Detecting Trending Skills in Information Technology Industry Using Job Ads

Fu-Yin Cherng, Khalil Mrini, Pierre Dillenbourg, Robert West
Ecole Polytechnique Fédérale de Lausanne, Switzerland
{fu-yin.cherng,khalil.mrini,pierre.dillenbourg,robert.west}@epfl.ch

1 ABSTRACT

Due to the rapid development in Information Technology (IT), knowing what skills are trending in the job market can not only help job seekers equip themselves with proper competencies but also help training and education association improve their training curricula. In this paper, we collected IT job advertisements (job ads) from an online job board and applied methods from natural language processing and trend detection on job titles and job descriptions to identify trending skills. We found that "Cloud Computing" is a trending skill for IT job from the results of all analysis. Also, there are more and more job ads that listed the requirement of business skills. In addition to investigating trending skills in the context of job ads, we also inserted these skills into Google Trends to see whether these skills are also trendy in the context of the public's interests. The comparison between Google Trends and job ads shows that a trend in the job market is not necessarily considered as a trend in the public's interests. We believe that our pipeline of how to analyze job ads can help the subsequent studies to better process the job ads and detect trending jobs and skills. The mismatch between Google Trends and job ads also shows the information gap between what people think the companies require and what the companies actually want.

2 INTRODUCTION

As the technology evolves, the demand on the job market changes rapidly. For instance, the jobs like data scientist and social media manager have emerged in these years, because of the popularity of Big Data and social networking websites. Aged society and long working hours also caused the increasing requests for the care provider. Understanding the trends in the job market is essential because it influences student's choice of discipline [3] and how the educational associations design their curricula.

What skills are trending in the job market is a question that researchers look into for a decade [9, 19, 25]. However, the prior research found that some results of these studies conflict with each others [15]. Moreover, the new skills and tools update fast in recent years, and it is indispensable and useful to keep updating the list of trending skills in IT job market.

The goal of this paper is to find emerging jobs by empirically analyzing job advertisements (job ads) in the online job board. Particularly, we investigated the changes of job ads in the category of Information Technology (IT) jobs.

We collected the two datasets of IT job ads which posted in Adzuna.com [1]. The first dataset is the IT job ads posted in 2013 and March-May 2017. The second dataset is the IT job ads posted in 2016 and 2017.

For the analysis of job ads, we first conducted descriptive analysis on job titles and job descriptions of the first dataset to investigate how the frequency of each job title and skill changes between 2013 and 2017. After conducting descriptive analysis, we used the methods of trend detection on the job descriptions of IT job ads of the second dataset to see what words and skills in the descriptions are identified as trending words by trend detection.

Finally, we compared our results with data of Google Trends[2]. We want to see whether the trends we found in the job market can also be found in the data represents the interests of the general audience, because we believe that the general audience can stand for the interests and opinions of job seekers.

The present research intends to make the following contributions. (1 We identified the emerging job titles and descriptive terms in IT job market through data-driven analysis. (2 We proposed an automatic method for trend detection of job advertisements. (3 We presented the comparison between the changes in job ads and Google Trends, which allows us to understand how the interests of the job market and general population correlated with each other.

3 RELATED WORKS

Since the 1990s, several studies have investigated what is training needs and required skills in IT-related job market [9, 19, 25]. Because of the dynamic nature of IT job market, these studies aim to understand and track the evolution of required competencies in IT job market. Moreover, they want to bridge the skill gap between education and job market to help job seekers and prospective students prepare themselves for this rapidly changing environment [18, 28].

There are several research methods and datasets using to investigate this problem. For the studies used survey-based approach [3, 5, 20], they designed questionnaires to students, faculties and employees from IT-related departments to collect their opinions about what competency is important. Furthermore, the researchers compared the survey results from students with the results from employees. They found the mismatch of skills' perceived importance between students and employees. For example, Richards, D. et al. [20] found that soft skills are perceived to be the most important skill with small knowledge gap between students and employees. However, technical skills are considered to be the less important skills but with largest knowledge gaps.

In addition to the results from the subjective aspect, there are the other set of studies using empirical methods to find out the required skill sets through analyzing job ads [2, 9, 14]. By collecting numerous job ads from online job boards, the results of these studies are more generalizable than survey-based studies, which can only recruit a limited number of participants.

[1] https://www.adzuna.co.uk

[2] https://trends.google.com/trends/

Debortoli, S. et al., applied LSA on job ads to extract required competencies for business intelligence and big data professionals, and by comparing the similarity and difference of their competency requirements, they clarified the confusion between these two professionals [9]. Moreover, there are studies applying longitude analysis on job ads to understand the evolution of required skills in IT job market [16, 18]. Prabhakar B. et al. collected job ads from Monster.com for three years (2002-2005) and examined the changes in required skills. They found that in the job ads, the frequency of web programming, Linux and SQL server skills increased from 2002 to 2005 [18].

Although there are series of studies aiming to address what is the in-demand skills for IT professionals, Litecky C. et al. mentioned that their results were sometimes conflicting with each other and previous summaries [15]. Therefore, more studies and analysis on larger datasets are needed for better understanding the trends of required competencies in IT job market.

4 RESEARCH GOAL AND QUESTIONS

The goal of this work is to find out what kind of jobs and competencies are demanded in IT industry by looking into the job titles and job descriptions in online job ads. By comparing the job ads collected in different periods, we listed the top job titles and words that their frequencies increased and decreased in the later periods. Therefore, the first research question we want to ask is: **RQ1. What are the emerging job titles and skills?**

In addition to investigating this question from the perspective of the job market, we also want to see whether the same trends appear in the data that can reflect the interests of the public. If the trends from job market match the trends from the public's interest, we can say that the trends are indeed in-demand in IT job market and people are aware of that. In contrast, if there is a mismatch between them, by observing the mismatches, we can reveal the lack of understanding between job market and the general audience. Moreover, making the public become more sensitive to what is trending in the job market can facilitate the process of recruitment and job search. Therefore, the second question we want to ask is: **RQ2. If we observed the increase and decrease of a certain job in job ads, can we find the same trend in the data that represents the public interests?**

5 DESCRIPTIVE ANALYSIS ON JOB ADS

We first performed frequency-based analysis on job titles and job descriptions in job ads.

