THE UNIVERSITY OF MELBOURNE

Profiling Electricians to Improve Public Safety

Data Science Industry Project

Group 9

Chen Zhou (987776)

Meghna Panda (1128829)

Vignesh Lakshminarayanan (1032043)

Xinyu Mao (1091647)

Master of Data Science

November 2021

Faculty of Science

Submitted in the total fulfilment for the Industry project

Master of Data Science at the University of Melbourne

Abstract

The internet has revolutionized our life's and it's our preferred medium for getting any services with the help of a few clicks. The online job platforms have been violated in relation to an unlicensed electrician accepting electrical work which is putting the people life at risk. The aim of the project is to extract the worker profiles and task listings on online platforms from websites like Airtasker, Gumtree etc to find out their demographics, licensing and registration details and then apply natural language processing methods on the aggregated data to find out the seriousness of work carried out by the tasker. The insights were interesting. There are 475 electricians not added their licensing details. People see the ratings or reviews to select the tasker rather than choosing a licensed electrician. The project was particularly interesting as it covered the end-to-end pipelines of data science starting from data collection, performing exploratory data analysis, modelling the data and visualization of results using Tableau. This project helped the client solve their problem.

Keywords: Web crawling, Web scrapping, Latent Dirichlet allocation, pyLDAvis, Tableau, Natural Language Processing (NLP).

Declaration

We certify that this report does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university and that to the best of my knowledge and belief it does not contain any material previously published or written by another person where due reference is not made in the text. The report is 9158 words in length (excluding text in images, tables, bibliographies and appendices).

Signed by:

vignesh lakshminarayanan

Meghna Panda

Xinyu Mao

Chen Zhou

Acknowledgements

We would like to express our gratitude to Chunhua Liu, our supervisor. She was always available for us when we had problems with our writing or the project. She was quite helpful in directing us in the proper route. We also want to express my gratitude to Jessie Nghiem and Priya Laxman from Energy Safe Victoria. This task would not have been possible without her invaluable counsel and experience. We would also want to express our gratitude to Micheal Kirley and Marika Benetti-Hille for their ongoing project advice.

Contents

1 Introduction

The internet has revolutionized our life's and it's our preferred medium for getting any services with a help of few clicks. There are many online ombudsman's websites like Airtasker, Gumtree that enable users to outsource everyday tasks such electrical, handyman's job from an online and mobile marketplace. Users can request by posting their task and quote a budget to complete the work, community members then bid to finish the task. It is illegal for an unlicensed tasker to undertake electrical installation work in Victoria as Unlicensed electricians lack the knowledge and skills necessary to do electrical work that meets installation standards, as well as the ability to test the work to ensure that it is electrically safe.

But over the recent years statistics show that Online jobs platform such as Airtasker has admitted to its community guidelines being breached in relation to an unlicensed electrician accepting electrical work which is putting the workers themselves and the public at risk. It has either resulted in severe injury or even death. Also, requiring re-working on the job to ensure the safety standards. Recently a man convicted after carrying out electrical work while being unlicensed. Australia's Competition and Consumer Commission (ACCC) has previously flagged the need for a "Digital Platforms Ombudsman" to regulate companies like Airtasker and Gumtree [1].

Energy Safe Victoria (ESV), our client, is the state's safety regulator for electricity, gas and pipelines. Under the Energy Safety Victoria Act 2005 ESV is responsible for monitoring, auditing, and enforcing compliance with the requirements as well as administering licensing, registration and approval systems that maintain safety standards and skills [2]. The main challenge for the client (ESV) is to identify all the unlicensed electricians who are carrying out electrical work in online platforms like Airtasker, Gumtree.

The project aims to look at the tasker profiles and task listings on online platforms like Airtasker, Gumtree websites to find out their demographics, licensing and registration details of electricians or handymen who offer doing electrical jobs. Then apply Natural language processing techniques to find the seriousness of work carried out by the tasker.

Some of the data science research questions include:

1) What are the characteristics that are required to identify licensed and unlicensed electricians who are working on online platforms?

