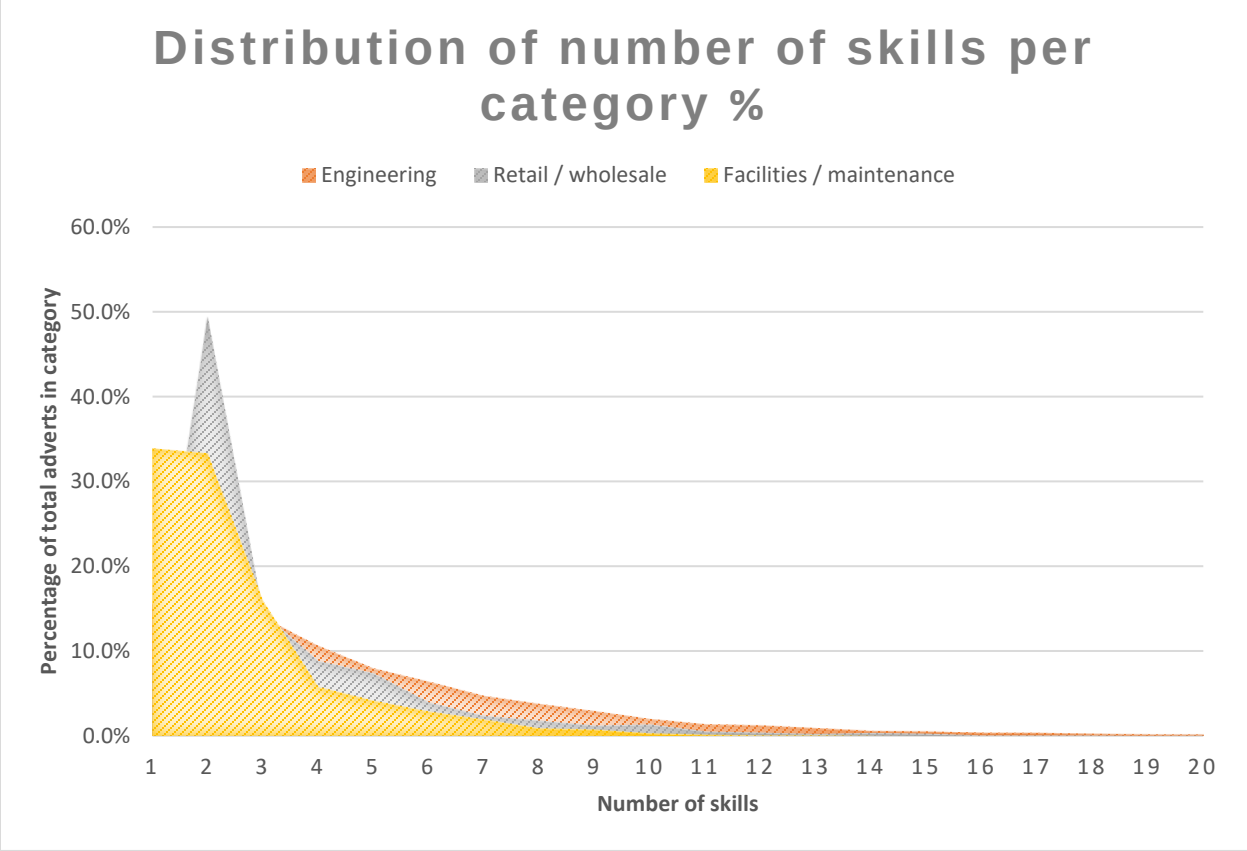
There were considerable differences in the average number of skills identified between different Adzuna categories, partially influenced by the vast differences in categories reporting no skills. The IT/Computing/Software category required the highest number of identified skills, with an average of 8, which was also the category least likely to report no skills, with only 11% of adverts having no skills identified.

Marketing/advertising/PR, Scientific/QA and Management/exec/consulting categories all required around 5/6 skills, whereas categories such as healthcare and social care and transport/logistics/warehouse required on average only one skill.

The distribution of number of skills also varied between categories. For all categories, most adverts had less than ten skills, but some categories had a wider distribution than others. For example, Facilities/maintenance has the smallest distribution of number of skills ranging from 1 to 14, whereas the Retail/Wholesale category has between 1 and 33 and the Engineering category has between 1 and 66. Figure 6 shows the different distributions of the number of skills, although the number has been capped at 20 due to the small proportion of adverts with more than 20 skills identified.

Most categories fall in a similar range to the Facilities and Maintenance, although the IT/Computing/Software category has the largest range of between 1 and 102 skills identified per advert, with 7.6% of adverts identifying more than 20 skills, compared with just 0.6% for the Engineering category.

Figure 6: Distribution of number of skills identified in each category, for 'Facilities/maintenance', 'Retail/Wholesale' and 'Engineering', as a percentage of total adverts in the category



Notes:

1. *Source is Adzuna*
2. *Number of skills capped at 20 although there are adverts in the categories with more than 20 skills identified*

As to be expected, there were also interesting differences in the most common skills requested in each category. For example, as shown in Figure 7 below, unsurprisingly, the top sought after skills for jobs in the Accounting/Finance category were ‘Accounting’, ‘Reporting’, ‘Finance’, ‘Taxes’ and ‘Audit’.