6 DATA COLLECTION

We collected the dataset of IT job ads posted on Adzuna.com. The job ads of 2013 obtained from Kaggle competition, Job Salary Prediction [3] which was hold by Adzuna, and job ads of 2017 were collected via JobApis [4] which is an open-source API for accessing data from multiple job boards. To avoid the confounding variable of geography, we only used the job ads from the UK. The job ads posted in Adzuna.com are labeled in 29 categories. We collected all the job ads that are categorized in Information Technology (IT). Although we only obtained the job ads from two separate years, 2013 and 2017, we believe that the interval of 4 years is sufficient for the job market to have observable changes. The information of two datasets is listed in Table 1.

Table 1: Information of data set of IT job ads in 2013 ans 2017.

Year	# Job	# Source	Location	Collecting Period
2013	65,795	Kaggle	The UK	all year
2017	13,757	Adzuna	The UK	March - May

6.1 Job Title Analysis and Normalization

First, we want to analyze the change of the frequency of different job titles between 2013 and 2017. Because a job title could represent a group of skills, we considered that observing the changes of job titles is the first step to understand the change of required skills.

6.1.1 Job Title Cleaning. Because each job title was assigned by different employers, the titles for the same job were often different. By browsing the job titles in IT job ads, we concluded some common variations for job titles of the same job.

The most common variation is the job title followed by locations. For example, the job "C# Developer" could be "C# Developer Birmingham" in other job ads. The second is the job title followed by additional requirements. For instance, the job "BI Analyst" could be "BI Analyst - 12 Month Fixed Term" or "BI Analyst - Retail Senior" in other job ads. The third is the change of order of words in a job title. For example, the job "C# Developer" could be "Developer C#" in other ads.

To deal with these variations and merge them, we removed the numbers, common words (e.g., stop words in English and words like "job" and "hour") and locations in the job titles. We also filtered out some punctuation, but kept the punctuation "" and "#." Lemmatization was conducted to normalize plural nouns and verb tenses.

6.1.2 Dictionaries for Job Title. Then, we built the dictionary from the corpus of job titles. In the dictionary, we only kept the words with more than 100 counts. After building the dictionary, we then removed the words that are not in the dictionary from the job title. In this way, we were able to remove the words that are rare (e.g., special requests and typo), and kept the words that are significant. The number of words in the dictionary for IT jobs is 235. We also used the dictionary to build word vectors for each job title. In the word vector, each dimension is one word in the dictionary, and the value of vector stands for the number of occurrences of this word in given job title. For example, "C# Developer" and "C# Developer" would have the same word vector, because "C" and "Developer" both appear once in both job titles. Therefore, the variation of job title caused by different word order was solved by using word vector to represent the job title.

[3] https://www.kaggle.com/c/job-salary-prediction/data
[4] https://www.jobapis.com/open-source/

Table 2: The data statistics of job description analysis.

Year	# job	# unigram	# bigram	# trigram
IT jobs - 2013	65,795	9,680	82,454	2,438,477
IT jobs - 2017	13,757	2,825	5,176	274,738

6.2 Result and Discussion of Job Title Analysis on IT Jobs

After normalizing the job titles of IT Jobs, we computed the percentage of each job title by dividing its count by the total number of the job positions in the given year. The changing percentage of each job title is computed by subtracting percentage of 2017 by percentage of 2013. Figure 1 shows the top 10 increasing and decreasing job titles by sorting the changing percentage of job titles. From the left figure in Figure 1, we can see that there are a group of jobs about front-end developer like "trainee web developer" and "javascript developer." We can also map some known trends in IT industry to our result. For example, the rise of Big Data and Data Science are reflected in the increasing job title "data analyst" and "data." The job "devops engineer" could be mapped to the growth of DevOps (DevOps, a engineering practice to unify software development (Dev) and software operation (Ops)) in recent years.

The right figure shows the top 10 decreasing job titles in IT job ads. The percentage of "engineer" decreased the most and followed by "administrator." The decrease of these job titles could be caused by the actual decreasing demand of the job, like "c aspnet developer", or the change of word usage for the same job, like using "developer" instead of "engineer." However, the value of decreasing percentage is pretty small. Therefore, whether the decrease is a false alarm or reflection of the actual drop of requirements needs more investigations.

6.3 Job Description of IT Jobs Analysis

The job description of job ads listed the information of required skills and the expectation for the position. Therefore, we believe that we could extract the trendy words and skills by analyzing the content of job description.

For cleaning the texts in job descriptions, we first removed all the stop words and the undesired punctuation. The lemmatization was then performed to normalized verb tenses and plural nouns. Second, we used n-grams model to compute the unigram, bigram, and trigram of descriptions of the job ads of 2013 and 2017. Third, after counting the frequencies of the unigram, bigram, and trigram in both years' job ads, we kept the terms with more than ten counts and divided them by the number of jobs in 2013 and 2017, respectively. Therefore, we can obtain the percentage of how many jobs used the n-grams (n = 1-3) in their descriptions. Finally, we subtracted the term percentage of 2017 jobs by the term percentage of 2013 jobs to get the changing percentage of each n-gram, and we sorted them by the value of the changing percentage. The statistics of the cleaned job description are shown in Table 2.

6.4 Result and Discussion of Job Description Analysis

Figure 2 shows top 10 increasing and decreasing n-grams in IT jobs by comparing job ads between 2013 and 2017.

From the left figure, we can see that the words like "dynamic", "ax," "nav" and "microsoft dynamics" increased by more than 15%, which means that the job ads mentioned these words increased by 15% by comparing ads of 2013 and 2017. We believe these words referred to the Microsoft software products "Microsoft Dynamic", "Microsoft Dynamic AX" and "Microsoft Dynamics AX." Microsoft Dynamics is a line of software products. The purpose of these products is for enterprise resource planning (ERP) and customer relationship management (CRM). This result indicates that more and more positions in IT market want their employees to have knowledge of business intelligence and skills related to Enterprise Systems (ES). Moreover, this finding matches the result from previous study [10]. From the result of surveying 70 Information managers, they concluded that the importance of ERP skills is expected to increase in the next five years. The importance of business skills also reflects in the word "customer" with 17% of growing percentage. This result might indicate that the abilities related to customer (e.g., customer relationship management) are mentioned in more job ads in 2017.

The right figure shows the top 10 decreasing words in the job descriptions. The word "net" decreased by 68%, which means there are 68% of jobs in 2017 no more required "net" in their job descriptions. By reading the raw job description, we confirmed that "net" refer to .NET, which is a software framework developed by Microsoft. "asp" and "asp net" also refer to ASP.NET which is one product from the product line of Microsoft .NET. This result could indicate that the usage of .NET framework decreased and the use of other alternative products increased. We would like to validate this interpretation by surveying software developers or measure the usage of .NET and alternative products.