2) To what extent the underlying electricians are actually licensed?

3) How to validate the information provided by the tasker?

The project's key hurdles were identifying the business problem, web scraping data from a variety of sources, cleaning the scraped data, and conveying the results to the client via interactive dashboards.

2 Related Work

In Victoria, having an electrician license ensures that the electrical installation is safe. Because each license holder in Victoria has the ability to perform all forms of electrical installation work without supervision. A completed Licensed Electrician's Assessment (LEA) and at least 12 months of electrical installation experience are two of the most critical qualifications for obtaining a license. [3] As a result, we can determine if an electrician can guarantee safety while performing electrical work based on the requirements of the electrical license.

Web Scrapping: Scraping data was one of the most important responsibilities. Web scraping, often known as web data extraction, is a data scraping technique that collects data from websites. To access the World Wide Web, web scraping software can use either the Hypertext Transfer Protocol (HTTP) or a web browser. Web scraping can be done manually by a software user, but it usually refers to automated activities carried out by a bot or web crawler. It's a sort of copying where specified data is obtained and copied from the internet, usually into a central local database or spreadsheet for later retrieval or analysis. [4] [5].

The process is broken down into three parts: the web scraper retrieves the needed connections from the web, the data is pulled from the source links, and the data is eventually saved in a csv file [4] [5]. The task can be completed using a variety of programming languages, such as SQL or Python.

Scraping data can be done using a variety of methods. We can use tools like Scrapy [5], a strong Web scraping framework for Python that automatically parses online pages and extracts web contents using XPath expressions, and Visual Web Ripper [6], which has additional features and allows users to scrape data from any website. Data can be exported to structured CSV, Excel, or XML format after the data scraping operation is completed. Beautiful Soup [7] is another Python tool for parsing and extracting data from HTML and XML files. It works with your choice parser to allow idiomatic parse tree navigation, search, and modification. Programmers frequently save hours or even days of effort.

Topic Modelling: The retrieval of Information is most necessitated field due to the continuous evolution of information available on the internet, the information is available in the form of text, images, videos, audios, documents and web pages. Over the years there has been a strong interest in text summarising. Topic modelling methods are widely used in natural language processing for topic finding and semantic mining from unstructured documents.

There are many types of topic modelling:

Latent Semantic Analysis (LSA): In this method, Documents are represented in vector-based format to form a semantic content. If a document is to be considered a Bag of Words (BoW), then few anchor words would be adequate to express the theme of the document. Latent Semantic Analysis (LSA) was considered the basic technique for thematic discovery from the text archives [8]. In this method the documents and words which are having similar concepts are mapped together. Words and documents can have many to many relationships, LSA encountered a problem, as a term has many meanings, which is quite common in massive text data. Singular Value Decomposition is used to fix this issue i.e., Words occur together and words can have only one meaning. The final simplification appears impractical and places a constraint on LSA. It is difficult to interpret the resulting dimensions. Even though LSA can predict topics for a given set of documents, however it does not help in document generation process, it requires a large corpus to yield a good result.

Probabilistic Latent Semantic Analysis (pLSA): The limitation of LSA is overcome by pLSA [9]. In LSA it associates a word with the concept, whereas pLSA associates a word with a probability. By allocating each word to a subject taken from a multinomial distribution over topics, the model is built in a more meaningful fashion. As a result, a document would be a matched to numerous concepts or themes, as opposed to a paradigm in which the document reflects only one concept. Asuncion et al. proved that pLSA performs better than LSA [10]. However, pLSA does not does not fit into the completely generative model. It does not offer a way to describe documents that aren't in the given collection.

Latent Dirichlet Allocation (LDA): LDA is one of the most popular topic modelling techniques [11]. It is based on unsupervised probabilistic model. Each document is represented by LDA as a collection of extracted themes. Because each topic is so important for each document that is learnt, it is feasible to determine which topic composition is the best using LDA, it encapsulates the essence of each document. To put it another way, each document has a topic combination that can be summarised which has the most significant weight.