Figure 7: Word cloud showing the most frequently occurring skills identified in job adverts in the Accounting/Finance category

Notes:

1. *Source is Adzuna*
2. *‘Reporting’ and ‘Recruitment’ skills have been excluded*

For jobs in the creative design/arts & media category, ‘Graphics’ and ‘Adobe Photoshop’ were amongst the most common skills identified, as shown in Figure 8.

Figure 8: Word cloud showing the most frequently occurring skills identified in job adverts in the Creative/Design/Arts & Media category

Notes:

1. *Source is Adzuna*
2. *‘Reporting’ and ‘Recruitment’ skills have been excluded*

When looking at less industry defined roles, such as those in the graduate category, the top skills requested were ‘Consulting’, ‘Sales’, ‘Recruitment’ and ‘Marketing’, perhaps providing insight into the types of industries to offer graduate positions.

As alluded to in section 4.1, there were also some surprises which warrant further investigation, such as ‘Recruitment’ and ‘Reporting’ consistently being identified as the most common skills requested in job adverts. Although this could be appropriate for positions in the HR & Recruitment or Accounting/Finance categories, it seems less likely these are the top skills asked for in Domestic Help or Education positions where they were also identified. This therefore seems more of a limitation in the way that the algorithm is currently applied, whereby common phrases found in online job adverts such as ‘we are recruiting for’ or ‘you’ll be reporting to’ are being misconstrued as skills. This presents a challenge to address, as expanding the existing data cleaning methods in place to include such phrases as stop words could reduce this noise from the dataset but also runs the risk of removing genuine instances of requests for recruitment/reporting skills. Plans to improve the current skill extraction algorithm to address challenges like these are included in section 6.

Regional breakdowns were also considered, although it is important to note that regional differences may be partially explained by differences in the industrial structure of different regions as some regions have a higher proportion of jobs advertised in specific categories.

The average number of skills identified in an advert did not vary substantially between regions, ranging from 2-4 skills. London and Northern Ireland were the only two regions for which four skills on average were identified.

The most common skills identified in adverts were also much more homogeneous across the different regions, with ‘Sales’, ‘Consulting’ and ‘Strategy’ included in the top 5 most common skills for almost all regions.

London was the only region for which ‘Leadership’ skills were in the top 5 most common skills sought after and Northern Ireland was the only region for which ‘Risk management’ and ‘Compliance’ were in the top 5 most common skills sought after.

4.3 Initial validation

To perform further validation of the algorithm’s performance, in addition to the summary statistics outlined above which provide a sense-check at a high-level, a subset of 30 job adverts were manually quality assured. It is important to note that this is a very small sample size for quality assurance, so no strong inferences can be drawn as to the performance of the algorithm overall and the authors may be applying subjective judgement. However, this quality assurance is still useful to identify some issues early on. As part of ONS’ future development work, a more extensive validation will be carried out which can then inform further improvements to the algorithm.

For each of the 30 adverts, it was considered if the skills extracted were correct, if there were any skills that were identified but shouldn't have been and if there were any missing skills.

4.3.1 Correctly identified skills

In all manually verified cases, the algorithm identified at least one correct skill. In 56% of cases, all skills identified by the algorithm were correct (although this does not mean the algorithm identified all skills – see section 4.34).

The algorithm was particularly accurate for jobs in the IT and creative categories, where it was able to identify specific programs and programming languages. The algorithm also successfully extracted language requirements well, for job adverts requiring fluency in English, Chinese, Arabic etc.

4.3.2 False positive matches

There were several instances where a phrase relating to a non-skill topic was identified as a skill, which will need to be addressed as part of further refinement of the model.

A prevalent example of this seen in multiple job adverts was the word 'social' which is often used in different contexts in job adverts such as highlighting regular social activity as a company benefit, using social media to recruit, or relating to encouraging applications from different social backgrounds. In each of these instances, the algorithm is identifying the context as a requirement for 'social science' skills.

As described in section 4.2, there is a known issue with false relevance matches such as 'recruitment' and 'reporting' skills regularly being identified when the semantic context in the advert is not related to a skill requirement. However, there were also instances where other noise in the data were being identified as skill requirements. For example, 'opportunities to learn other disciplines such as photography' is identified as a requirement for photography skills or 'you will be mentored by a team leader' is identified as a requirement for mentoring skills.

These nuances represent real challenges in working with online job adverts which by default, often contain a lot of noise in terms of company information and job benefits. These may also be further challenged by lemmatization as a standardisation technique. There is also a conceptual angle to consider around whether if a job advert lists key duties as using specific programs or techniques, but doesn't list them as specific skill requirements, they should be recorded as such given the potential role of on-the-job training. Depending on the use of such skills extraction, the answer may differ. For example, to understand skills-occupation relationships, one would likely want to account for such information as skills, while tracking skills needs and how they can be met by the skills of the current workforce may not require such training-led parts of the job.

4.3.3 Missing skills

In terms of skills that weren't identified by the algorithm, in 80% of the adverts manually reviewed, at least one skill was missing. As highlighted in section 4.1 when reviewing the

adverts for which no skills were identified, the most common skills not being picked up by the algorithm were the ‘soft skills’. Most prominent in this category were communication/interpersonal skills which were mentioned (but not identified by the algorithm) in 40% of the adverts that were manually reviewed. Other skills not identified included time management, organization, attention to detail and teamworking skills. Additionally, broader skills such as ‘analytical’ and ‘problem solver’ were not identified. This is despite some of these skills being part of the original list used to build the model in step 1.