In addition to analyzing the descriptive terms in job ads, we are also interested in knowing what are the emerging skills in IT job market. What kind of skills become trendy in 2017 by comparing with 2013. Because we considered that accurately extracting skill terms from the job description requires the human intervention, we decided to use the predefined list of skills to extract the skills and see how these known skills change between 2013 and 2017.

To define the list of the known skills, we collected the name of skills from online reports [5] and prior studies [2, 15]. We created a list of 40 technical skills that are mentioned in these articles. According to the list of skills, we extracted these skill terms from our unigram result and sorted by their difference of percentage, which obtained by subtracting percentage in 2017 by 2013.

The top 10 increasing and decreasing skills among our skill lists are shown in Figure 3. In the top 10 increasing skills, we first want to discuss the words "cloud", "AWS", and "azure." "AWS" stands for Amazon Web Service, which provides personal cloud computing platform. Azure is also a could computing platform developed by Microsoft. The increasing popularity of the two platforms and "could" match the trend of cloud computing [29].

[5] http://www.cio.com/article/3060812/it-skills-training/10-fastest-growing-tech-skills.htmlslide2; https://stackoverflow.blog/2017/03/09/developer-hiring-trends-2017/

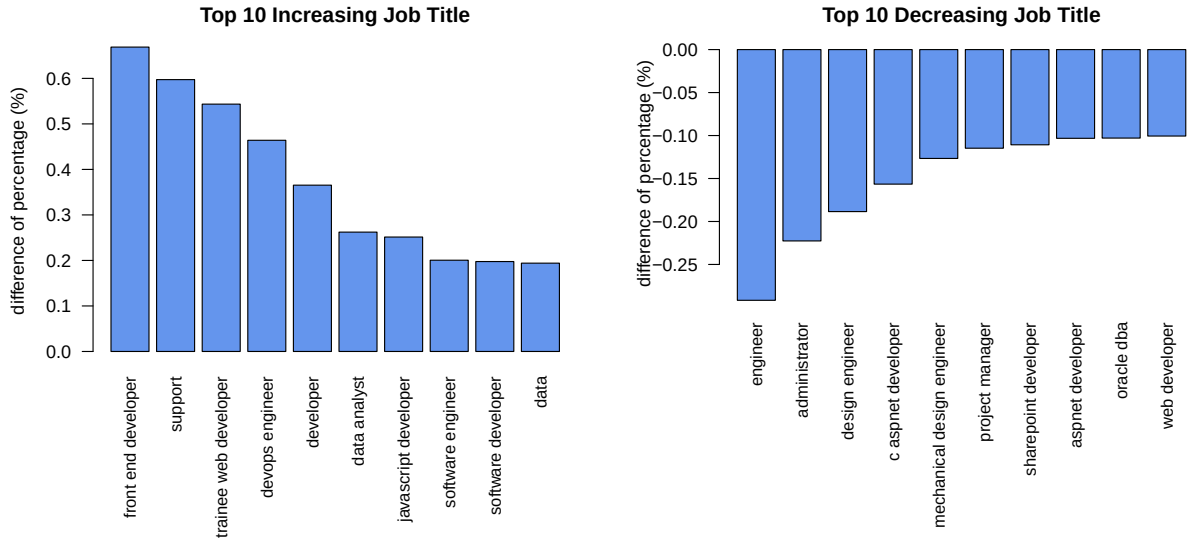

Figure 1: Top 10 increasing and decreasing normalized job titles in IT job ads.

The skill “Devops" stands for DevOps. The increase of DevOps in job description analysis matches our result in job title analysis, which we also found the increasing percentage of “DevOps engineer." We considered that this result also supports the emerging importance of business skills in IT job market that we discussed above. The increase of skill “salesforce", which is a cloud computing company focus on customer relationship management (CRM) product and commercial application, also support the trend of business skills.

For the other increasing skills, we can see that the percentage of Microsoft increased by 15%. Although the decreasing of Microsoft .NET, Microsoft still increased a lot because of lots of other products and services. The skills “python" and “Go" are two programming language that became more and more popular among developers because of their simplicity and extensive support libraries.

For the decreasing skills, .Net, SQL, and C also show up in the list followed by HTML, PHP and Java. The decrease of these programming languages might be caused by the release of competitive programming languages. For example, programmers tend to choose Python instead of C for software development. However, the current result could only show what skills are increasing and decreasing by comparing 2013 with 2017. If we want to evaluate the reasons for these changes, more analysis and studies need to be performed. For example, by fixing the job title, we could observe how the required skills for this job change over the years.

6.5 New skills in IT job market in 2017

In addition to showing the increasing and decreasing of n-grams that both exist in 2013 and 2017 job ads, we are also interested in understanding those terms that newly appeared in 2017.

There are 105 new words that showed up in 2017 job ads. We manually identified several new IT skills from them. Based on the purpose and functionality of the skills, we classified them into five groups. Each group and the skills it contained are listed in Table 3.

Table 3: The skills that newly appeared in job ads of 2017. SGDB stands for database management system.

Group	Skills	Total %
Data	Kafka, SGDB	1.7%
Deployment Management	Ansible, Kubernetes	3.0%
Management Tool for Software Development	Docker, gulp.js, Gitlab, Xamarin	4.1%
Web-applications	Laravel, Reactjs, Webpack, Redux	2.2%
Others	IOT, Blockchain, Microservices	3.4%

There are some possible reasons that why these skills showed up in 2017 and didn’t exist in 2013. The first and the most intuitive reason is that these skills were released after 2013, e.g., kubernetes, gulp and Redux. The second reason we considered is that the functionality of these skills became more and more valuable as the years go by. For example, due to the popularity of telecommuting and online service, the tools like Gitlab and Reactjs are almost indispensable by developers and employees in IT industry. Therefore, there are several skills belong to the group “Management Tool for Software Development" and “Web-application." The skills in “Data" also followed the trend of Big Data and Data Mining.

Although we discussed the potential reasons of the appearance of these skills, we still need further studies and more collection of job ads to verify these statements and obtain more insights.

7 ADVANCED ANALYSIS ON JOB ADS

The studies about detecting trending topics and keywords on Twitter [8] and discussion forum [17] offer more advanced methods for trend detection than the methods in the frequency-based analysis. Therefore, in the following sections, we applied these advanced methods on job ads, which was collected in the different period, to detect trendy words in job descriptions.