The problem with LSA is that its hard to determine optimal number of dimensions, whereas LDA is usually preferable to pLSA because LDA can quickly generalise to new documents. whereas The document probability in pLSA is a constant in the data set. Due to following reasons we have adopted LDA as topic modelling technique.

3 Methodology

This section explains the steps taken, right from the data collection to displaying our results, in detail. Section 3.1 talks about the exploration of four job advertisement websites - Airtasker, Gumtree, ServiceSeeking and HiPages from which relevant electrician profile and tasks were extracted. After which the data was properly cleaned and filtered to get it ready for analysis which is mentioned in Section 3.2. For Modelling, Latent Dirichlet Allocation (LDA) algorithm was deployed to find out about the main tasks of the workers, which is explained in detail in section 3.3. Finally, all analysis were clubbed together as Tableau story and Python plots to demonstrate our results, which is demonstrated in the Result Section of this report.

3.1 Data Collection

The procedure of obtaining, evaluating, and reviewing correct findings using established methodologies is characterised as data collection. Based on the evidence received, all analysis and tests have been performed. The primary goal of data collection in this project was to gather details and accurate features in order to make statistics decisions for analysis and perform relevant modelling activities. There are various methods for scraping data, but we chose the most effective one. Beautiful Soup and Selenium, both Python packages, are used to scrape all the data. BeautifulSoup is best suited as it can scrape data from JavaScript-enabled websites in a simple and effective manner. Selenium is a bit complex but is useful for retrieving data which are hidden or are dynamic in the websites.

All of the electrician data is extracted from four different websites. We have extracted two types of information - the worker profile information and the tasks that the workers have performed. The information is extracted from 4 different websites, as recommended by ESV - Airtasker, Gumtree, Service - Seek and HiPages. These four websites are most used by people to post and discover tasks and services. All the data available in the websites has been extracted and then relevant cleaning and filtering has been done.

The websites we used in this project are Airtasker, Gumtree, ServiceSeeking and HiPages. Airtasker is a trustworthy community platform that has linked around 2.1M [12] Australian's looking to outsource jobs and locate local services with those looking for work . Airtasker can assist anyone with easy to complex chores such as house cleaning, handyman labour, admin work, photography, electrical work, and even website development. Gumtree is the chosen marketplace of 7 million Australians. With over 2.4 million ads [13] to choose between across multiple product categories, Gumtree connects market participants in the locality. Anyone may purchase, trade, or seek everything they are looking for. Service Seek and Hipages are two other platforms that serve the same purpose with around 10M posters combined.

The data collection process was divided into two main parts. One is the task data collection, which is only applied in the Airtasker platform because of the particular structure of Airtasker. Furthermore, the information of tasks can be collected for more analysis. Another part is the worker data collection. First, get the personal page URLs from each platform and then obtain workers' information through their personal page.

3.1.1 Task Data Collection

For the task section, the task page link was collected to extract the workers we wanted. Moreover, the task information is also helpful to do further analysis. More details are provided in Section 3.1.1.1 and 3.1.1.2.

3.1.1.1 Task Page Crawling

When surfing the Airtasker platform, we can only get the information of electricians by crawling each task that meets the requirements and then getting into the personal page of workers who want to participate in this work. Thus, the first step we crawl in this platform is tasks crawling.

We get into the tasks list of Airtasker by using the 'browse Task' function on the main page, choosing the area we want, and inputting the task keywords in the search bar. Since we are discussing the safety of electrician tasks, we only need to capture electrician related tasks. After discussion with team members and confirmation with the client, the process we obtained the appropriate task list is as follow:

1) Account Login
   As a visitor without the Airtasker account, people can only see a few sample tasks. Therefore, the step before we start searching is to log in to Airtasker.