4.4 Parameter sensitivity

There are a number of parameters embedded in the skill extraction process, which when adjusted, can influence results. More work is required to fully understand the optimal level of each parameter, but initial investigations into this are included in the section.

4.4.1 Construction of skills dictionary

As outlined in section 3.3, the first step to constructing the skill extraction algorithm was to identify a list of pre-defined skills and link them to descriptions of said skills found on Wikipedia. These skills and their descriptions can be referred to as a ‘skills dictionary’. Initially, it was decided to pull through a lot of information from Wikipedia, but upon verifying the skills dictionary, it was clear that non-skill related information had also been linked. This meant that when applying the algorithm to the job adverts, job descriptions were being assigned non-sensical skill values. For example, a job advert looking to recruit in the justice system, was returning ‘Justice League’ as a skill.

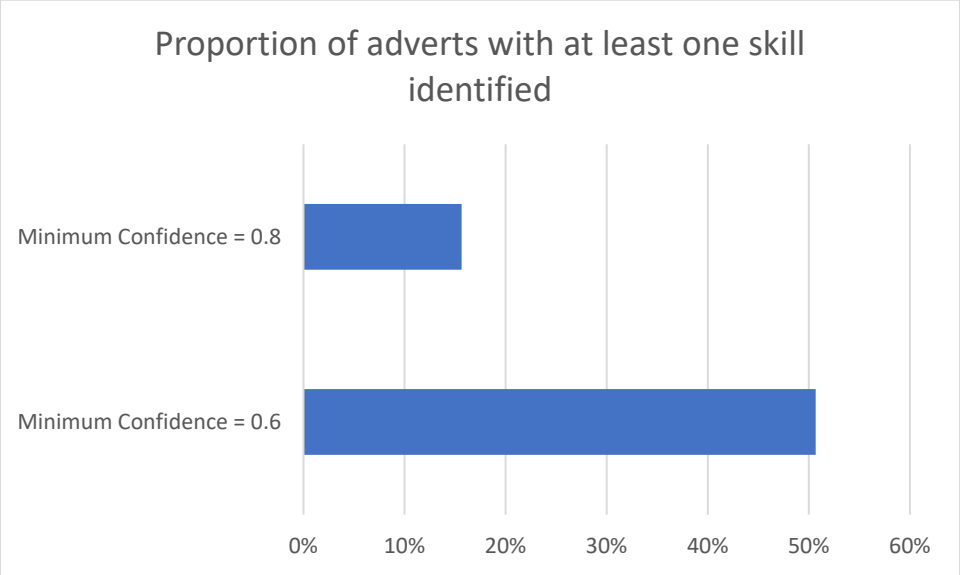
Manual verification was subsequently applied to the skills dictionary which removed these non-sensical references, thus improving the accuracy of the algorithm.

4.4.2 Confidence parameter applied to returned skills

When applying the skill extraction algorithm to online job adverts, a ‘minimum confidence’ parameter can be adjusted. This parameter, which has a value ranging from 0 and 1, represents a cut-off relevance score applied to terms discovered in the job advert, when they are verified against the skill dictionary.

All results discussed in this paper above are based on a minimum confidence level of 0.6. However, to test the sensitivity of the parameter, the algorithm was applied to the same subset of data, but with a minimum confidence value of 0.8. One clear impact of amending the minimum confidence was the increase in the number of adverts for which no skills were identified by the algorithm, highlighted in Figure 9.

Figure 9: Proportion of adverts with at least one skill identified at varying minimum confidence levels



Notes:

1. *Source = Adzuna*

Further investigation into the results of the algorithm based on different minimum confidence parameters, similar to the validation carried out in section 4.3 before a clear decision can be made on which is the optimal parameter.

A more systematic 'search' across the parameters outlined above to the global maximum in the quality metrics phase-space will help identify appropriate combinations of parameters. This will form part of future developments of the algorithm and ONS' next steps in this space.

5. Dissemination

Considering how best to disseminate measures of employer skill demand is very important, as there are a vast number of skills and the end product must find a balance between granularity and interpretability, to best meet user needs. This section highlights some general user requirements expressed through stakeholder engagement, as well as ONS' considerations to date around data presentation.

5.1 User requirements and applications for employer skill demand

As part of this work, ONS has conducted several stakeholder engagement sessions to better understand the applications of these analyses, as well as specific user requirements. This has focused on domestic policy users, particularly across government departments.

There is a clear user need for understanding employer skill demand at a more granular level. For example, understanding employer skill demand over time, across different

geographies would provide greater opportunity for policy intervention at a local level, as it can inform both local upskilling programmes and skills provision investment more generally. Other useful breakdowns cited have been measuring employer skill demand by occupation and by industry, as this can help disentangle the driver for certain skills being certain parts of the economy, or whether new requirements for occupations are occurring as job roles change.

As well as granular breakdowns, there is user interest in specific skill types. For example, understanding if there is growing trend of demand for 'green' skills, and if so, which occupations these skills align to. In the UK, there is a specific government goal to recruit 2 million people into 'green jobs' by 2030 (BEIS, 2021). In a similar vein, emerging skills have been highlighted as a priority in better understanding how the workforce can prepare for the future, and how it may respond to automation. Existing research by NESTA has highlighted how employment is likely to change in future (Bakhshi & Schneider, 2017), and understanding which skills are likely to be in high demand is crucial.