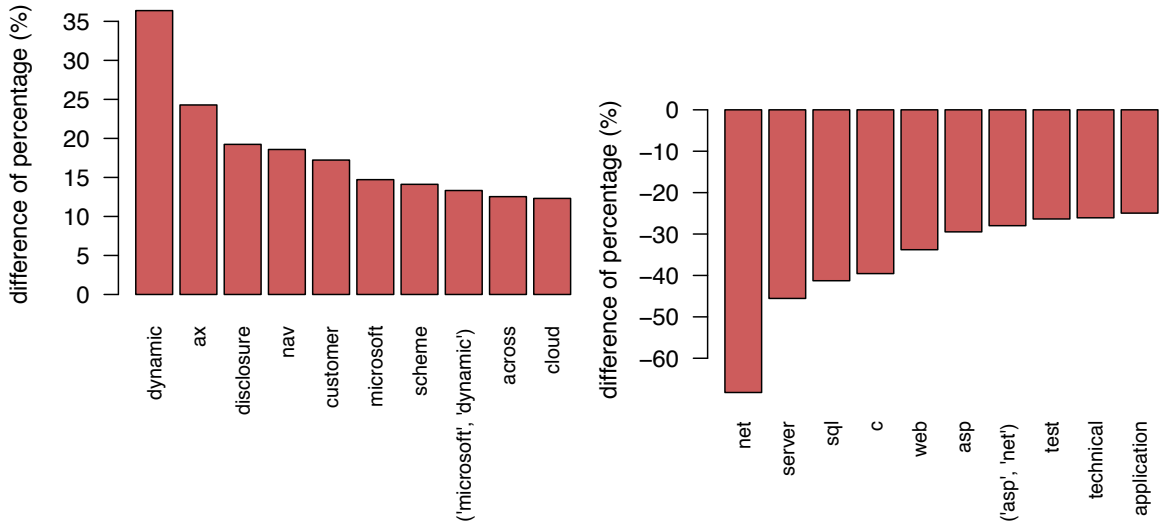

Figure 2: Top 10 increasing and decreasing n-grams in job descriptions of IT job ads.

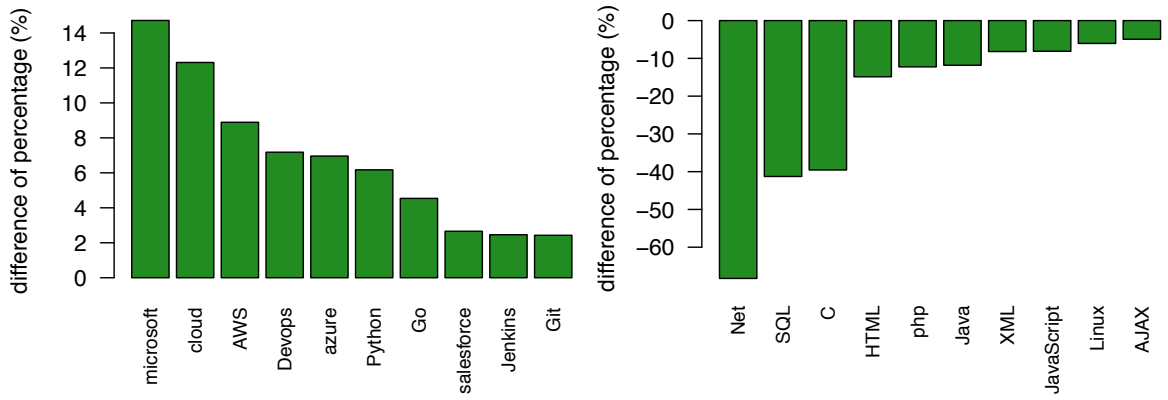

Figure 3: Top 10 increasing and decreasing skills of IT jobs.

7.1 Data Collection

The job ads used in this paper were collected in mid-June 2017 on Adzuna[6], an online search engine for job ads, using their official API. This API offers job ads in 16 countries, in 8 languages, and the jobs are labeled in 29 categories. Among all the data we collected, we selected IT job market in the UK to apply our methods. Based on the date of posting we can trace back, the time periods we selected to compare is the first semester in 2016 and the first semester in 2017. There are totally 163 jobs in 2016 and 74, 123 jobs in 2017.

7.2 Pre-processing of job description

Three steps were used for pre-processing: stop words removal, lemmatization, and cross-domain filtering. Except cross-domain filtering, the rest processes are the same as the previous sections.

Given that we can collect multiple job domains (categories) for the same country, words that appear across domain in a similar frequency are not domain-specific and should be deleted. This procedure is called cross-domain filtering.

For that, words should first be ranked by their domain-wide TF-IDF weight [21] in each of the domains. Li et al. propose in [13] a formula for computing a corpus-wide TF-IDF weight for a news corpus to get keywords, based on the assumption that keywords often appear in a news article. This is not necessarily the case for job ads where skill requirements may be mentioned only a few times. Wartena et al. present in [27] a corpus-wide TF-IDF weight which term frequency is absolute and therefore gets more importance in the overall result.

Descriptions of ads can be of varying length. Therefore, on top of reducing the weight of term frequency in the TF-IDF formula, we want to reduce the weight of a document's length, such that long job ads will not overshadow shorter ones. We adopt the TF-IDF formula in Equations 1, 2 and 3 introduced by Lee and Kim in [11].

[6]Adzuna is accessible on https://www.adzuna.com/

$$TF(w) = \log\left(\frac{1}{|D|}\sum_{d\in D}\frac{n(w,d)}{\max_{w'\in d} n(w',d)}\right) + 1 \quad (1)$$

$$IDF(w) = \log\left(\frac{|D|}{|\{d \in D : w \in d\}|}\right) \quad (2)$$

$$TFIDF(w) = TF(w) \cdot IDF(w) \quad (3)$$

We performed cross-domain filtering [11] by first ranking the words by their TF-IDF weight, and then we computed the standard deviation of their ranks across the domains. We setted a threshold for the standard deviation, here defined as $min(1000, |w \in D|)$ with D being the job ads in the six domains for a given country. If the standard deviation is below that threshold, the word w is a word that does not bear domain-specific information and should therefore be deleted.

7.3 Trend Detection

To detect trends, two collections were created: one for the first half of 2016, and the other for the first half of 2017. We first formed n-grams to collect multi-word phrases and improve the vocabulary, and then we further restricted it to only meaningful, information-rich words by filtering based on named entities and part-of-speech tags.