2) Options Selection
   There are five options we can select on the task browsing page.
   The first choice is the way the task can be done. These tasks can be done in person and remotely. We select the in-person type according to the client requirement.
   The second option is the suburb where people want to hire workers and the distance centred around the area people choose. We cannot select 'Victoria' as the suburb name because the longest distance we can choose is 100km. It is clear that Victoria's entire missions cannot be captured within 100km of Victoria's central point. In order to cover the whole of Victoria as much as possible, we have selected the following seven suburbs

as the centres of the circles for 100km.

Surburbs List Used in Airtasker	
Melbourne VIC, Australia	Geelong VIC, Australia
Bendigo VIC, Australia	Horsham VIC, Australia
Warrnambool VIC, Australia	Shepparton VIC, Australia
Traralgon VIC, Australia	

Table 1: Surburbs List

Besides, there are many ‘Melbourne’ and ‘Geelong’ tasks, so adding these two locations to the searching area can prevent missing tasks. This process will produce many duplicate values, so eliminating duplicate values when crawling the task page list is an essential step.

The third option is the range of price in tasks. To cover all the electrical tasks, we keep the default value, ranging from AU$5 to AU$9999.

The final option is the keyword input we used to search the tasks. In general, different websites have different keyword search mechanisms. For example, some websites can only find out the tasks whose names contain keywords. However, the Airtasker website can find all the tasks whose names and descriptions contain keywords people want. This mechanism allows us to find as many electrical tasks as possible using only electric related extensions as search keywords. Thus, the keywords are [’electrician’, ’electric’, ’electronic’].

3) Tasks Status Determining

After determining what type of task we want, we can crawl the list of all task page URLs. Next will be to crawl the worker page list on each task page. We get the worker page link in the ’OFFERS’ section of each task page. In general, we only need to crawl the worker list in the ’OFFERS’ section because they show the workers who have given the poster quotation. Nevertheless, during this part, we found that only the task assigned worker remained in the ’OFFERS’ area when a task had been assigned. At the same time, others in the competition will be removed from the task page. To avoid missing workers, we have to get the current state of the task while crawling and then separate closed tasks from open tasks. Besides, we need to constantly grab these open tasks and get the list of their workers until the task is closed—usually, the capture interval range from one day to one week. Thus, determining the current status of each task when crawling is a critical stage.

’Open’, ’Assigned’, ’Expired’, ’Cancelled’ and ’Completed’ are five kinds of task status. The corresponding measure we deal with different status of tasks have been shown in the following table.

status	measures
Open	The number of workers in task page will keep increasing, need to be followed up
Assigned	The information in the task page would not change anymore. The worker list in task can be extracted directly
Completed	
Expired	
Cancelled	

Table 2: Measures of Different Task Status

4) Tasks Pages Crawling

In crawling the task page and extracting the workers’ URLs, we found that the number of tasks will increase over time, and the workers’ page will be updated in open tasks. Thus, writing a code that can run the entire tasks fetching process at once is necessary. However, the whole process includes login step, options selection and status determination, which requires a lot of manual keyboard input operations and mouse selection operations. Moreover, the tasks list is refreshed for 30 more tasks when the wheel reaches the bottom, requiring people to scroll the mouse wheel constantly until there are no more tasks to refresh. So, getting code to do this automatically is a big challenge in the process.

The selenium package in Python solved these challenges. It can automate web browser interaction from Python, load the specific webpage people want to crawl, simulate keyboard input and mouse action.

For the login step, selenium opens a new Chrome browser and load the Airtasker main page. Then, the code finds the login elements using the Xpath and sends the username and password to the corresponding position. The time.sleep() function will be introduced to prevent the browser from jumping to the corresponding page not in time due to network latency. After logging into the Airtasker, the selenium loads the 'tasks browser' page and then simulates the mouse action to select the options we want.

For the tasks refresh mechanism in Airtasker, selenium also addresses this infinite scrolling scrapping problem perfectly. The code scrolls the website straight to the bottom. After waiting for 1s, if the position of the scroll wheel because the task refresh is not at the bottom, continue scrolling to the bottom using the code. This loop continues until there are no more updates.