Two final user requirements, which are to be longer-term aspirations for ONS due to their complexity and wider dependencies, are around understanding skill shortages and differing skill levels (e.g., advanced, expert, basic etc.). Furthermore, as identified in the introduction, considering skills mismatch more explicitly using such data is a long-standing ambition of several users.

5.2 Reviewing existing taxonomies

As well as considering specific user requirements, it is important to think more generally about how to appropriately group skills together for dissemination purposes. There is a wealth of existing research available which has looked to classify skills into taxonomies, and so to understand how best ONS can present these data, colleagues in King's College London have been working in partnership with ONS to undertake a review of existing taxonomies (Romanko, forthcoming with permission).

To conduct this review, the following metrics are being considered:

- *Coverage*
  A key factor to consider when analyzing existing taxonomies is coverage, and in particular any gaps in the types of skills that are covered. For example, some taxonomies may have strong coverage of technical skills, but less coverage of softer skills.

- *Timeliness & ongoing management*
  With new skills emerging over time, it is important to consider when the taxonomy was last updated, and if the company owning said taxonomy regularly reviews it to ensure it is still fit for purpose, or if it is semi-automated and data driven.

- *Transparency*
  Additionally, it is important that the company who create the taxonomy are transparent in how it was created/is maintained. This allows users to have confidence in the methodology that has been applied, and that it has been considered with minimal implicit subjectivity.

To date, the taxonomies which have been included in the review are:

- O*Net which provides a list of around 34,000 skills and competencies (abilities, skills, knowledge, tools and technologies) (O*NET, 2021)
- The European Skills/Competences, qualifications and Occupations which comprises 13,485 concepts including knowledge, skills, attitudes and values and language skills and knowledge (ESCO, 2020)
- A list of over a million skill concepts from JANZZon! (Janzz.Technology, 2021)
- The data driven taxonomy produced by NESTA (NESTA, 2018)
- The European Dictionary of Skills and Competences online thesaurus (DISCO, 2021)
- Skills-ML which provides open-source software for applying natural language processing techniques to labour market data (Skills-ML, 2018a)
- A skills and competency framework with a specific focus on digital skills (SFIA, 2018)
- A list of 250 competences obtained from analysis of 800+ early career UK space jobs (Space Skills Alliance, 2020)
- A Canadian skills and competencies taxonomy (Government of Canada, 2021)
- A skills and occupation taxonomy published by It's Your Skills (IYS, 2021)
- The Burning Glass Skills Taxonomy (Burning Glass, 2019)
- An independent skill list created by three developers (Skill Project, 2021)
- The Skills and Recruitment Ontology (Sibarani et al, 2020)
- A skill list developed by Textkernel (Textkernel, 2021)
- A skills taxonomy produced by Skills engine (SkillsEngine, 2021)
- List of over 30,000 skills included in the EMSI's Open Skills Library (EMSI, 2021)
- The repository for all approved National Occupational Standards (NOS, 2021)
- Skills framework providing information on existing and emerging skills required for job roles (Skills future, 2021)
- Skill list specifically for the IT sector (IT Jobs Watch, 2021)

Although the full findings of this review are not yet available, some initial observations have highlighted that there are varying coverage levels for each taxonomy. For example, some lists such as those produced by IT Jobs Watch and SFIA have a specific focus, with IT Jobs Watch limited to skills required for jobs in the IT sector, and SFIA focusing on digital skills only. Other taxonomies such as ESCO cover a wider breadth of skills but underrepresent technical and digital skills. Combining multiple taxonomies may therefore

be a useful exercise to achieve optimal coverage, as well as supplementing with other resources such as the skill term thesaurus designed by DISCO.

The timeliness of each list also varies substantially with some lists such as EMSI being updated very frequently (every two weeks) and JANZZon being updated daily. Whereas others are much more static, such as O*NET which is updated on an annual basis or ESCO which is updated every 1-2 years. This is partially influenced by differences in methodologies, for example EMSI is driven by job advert data, whereas O*NET use surveys of workers.

There are also substantial differences between how each resource may be used. For example, as outlined above, DISCO provides a skill thesaurus which would be useful in enhancing other taxonomies but not as useful in isolation as it was last updated in 2012. Some resources such as EMSI and Dice provide a list of skills that can be used to inform skills extraction algorithms, but do not present the skills in taxonomy form or aggregate them in any way, hence they are less useful for determining an effective presentation of skills. Several of the taxonomies including O*NET, ESCO, NESTA and Janzzon classify the skills by occupation, which is very useful and one of the user requirements identified in section 5.1, whereas others such as ESDC have classified skills according to their own groupings.

As already highlighted, this review is not yet fully complete, but it is clear that the sources/lists considered are hugely influenced by their intended purpose. Like with the use of government tax and other administrative data, the creation/'collection' process must be considered when analysts and researchers try and incorporate such data into analytical pipelines, and although the skills taxonomies described above are generally qualitative in nature, such a similar thought process should be applied.

ONS' interest in skills taxonomies/lists is twofold. Firstly, to ensure the most comprehensive list of skills is used to inform a skill extraction algorithm, and secondly, to ensure skills are grouped in the most efficient way for future publications. When available, the results of this review will be used in conjunction with stakeholder engagement to inform the future developments of the skill extraction algorithm and dissemination of employer skill demand.