Lent et al. propose in [12] an adaptation of sequential patterns mining to textual data. Agrawal and Srikant introduced the Apriori algorithm in [1] for mining sequential patterns and later improve it in [23] by introducing the Generalized Sequential Patterns (GSP) algorithm.

We use the GSP algorithm to collect all n-grams with $n > 1$ with a minimum support of 0.2. We apply the algorithm on the union of the two collections of job ads compared so that they can share the same n-grams for better trend detection.

Named entities such as locations and persons should not come up as trends. Indeed, finding that *London* is trending for job ads in the UK is not useful. However, if we find that companies, which are tagged as organizations, are trending, it could be an interesting piece of information.

To perform Named-Entity Recognition (NER) in the job ads in their respective languages, we used Polyglot-NER [4]. We then deleted each word or multi-word expression tagged as location or person.

Verbs and adverbs are examples of parts of speech that will not be meaningful as trends. We, therefore, decided only to keep words that are tagged as nouns or adjectives or which part of speech could not be determined. The latter case would surface for instance for n-grams or newly coined words, that we would like to keep.

7.3.1 Comparison of TF-IDF weights. We first computed the TF-IDF weights according to the formula in Equations 1, 2 and 3 for each word separately on the two collections. That formula normalizes the term frequency by the maximum term frequency in each job ad and by the size of the collection, hereby tackling the imbalance in the number of job ads between 2016 and 2017.

For each word present in both collections, we computed the trend score by subtracting the TF-IDF weight for the first semester of 2016 from its 2017 counterpart. A positive trend score indicates a word that is more trending in the first semester of 2017 than in the same period in 2016.

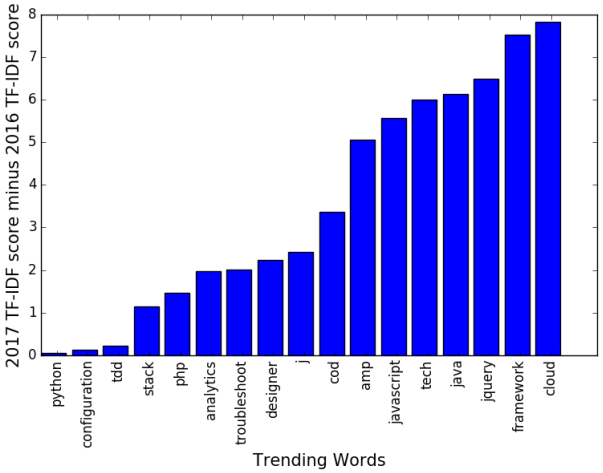

Figure 4: Trending Words in IT jobs in the UK plotted with their 2016-2017 TF-IDF evolution

7.3.2 Clustering of Trending Words. After obtaining the trending words, it is possible to cluster them to get an idea of their similarity.

Wartena and Brussee cluster documents in [26] by using the similarity between their keywords. They expressed that similarity in three ways: the cosine similarity of their document distributions, the Jensen-Shannon divergence of their document distributions and the cosine similarity of their TF-IDF vectors. The latter is widely used in information retrieval and "*has proven to be a robust metric for scoring the similarity between two strings*" [24].

We clustered trending words in a dendrogram with the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [22]. The distance between two trending words is the Euclidean distance between their TF-IDF vectors of length $|D|$ where D is the collection of documents. The distance between two clusters C_1 and C_2 of trending words is the average of all the pairwise distances of the words they contain, as shown in Equation 4.

$$d(C_1, C_2) = \frac{1}{|C_1| * |C_2|}\sum_{a\in C_1}\sum_{b\in C_2}\sqrt{\sum_{d=1}^{|D|}(a_d - b_d)^2} \quad (4)$$

7.4 Trend Results and Discussion

Based on the pipeline afore-described, we obtained trending words for the UK's Information Technology job ads.

Out of the 217 common words between the 2016 and 2017 collections, 7.8% (2017) have a positive difference between their 2017 TF-IDF score and their 2016 one and are therefore considered to be trending words. They are shown in Figure 4.

The trending words contain skills such as programming languages (Python, PHP, Java), abbreviations (AMP, which stands for Apache, MySQL and PHP, a commonly used solution stack; TDD, Test-Driven Development) or other concepts in Computer Science (Cloud, analytics, troubleshoot). We also found the trends of Could

and Python in the descriptive analysis by using the data sets of 2013 and 2017. This finding indicates that the importance of these two skills grows steadily from 2013 - 2017.

Figure 5 shows the clustering of the trending words in a dendrogram. This TF-IDF-based clustering captured a few word couples that have similar frequency patterns, like configuration and troubleshoot, or TDD and jQuery.

There is a large part of the words collected that are not trending, although the TF-IDF formula normalized term frequency for each text, and as well for the length of the corpus. It is likely caused by the imbalance in the number of job ads between 2016 and 2017. We believe that if we collected more balanced and comprehensive IT job ads, we could obtain more insights and comparisons between multiple time points.

8 COMPARISONS WITH CHANGES IN GOOGLE TRENDS

The goal of the comparison in this section is to understand if we observed a certain job increased in the job market, whether we can also find the same increase of this job in a more general context. In the other words, we want to know whether the public would also be interested in a certain job or skill that are trending in the job market.

We used Google Trends as the data represents the interests of the public. Google Trends is a public dataset released by Google Inc. that shows how often a word is searched in Google Search Engine. The value of the input word in Google Trends does not give absolute data about search query quantities, but relative data. The value is normalized to 100, and 0 meaning that there was no search query. The Google Trends has been used in various studies and applications, including Google Flu Trends [6] and prediction of automobile sales [7].

In the following sections, we conducted two kinds of comparison between data from Google Trends and job ads. First, we computed each skill's difference between the annual frequency of Google Trends and compared them with the changing percentage of the skills in job ads. Second, we used the formula from [8] to compute the trending energy of the skills by using Google Trends data. The skills are from the increasing skills we obtained in the result of the advanced analysis. Through these comparisons, we want to know whether these trending skills would also be trending in Google Trends.

8.1 Google Trends: Difference of Percentage

The skills used in this comparison are selected from the list of the known skills in the descriptive analysis on the job description. We chose 20 skills from the list. For the computation of Google Trends data, we computed the percentage of annual frequency of each skill by averaging the values in the given year. We then subtracted the values of 2017 to 2014 by the value of 2013, and we obtained four values for the differences of annual frequency.