5.3 Alternative dissemination ideas

As well as meeting specific user requirements, the nature of the application of the current methodology means that further insights can be drawn from these outputs, which may also be useful to disseminate to users in future. Colleagues in King's College London were able to web scrape a number of job platforms and apply the same methodology as outlined in section 3.3, to explore possible further outputs.

For example, network analysis of the results of the algorithm allows skills co-occurrence to be evaluated (Romanko, forthcoming with permission). This allows insights to be drawn of which skills frequently occur together. Due to the nature of the algorithm, it is also possible to identify which skills are semantically similar using clustering techniques,

independently of the specific demand in specific weeks. This may be particularly useful in grouping sets of skills together into hierarchies, rather than presenting thousands of skills independently from each other. This may also help identify market prices for groups of skills rather than individual skills, given the potential non-linear impact to salary of the presence of various skills together.

6. Conclusion and next steps for ONS

To conclude, this initial application of a skill extraction algorithm has produced some interesting results which provide an indication of what is possible when using online job adverts data to produce an alternative measure of human capital.

There has been success in terms of extracting skills for some adverts, such as those in the IT sector. Even indicative results highlight potential new insights, such as the larger focus of leadership-type skills in London and higher levels of risk-assessment skills in Northern Ireland than the rest of the UK. Additionally, highlighting common skill overlaps shows how some soft skills may be more likely to occur together, whereas digital skills are more distributed across a range of software, and this may impact how people choose to train and learn a broader or narrower range of skills. Additionally, some of the validation and analysis of limitations of the methods starts to open up the understanding of the limitations of such type of data. For example, there may be significant proportions of jobs in certain categories that genuinely do not refer to skills needed and instead capture the bare essentials for people to consider applying (qualifications, location, salary, company and benefits information and a job title) This may limit the amount of insights that such big data tends to promise, at least in terms of representability and granularity of analysis, and may inform future more targeted data collection.

However, before such inferences can be made, it is clear there is much more to do to improve the current methodology to reduce bias and develop more robust estimates. There are clear limitations such as the ineffective identification of soft skills, and further work to do to identify the optimal outputs that should be produced as an end product of employer skill demand. The methods which will be applied to derive market prices for skills also require consideration, as well as general standardisation of the salary field and a method defined to account for missingness.

As a result, ONS has both short-term and long-term aspirations for the general use of online job advert data. This section outlines the next steps in terms of specific skill extraction algorithm improvements as well as wider considerations such as scalability and compatibility with existing infrastructure and future dissemination.

6.1 Specific areas of improvement for existing algorithm

In the immediate term, ONS plan to improve the current skill extraction algorithm to address some of the issues highlighted in this paper. The following sections detail the specific improvements ONS will make in the coming months.

6.1.1 Expanding our initial skill list

As highlighted in section 3.3 and 5.2, ONS have started to consider which is the optimal skill list that should be used to inform our skill extraction algorithm. It's clear from investigations into skills identified that the algorithm is currently missing key skills which should be identified. In particular, ONS intend to expand its coverage of softer skills and associated skills dictionary entries, as well as considering any wider gaps of skills that may not be covered currently. Longer-term, research may consider optimal combination of independent skills taxonomies for a more complete list bringing together both data-driven and expert-led approaches.

6.1.2 Removing noise from job adverts in pre-processing stage

Another key improvement that could be applied to general extraction of skills from online job advert data is to expand our current data pre-processing steps to identify and remove additional noise in the data. For example, stripping out excess information on the company, which is advertising the role, may lead to a reduction in false positive skill identification. Specifically, it would be useful to identify for each job advert the specific paragraph which relates to skill requirements using existing data cleansing algorithms where possible (Skills-ML, 2018b). This would allow the skill extraction algorithm to then be applied to the skills-related paragraph in isolation.

ONS plan to fully consider the best techniques to apply to the data ahead of the skill extraction phase to address these issues.

6.1.3 Expanding the scope of requirements identified by the algorithm

As highlighted in section 4.1, in many job adverts there are requirements which may not explicitly be identified as 'skills', although this depends on the definition used, but are still a requirement for the applicant. A key example is where job adverts require a certain number of years' experience in a specific field. Additionally, job adverts often list what may be referred to as 'characteristics' as requirements for the job such as being 'friendly' or 'having a positive attitude'.

In future, ONS would look to expand the algorithm to capture these kinds of requirements, to consider if they may be translated into skills requirements from more complete job adverts, and to expand the definition of what is measured, beyond just 'skills'. In the same vein, critical qualifications or certifications that are requirements of the job will be investigated to be extracted, particularly as there is an expectation that they would have an impact on the proposed salary and hence on the market price of the remaining skills-focused requirements.

6.1.4 Improving the skill extraction algorithm itself

As highlighted in section 3.2, the currently applied methodology is one of many techniques that could be applied to these data, and hence some improvements that are made to the algorithm may relate to expanding the approach or considering alternative methods entirely.

ONS plan to explore alternative methods such as topic modeling, or even non-natural language processing techniques such as generative adversarial models to understand the differential results depending on which technique is applied. Additionally, the current algorithm could be expanded to take more of a data-driven approach by introducing the use of noun phrases extracted from job adverts to train a neural network instead of pre-defined lists as is currently used.