Figure 6 shows the scatter plot, which x-axis stands for the changing percentage in job ads and y-axis stands for changing percentage of frequency in Google Trends.

Because we can get the yearly data from Google Trends, we want to see how the plot would change if we changed the comparison years for the y-axis. Therefore, we overlapped the results of the comparison of two periods in Figure 6. In Figure 6, the blue points show the result of comparison between job ads (2013 vs. 2017) and Google Trends (2013 vs. **2015**). The red points show the results of comparison between job ads (2013 vs. 2017) and Google Trends (2013 vs. **2017**). We first discussed the comparison which used the Google Trends and job ads of 2013 and 2017 (red points). After that, we discussed how the different period affect the results of the comparison.

Table 4: Pearson correlation coefficient between changing percentage in Google Trends and job ads.

PCC	2013 - 2014	2013 - 2015	2013 - 2016	2013 - 2017
IT Job Skill	r = .4370	r = .4542	r = .4782	r = .4973

For the red points in Figure 6, we first noticed that the skills like "spark", "JSON" and "MongoDB" increased more in Google Trends than in job ads analysis. This finding could indicate that the people have high interests in these skills, but these skills are not as trending as people think in the job market.

The similar situation happens in the decreasing skills. For example, "C" decreased by 40% in job ads, but their frequency only dropped 9% in Google Trends. "Android" only decreased 2% in job ads, but its frequency decreased 18% in Google Trends. This finding indicates that these skills' extent of decline in Google Trends and job ads do not similar to each other.

The other interesting case is that people's interests for "microsoft" decreased by comparing its frequency of Google Trends in 2013 with 2015 and 2017, but in the result of job ads, the requirements for "microsoft" increased. This case demonstrates the misconception between what the people think is trending and what is actually needed in the job market. This mismatch between Google Trends and job ads could help us identify the reason of why some positions cannot find a qualified employee.

After discussing the comparison between two datasets in 2013 and 2017, we would like to know whether the mismatch between Google Trends and job ads can be different by changing the comparison years of Google Trends data. By observing the blue points (Google Trends: 2013 vs. 2015), we found that the values of changing percentage of Google Trends are smaller than red points, but the changing direction remains the same. The finding shows that the trend of what skills would increase and what skills would decrease can be already observed in the comparison between 2013 and 2015. We also computed the Pearson correlation coefficient (PCC) between the values of job ads and Google Trends. From the Table 4, we found that there is positive correlation between Google Trends and job ads. As the latter year goes from 2014 to 2017, the positive relationship between Google Trends and job ads become stronger and stronger.

8.2 Google Trends: Trending Energy

In this section, we selected the trending words that we obtained in the advanced analysis on job ads. These words were identified as trending descriptions in job ads by comparing their TF-IDF weights between 2016 and 2017. Therefore, we would like to see whether these words are also trending in Google Trends. Unlike the previous

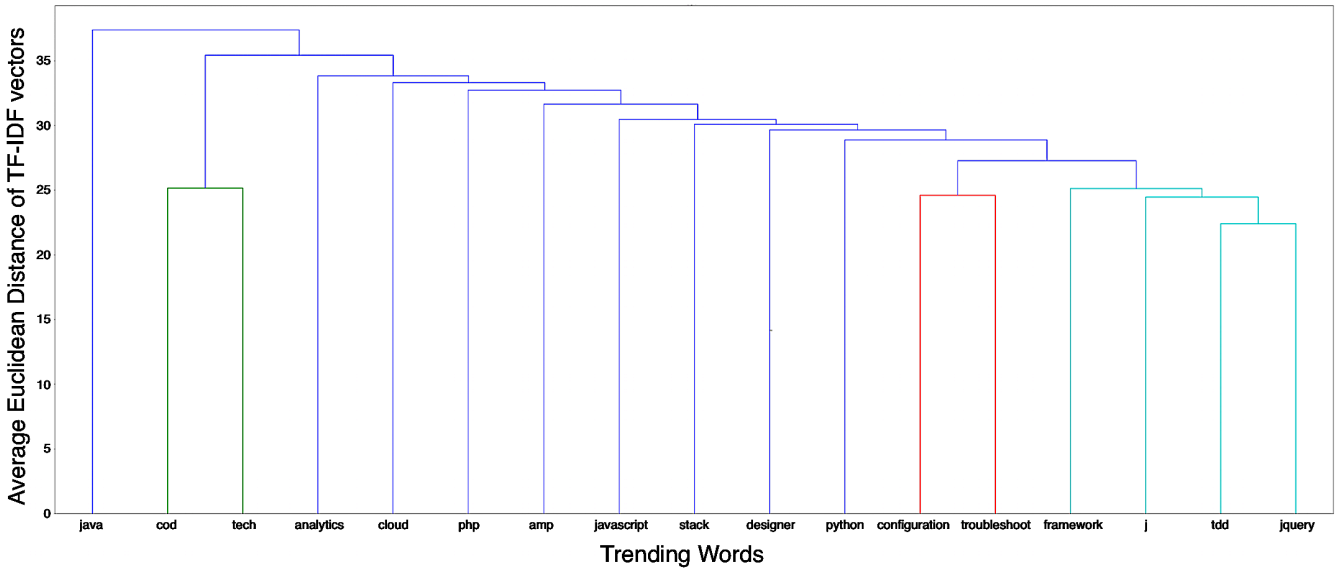

Figure 5: Dendrogram of Trending Words in IT jobs in the UK clustered according to their 2017 TF-IDF vectors

section that computed the difference of the annual frequency of Google Trends, in this section, we used the formula provided by [8] to compute trending energy of Google Trends.

For each case with trending words, we queried Google Trends for a given trending word w during the period from January 2016 to June 2017. Each data point obtained is an integer representing one month. We computed the trending energy of a word w in a set of months M using the formula in Equation 5. It takes into argument the vector v of $|M|$ Google Trends values. A negative value indicates decreasing popularity, and vice-versa for a positive value. According to the result from the previous section, we hypothesized that the trending energy of the trending words would be positive for the most of the cases.

$$energy(v) = \sum_{m=1}^{|M|-1} \left(\left({v_{|M|}}^2 - {v_m}^2 \right) * \frac{1}{|M| - i} \right) \tag{5}$$

Figure 7 shows the trending energy of Google Trends for the trending words that we obtained by analyzing UK's IT job ads. Opposite to our hypothesis, we noticed that the Google Trends energies are negative for the most of the trending words, except for TDD (Test-Driven Development) and Stack. This finding indicates that the trending words in the job market are not considered trending to the public according to the words' trending energies of Google Trends. It is worth noting that our results are definitively domain-specific, whereas word sense disambiguation cannot be done in a Google search query for polysemic trending words, such as python, stack, amp, and framework. Moreover, whereas the trend energies obtained with Google Trends are representative of queries done by the general audience, our trends stem directly from, and thus give insights of, the IT job market demand, and more broadly domain-specific job market demands. In addition to the influence of domain-specificness and polysemic words, we need more analysis to explain this conflicting result. For example, we could perform the same analysis in the condition of selecting the homonyms words or changing the period of Google Trends we collected.