An additional key part of the work to improve upon the algorithm is to further explore the different parameters outlined in section 4.3 in order to inform on which are the optimal combinations to achieve the highest quality results.

6.2 Scaling up processing

Another key, short-term consideration for the future is how ONS can scale up the current processes so that skills can be extracted from job adverts in the most optimal way.

ONS currently has access to weekly extracts of data from Adzuna, representing a high volume of data when considering the time-series is available from February 2018 to August 2021 and on-going. In order to optimize the way that the data is processed, ONS are working within a distributed computing system. However, currently the algorithm utilizes Python packages not designed to take advantage of said infrastructure. At present, the length of time taken to apply the skills extraction algorithm to an extract of around 1 million job adverts is 6 hours. ONS therefore plans to adapt the algorithm to be compatible with PySpark technologies which will hopefully improve run-time and efficiency.

6.3 Addressing general bias in the data

ONS acknowledge that online job adverts in general have some bias in the coverage in that not all job adverts will be posted online. Additionally, this may impact some sectors more than others. As part of future development work, ONS plans to consider representability of the data, and if possible develop methods to benchmark to other sources, with the aim to remove some of the bias and give a more accurate picture of skills demand. This will also allow for future aggregate comparisons of skill demand with skill supply, and aid skill mismatch analyses.

6.4 Defining methods for extracting market price of skills

As outlined in section 1 of this paper, longer-term, as well as producing measures of employer skill demand ONS aspire to generate a market price for the extracted skills, or groups of skills which can be applied to ONS measures of human capital stock. The focus of this paper has been on the methods applied for extracting skills, but longer-term ONS will consider appropriate statistical techniques for analysing the relationships between skills and salaries, to arrive at said market price values.

6.5 Publication and dissemination

As highlighted in section 5.1, there are many user requirements in terms of ONS estimates of employer skill demand in terms of breakdowns by location/industry as well as more specific interests into areas such as ‘green skills’.

ONS intend to make outputs from these kinds of data sources publicly available in future, but this will depend on future data sharing agreements with our commercial providers. These may range from regular analysis of trends and aggregated series over time of shares of skills appearing in adverts (or in certain occupations/industries benchmarked to be representative of the economy) to equivalent market prices of specific skills over time. They may also span more detailed statistical analyses, such as one-off pieces on the relationships between certain skills and other factors such as company output, productivity or export intensity, variation of skills per occupations, common skills co-occurrences or larger skills clusters, or even skills similarities through network analysis, and other emerging policy-driven analysis. However, a lot of these outputs are yet to be defined and where ONS can, it will consider dissemination to meet a wide range of policy and research user needs.

7. Acknowledgements

ONS would like to thank all those who have contributed to this work to date, including:

Special thanks to:

- Martin Wood and Ryan Schofield, colleagues in ONS who contributed to early iterations of the skill extraction algorithm.
- Colleagues in NESTA for their ongoing collaboration in sharing of best practice when working with online job advert data and their engagement in Economic Statistics Centre of Excellence (ESCoE) projects
- Fred Navruzov for valuable insight into machine learning techniques used in the earlier development of the skill extraction model, as well as programming inputs which were partially reused in the current model.

8. References

- Adzuna. (2021). *Adzuna about us*. [online] Available at: https://www.adzuna.co.uk/about-us.html [Accessed 29 Jul. 2021]
- Alvarez, J., 2017. *A review of word embedding and document similarity algorithms applied to academic text*. Undergraduate. University of Freiburg.
- Arthur, R., 2021. Studying the UK job market during the COVID-19 crisis with online job ads. *PLOS ONE*, 16(5), p.e0251431.
- Bakhshi, H., Downing, J., Osborne, M. and Schneider, P. (2017). *The Future of Skills: Employment in 2030*. London: Pearson and Nesta
- Bank of England, (2018). *Using online job vacancies to understand the UK labour market from the bottom-up*. [online] Available at: https://www.bankofengland.co.uk/-/media/boe/files/working-paper/2018/using-online-job-vacancies-to-understand-the-uk-labour-market-from-the-bottom-up.pdf [Accessed 29 Jul. 2021]
- BEIS. (2021). *Green jobs taskforce report*. [online] Available at: https://www.gov.uk/government/publications/green-jobs-taskforce-report [Accessed 05 Aug.2021]

- Burning Glass. (2019). *Mapping the genome of jobs: The Burning Glass Skills Taxonomy*. [online] Available at: https://www.burning-glass.com/research-project/skills-taxonomy/ [Accessed 03 Aug. 2021]
- Dice. (2021). *Connecting futures now*. [online] Available at: https://dhigroupinc.com/home/default.aspx [Accessed 29 Jul. 2021]
- DISCO. (2021). *European Dictionary of Skills and Competences*. [online] Available at: http://disco-tools.eu/disco2_portal/ [Accessed 03 Aug. 2021]
- Djumalieva, J. and Sleeman, C. (2018). *Linking skills to occupations: using big data to build a new occupational taxonomy for the UK*. [online] Available at: https://www.nesta.org.uk/blog/linking-skills-to-occupations-using-big-data-to-build-a-new-occupational-taxonomy-for-the-uk/ [Accessed 29 Jul. 2021]
- EMSI. (2021). *EMSI Skills API*. [online] Available at: https://api.emsidata.com/apis/skills [Accessed 03 Aug. 2021]
- ESCO. (2020). *European Skills/Competences, qualifications and Occupations*. [online] Available at: https://ec.europa.eu/esco/portal/skill [Accessed 03 Aug. 2021]
- Fasttext. (2020). *Word representations*. [online] Available at: https://fasttext.cc/docs/en/unsupervised-tutorial.html [Accessed 05 Aug. 2021]
- Gilyadov, J. (2021) *Word2Vec Explained*. [online] Available at: https://israelg99.github.io/2017-03-23-Word2Vec-Explained/ [Accessed 29 Jul.]
- Government of Canada. (2021). *Taxonomy*. [online] Available at: https://noc.esdc.gc.ca/SkillsTaxonomy/TheTaxonomy?GoCTemplateCulture=en-CA [Accessed 03 Aug. 2021]
- Hammond, P. (2018). *Spring Statement 2018: Philip Hammond's speech*. [online] Available at: https://www.gov.uk/government/speeches/spring-statement-2018-philip-hammonds-speech [Accessed 02 Aug. 2021]
- Harper, R., 2012. The collection and analysis of job advertisements: A review of research methodology. *Library and Information Research*, 36(112), pp.29-54.
- HM Treasury. (2021). *Build Back Better: our plan for growth*. [online] Available at: https://www.gov.uk/government/publications/build-back-better-our-plan-for-growth/build-back-better-our-plan-for-growth-html#skills [Accessed 29 Jul. 2021]
- Institute for Fiscal Studies. (2020). *Human capital*. [online] Available at: https://ifs.org.uk/research/29 [Accessed 29 Jul. 2021]
- IT Jobs Watch. (2021). ISO 9001 Jobs. [online] Available at: https://www.itjobswatch.co.uk/jobs/uk/iso9001.do#Skill-Set-Cloud-Services [Accessed 03 Aug. 2021]
- IYS. (2021). *What is the IYS Skills and Occupation Taxonomy*. [online] Available at: https://www.itsyourskills.com/iys-skills-taxonomy/ [Accessed 03 Aug.2021]
- Janzz.Technology. (2021). *JANZZon! The unique, multilingual and most comprehensive ontology. Worldwide*. [online] Available at: https://janzz.technology/janzz-on/ [Accessed 03 Aug. 2021]
- Kautz, T. et al. (2014). *Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success*. [online] Available at: https://www.nber.org/system/files/working_papers/w20749/w20749.pdf [Accessed 05 Aug. 2021]

- Liddy, E.D. (2001). *Natural Language Processing*. [online] Available at: https://surface.syr.edu/cgi/viewcontent.cgi?article=1019&context=cnlp [Accessed 02 Aug. 2021]
- NESTA. (2018). *The first publicly available data-driven skills taxonomy for the UK*. [online] Available at: https://www.escoe.ac.uk/the-first-publicly-available-data-driven-skills-taxonomy-for-the-uk/ [Accessed 03 Aug. 2021]
- NLTK, (2021). *NLTK 3.6.2 documentation*. [online] Available at: http://www.nltk.org/ [Accessed 29 Jul. 2021]
- NOS. (2021). *Repository for all approved National Occupational Standards*. [online] Available at: https://www.ukstandards.org.uk/ [Accessed 03 Aug. 2021]
- O*NET. (2021). *O*NET Online*. [online] Available at: https://www.onetonline.org/find/descriptor/browse/Skills/ [Accessed 03 Aug. 2021]
- OECD. (2017). *Boosting skills would drive UK growth and productivity*. [online] Available at: https://www.oecd.org/education/boosting-skills-would-drive-uk-growth-and-productivity.htm [Accessed 29 Jul. 2021].
- OECD. (2021). *Navigating skill demands in turbulent times*. [online] Available at: https://www.oecd-ilibrary.org/sites/01a857c7-en/index.html?itemId=/content/component/01a857c7-en#chapter-d1e27218 [Accessed 29 Jul. 2021]
- ONS, (2016). *UK SIC 2007* [online] Available at: https://www.ons.gov.uk/methodology/classificationsandstandards/ukstandardindustrialclassificationofeconomicactivities/uksic2007 [Accessed 05 Aug. 2021]
- ONS. (2012). *Vacancy Survey QMI*. [online] Available at: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/methodologies/vacancysurveyqmi [Accessed 02 Aug. 2021]
- ONS. (2018). *Human capital workplan: 2018*. [online] Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/articles/humancapitalworkplan/2018 [Accessed 29 Jul. 2021]
- ONS. (2019). *Human capital estimates in the UK: 2004 to 2018*. [online] Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/articles/humancapitalestimates/2004to2018#methodology-developments [Accessed 02 Aug. 2021]
- ONS. (2020). *Coronavirus and the latest indicators for the UK economy and society: 28 May 2020*. [online] Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/coronavirustheukeconomyandsocietyfasterindicators/28may2020 [Accessed 05 Aug. 2021]
- ONS. (2021a). *Using Adzuna data to derive an indicator of weekly vacancies: Experimental Statistics*. [online] Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/methodologies/usingadzunadatatoderiveanindicatorofweeklyvacanciesexperimentalstatistics [Accessed 29 Jul. 2021]
- ONS. (2021b). *Economic activity and social change in the UK, real-time indicators*: 29 July 2021. [online] Available at: https://www.ons.gov.uk/economy/economicoutputandproductivity/output/bulletins/economicactivityandsocialchangeintheukrealtimeindicators/29july2021 [Accessed 29 Jul. 2021]