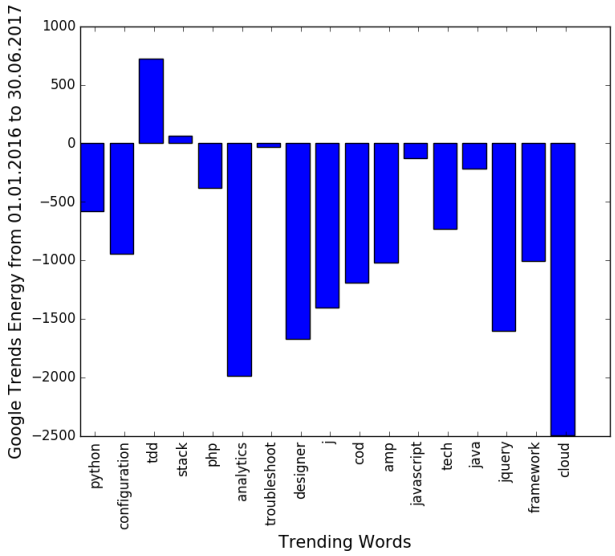

Figure 7: Trending Words in Information Technology jobs in the UK plotted with their Google Trends energy for the time period between 01.01.2016 and 30.06.2017

9 LIMITATIONS AND FUTURE WORKS

The first limitation is the lack of the job ads in intermediate years between 2013 and 2017. Our results can be better supported if we found the similar observation in job ads of intermediate years. The methods from time series analysis can also be applied on job ads. Moreover, because the job ads we used in this study is limited

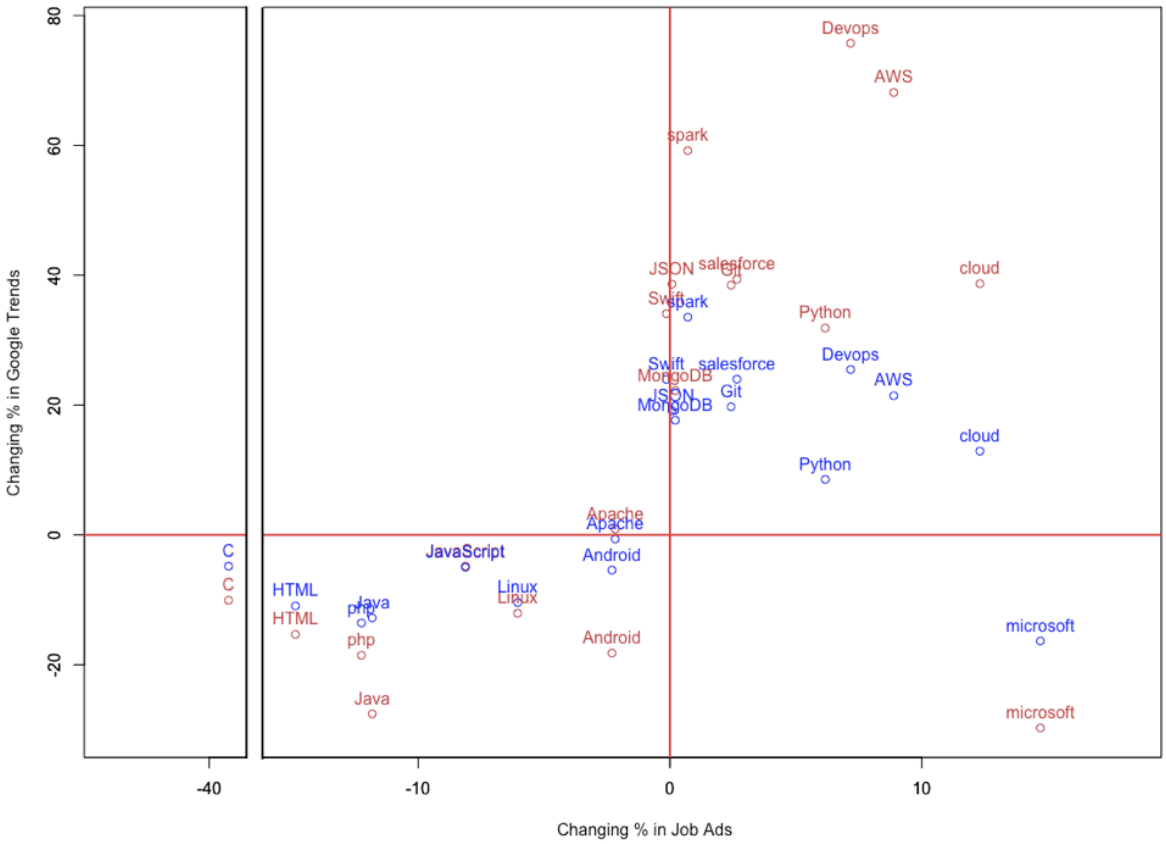

Figure 6: Scatter plot of the changing percentage in Google Trends and job ads.

in the UK and Adzuna.com, the observations in this study do not necessarily reflect the situations of the job market in other countries. More data of job ads need to be collected and analyzed to reach the general results.

The second limitation is the normalization method on job titles. Although we solved the problems mentioned in section 6.1, we still cannot merge the jobs like “C# developer" and “C# engineer" which the meaning of developer and engineer are very similar. To solve this problem, we could introduce the lexical dictionary in the normalization process to gather the words that are semantically similar.

Besides the aforementioned future works, we would like to conduct more studies to investigate why this skill is emerging and why the other skill becomes less popular. We plan to combine more datasets, like Stack Overflow, Wikipedia and Online training course, together to see how the same skill performs differently in these datasets.