- ONS. (2021c). *Coronavirus and the latest indicators for the UK economy and society: 29 April 2021*. [online] Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/coronavirustheukeconomyandsocietyfasterindicators/29april2021#online-job-adverts [Accessed 02 Aug. 2021]
- ONS. (2021d). *Vacancies and jobs in the UK: July 2021*. [online] Available at: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/bulletins/jobsandvacanciesintheuk/july2021 [Accessed 02 Aug. 2021]
- ONS. (2021e). *Business and individual attitudes towards the future of homeworking, UK: April to May 2021*. [online] Available at: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/articles/businessandindividualattitudestowardsthefutureofhomeworkinguk/latest [Accessed 05 Aug. 2021]
- Pascual, F. (2019). *Topic modeling: An introduction*. [online] Available at: https://monkeylearn.com/blog/introduction-to-topic-modeling/ [Accessed 05 Aug. 2021]
- Prabhakaran, S. (2018). *Lemmatization Approaches with Examples in Python*. [online] Available at: https://www.machinelearningplus.com/nlp/lemmatization-examples-python/ [Accessed 02 Aug.2021]
- Pure Speech Technology. (2021). *A guide: Text Analysis, Text Analytics & Text Mining.* [online] Available at: https://www.purespeechtechnology.com/text-analysis-text-analytics-text-mining/#TextAnalysisTextMiningTextAnalyticsDifference [Accessed 02 Aug. 2021]
- Riva, M. (2021) *Word embeddings: CBOW vs Skip-Gram* [online] Available at: https://www.baeldung.com/cs/word-embeddings-cbow-vs-skip-gram [Accessed 05 Aug. 2021]
- Romanko, O ; O'Mahony, M. (Forthcoming with permission). *The use of online job sites for measuring skills and labour market trends: A review*. In: Economic Statistics Centre of Excellence (ESCoE) Discussion Papers.
- Nikolenko, S., Koltcov, S. and Koltsova, O., 2016. Topic modelling for qualitative studies. *Journal of Information Science*, 43(1), pp.88-102.
- SFIA. (2018). *The global skills and competency framework for a digital world*. [online] Available at: https://sfia-online.org/en/about-sfia/browsing-sfia [Accessed 03 Aug. 2021]
- Sharma, N. (2019). *Job skills extraction with LSTM and Word Embeddings.* [online] Available at: https://confusedcoders.com/wp-content/uploads/2019/09/Job-Skills-extraction-with-LSTM-and-Word-Embeddings-Nikita-Sharma.pdf [Accessed 29 Jul. 2021]
- Sibarani, E. et al. (2020). *Skills and Recruitment Ontology*. Available at: https://elisasibarani.github.io/SARO/ [Accessed 03 Aug. 2021]
- Skill Project. (2021). *We are mapping every human skill out there. And we need your help*. [online] Available at: http://www.skill-project.org/ [Accessed 03 Aug. 2021]
- Skills Engine. (2021). *Standards-Based Skills Taxonomy*. [online] Available at: https://www.skillsengine.com/skills-taxonomy [Accessed 03 Aug. 2021]
- Skills future. (2021). *Skills Framework.* [online] Available at: https://www.skillsfuture.gov.sg/skills-framework/ [Accessed 03 Aug. 2021]
- Skills-ML. (2018a). *Skills-ml documentation*. [online] Available at: https://workforce-data-initiative.github.io/skills-ml/skills_ml_tour/ [Accessed 05 Aug. 2021]

- Skills-ML. (2018b). *Skills-ML: An open source Python library for developing and analyzing skills and competencies from unstructured text*. [online] Available at: http://dataatwork.org/skills-ml/SkillsMLWhitepaper.pdf [Accessed 05 Aug.2021]
- Space Skills Alliance. (2018). *Towards a space competencies taxonomy*. [online] Available at: https://spaceskills.org/towards-a-space-competencies-taxonomy [Accessed 03 Aug. 2021]
- Tang, J., Meng, Z., Nguyen, X., Mei, O., Zhang, M. (2014). *Understanding the limiting factors of topic modeling via posterior contraction analysis*. In proceedings of the 31st International Conference on Machine Learning – Volume 32 (ICML'14). JMLR.org, I-190-I-198.
- Textkernel. (2021). *Skills: Talent Pipeline and Database Enrichment*. [online] Available at: https://www.textkernel.com/skills/ [Accessed 03 Aug. 2021]
- Towards data science, (2020). NLP in Python-Data cleaning. [online] Available at: https://towardsdatascience.com/nlp-in-python-data-cleaning-6313a404a470 [Accessed 29 Jul. 2021]
- Le, Van-Duyet & Vo, Minh-Quan & Quang-An, Dang. (2017). *Skill2vec: Machine Learning Approach for Determining the Relevant Skills from Job Description*. [online] Available at: https://arxiv.org/pdf/1707.09751.pdf [Accessed 04 Aug. 2021]
- Zhao, M. et al. (2015). *SKILL: A System for Skill Identification and Normalization* [online] Available at: https://www.aaai.org/ocs/index.php/IAAI/IAAI15/paper/viewFile/9363/9907 [Accessed 04 Aug. 2021]