10 CONCLUSION

In this paper, we collected over 150,000 job ads of IT job market that were posted in the online job board. We used the standard procedure of Natural Language Processing to clean texts in job ads. Using the cleaned job ads, we extracted skill terms by build n-grams model and the predefined list of skills. We calculated the frequency and TF-IDF of each n-gram, and by observing how these values change over time, we identified several trending words and skills in current IT job market. Furthermore, we also investigated whether people are aware of these trending skills in the job market by looking the Google Trends data of these skills. We found that some skills are trending in both job market and Google Trends, like DevOps and SWA, while some skills performed differently in Google Trends and job ads, like Microsoft and AMP. We believe that these findings give us a good starting point to investigate structural unemployment from the perspective of data mining. In the end, we plan to continue collecting online job ads to refine the current dataset and to strengthen the current findings. After we collect a larger set of job ads, we would like to apply methods from time series analysis and forecasting to identify trending skills from historical information.

REFERENCES

[1] Rakesh Agrawal and Ramakrishnan Srikant. 1995. Mining sequential patterns. In *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*. IEEE, 3–14.

[2] Andrew Aken, Chuck Litecky, Altaf Ahmad, and Jim Nelson. 2010. Mining for computing jobs. *IEEE software* 27, 1 (2010), 78–85.

[3] Andrew Aken and Michael D Michalisin. 2007. The impact of the skills gap on the recruitment of MIS graduates. In *Proceedings of the 2007 ACM SIGMIS CPR conference on Computer personnel research: The global information technology workforce*. ACM, 105–111.

[4] Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2015. Polyglot-NER: Massive Multilingual Named Entity Recognition. *Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, British Columbia, Canada, April 30 - May 2, 2015* (April 2015).

[5] Samer M Barakat, Khalil Yaghi, and Zaina Hamdan. 2011. MIS Students Perception of Most Wanted MIS Job Market Skills. *Computer and Information Science* 4, 3 (2011), 33.

[6] Herman Anthony Carneiro and Eleftherios Mylonakis. 2009. Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clinical infectious diseases* 49, 10 (2009), 1557–1564.

[7] Yan Carrière-Swallow and Felipe Labbé. 2013. Nowcasting with Google Trends in an emerging market. *Journal of Forecasting* 32, 4 (2013), 289–298.

[8] Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. 2010. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the tenth international workshop on multimedia data mining*. ACM, 4.

[9] Stefan Debortoli, Oliver Müller, and Jan vom Brocke. 2014. Comparing business intelligence and big data skills. *Business & Information Systems Engineering* 6, 5 (2014), 289–300.

[10] Kyootai Lee and Dinesh Mirchandani. 2009. Analyzing the dynamics of skill sets for the US information systems workforceusing latent growth curve modeling. In *Proceedings of the special interest group on management information system's 47th annual conference on Computer personnel research*. ACM, 113–120.

[11] Sungjick Lee and Han-joon Kim. 2008. News keyword extraction for topic tracking. In *Networked Computing and Advanced Information Management, 2008. NCM'08. Fourth International Conference on*, Vol. 2. IEEE, 554–559.

[12] Brian Lent, Rakesh Agrawal, and Ramakrishnan Srikant. 1997. Discovering Trends in Text Databases.. In *KDD*, Vol. 97. 227–230.

[13] Juanzi Li, Qi'na Fan, and Kuo Zhang. 2007. Keyword extraction based on tf/idf for Chinese news document. *Wuhan University Journal of Natural Sciences* 12, 5 (2007), 917–921.

[14] Chuck Litecky, Andrew Aken, Bipin Prabhakar, and Kirk Arnett. 2009. Skills in the MIS job market. *AMCIS 2009 Proceedings* (2009), 255.

[15] Chuck Litecky, Amy J Igou, and Andrew Aken. 2012. Skills in the management oriented IS and enterprise system job markets. In *Proceedings of the 50th annual conference on Computers and People Research*. ACM, 35–44.

[16] Chuck Litecky, Bipin Prabhakar, and Kirk Arnett. 2006. The IT/IS job market: a longitudinal perspective. In *Proceedings of the 2006 ACM SIGMIS CPR conference on computer personnel research: Forty four years of computer personnel research: achievements, challenges & the future*. ACM, 50–52.

[17] Yingjie Lu, Pengzhu Zhang, Jingfang Liu, Jia Li, and Shasha Deng. 2013. Health-related hot topic detection in online communities using text clustering. *Plos one* 8, 2 (2013), e56221.

[18] Bipin Prabhakar, Charles R Litecky, and Kirk Arnett. 2005. IT skills in a tough job market. *Commun. ACM* 48, 10 (2005), 91–94.

[19] Bipin K Prabhakar, Charles R Litecky, and Kirk Arnett. 1996. A longitudinal analysis of job skill trends in the MIS job market. In *Proceedings of the Second Americas Conference on Information Systems*.

[20] Deborah Richards and Mauricio Marrone. 2014. Identifying the education needs of the business analyst: an Australian study. *Australasian Journal of Information Systems* 18, 2 (2014).

[21] Gerard Salton and Michael J McGill. 1986. Introduction to modern information retrieval. (1986).

[22] Robert Sokal and C Michener. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin* 38 (1958), 1409–1438.

[23] Ramakrishnan Srikant and Rakesh Agrawal. 1996. Mining sequential patterns: Generalizations and performance improvements. *Advances in Database Technology - EDBT'96* (1996), 1–17.

[24] Sandeep Tata and Jignesh M Patel. 2007. Estimating the selectivity of tf-idf based cosine similarity predicates. *ACM Sigmod Record* 36, 2 (2007), 7–12.

[25] Peter A Todd, James D McKeen, and R Brent Gallupe. 1995. The evolution of IS job skills: a content analysis of IS job advertisements from 1970 to 1990. *MIS quarterly* (1995), 1–27.

[26] Christian Wartena and Rogier Brussee. 2008. Topic detection by clustering keywords. In *Database and Expert Systems Application, 2008. DEXA'08. 19th International Workshop on*. IEEE, 54–58.

[27] Christian Wartena, Rogier Brussee, and Wout Slakhorst. 2010. Keyword extraction using word co-occurrence. In *Database and Expert Systems Applications (DEXA), 2010 Workshop on*. IEEE, 54–58.

[28] Richard W Woolridge and Rachida Parks. 2016. What's In and What's Out: Defining an Industry-Aligned IS Curriculum Using Job Advertisements. *Journal of Higher Education Theory and Practice* 16, 2 (2016), 105.

[29] Shuai Zhang, Shufen Zhang, Xuebin Chen, and Xiuzhen Huo. 2010. Cloud computing research and development trend. In *Future Networks, 2010. ICFN'10. Second International Conference on*. Ieee, 93–97.