5) Workers URL Getting According to Tasks Status

When scrapping the workers' URLs on each task page, we mainly divide them into two categories. One is the closed tasks, consisting of the assigned tasks, completed tasks, expired tasks and cancelled tasks. They are closed because the workers' list in the 'OFFERS' section would no longer change. All we should deal with these closed tasks is extract the workers' page they have. Another one is the open tasks. The number of workers will keep increasing until the task closes. Thus, we need to continuously extract workers in these tasks until the task is closed to update our comprehensive list of workers.

3.1.1.2 Task Information Crawling

Based on the electrical work tasks posted in Airtasker, we would like to extract them from the following features(Poster name, address, due date, description, prize and their title). Since we have been crawled the lists of all the electrical websites URLs due to the current date, based on these URLs, we scrape the information from each URL using the python tool beautiful soup. In the end, we would summarize the data into a CSV file for later analysis.

The purpose of storing the information is to analyse the task description using the LDA topic modelling. Based on the result of LDA modelling, we need to identify the seriousness of electrical tasks for those licensed electricians versus unlicensed workers such as handymen.

3.1.2 Worker Data Collection

The worker data collection was divided into two parts. One is the worker URL crawling, which is used to get the link from each platform. Another one is the information scrapping. We crawl all the worker information we need from each worker's personal page. More details are provided in Section 3.1.2.1 and 3.1.2.2.

3.1.2.1 Worker Personal URLs Crawling

Gumtree Gumtree is the most popular marketplace in Australia, with over seven million Gummies join. It connects buyers and sellers in the local community, with over 80,000 new ads added every day in categories like Home Garden, Baby Children, Sport Fitness, Clothing Jewellery, and Gumtree Jobs (Gumtree Australia Support Knowledge base - Basics - What Is Gumtree?, n.d.) [14].

Because of the complexity of web services, it is not easy to find the electrical information we need. For example, when people search for an electrician, the results come up not only for electricians but also for electrical jobs and appliances for sale. Besides, the workers' page URLs cannot be extracted from the electrical task page because of the secrecy between workers and posters. All the messages between workers and posters are only visible to each other. Thus, a page crawler cannot extract workers' information on the task page.

After a series of tests, the workers' page URLs can be obtained from the following path: Services For Hire → electrician → Victoria → Offer Type: Offering. Because Gumtree is mostly about display ads, the personal pages that get crawled are all about companies that provide electrical services to people.

The next step is crawling the search result page and scraping the workers' pages listed on that. However, the number of the result pages is not only one, which means that the page-turning capability needs to be present in the code. We found all the valuable page URLs can be get in the 'div[class="page-number-navigation"]' section. Storing the page URLs down first and scrawling them one by one can be easily implemented.

Service Seeking Service Seeking is an online marketplace for consumers and businesses to trade services. Consumers can submit their jobs on Service Seeking's websites, and firms can respond with quotes. Accounting, construction, graphic design, photography, and furniture removal are their most famous work categories (Mason, 2012) [15].

The electricians' list can be easily found on this platform because it has a separate category for electricians. However, the list is difficult to extract. The site can only get the list of nearby workers based on the designated suburb, and the number of suburbs in Victoria is too much. Fortunately, there is a 'browse by location' page stores all the suburbs searching link. There 101 suburbs are listed on that page. Although this is not all suburb names for Victoria, it contains almost all worker lists because the platform recommended it. Thus, scraping these regional links off the recommendation page is the first thing to do.

Then in the process of scraping the workers' page URLs, the 'View More' button prevents us from fetching all worker links directly from the regional search results websites. The selenium package in Python is used to solve this problem. The code uses the Chrome web driver to load the location result page respectively, then checks whether the 'View More' button exists or not. If the button exists, simulate the mouse clicking on the button to let the website load more workers' URLs. Moreover, this operation continues until the code cannot detect the 'View More' button.

HiPages The HiPages is an online platform that links Australians with reliable tradies to make home improvement easier for them while also assisting workers in growing their businesses. Homeowners can find reliable local workers by posting a job on the marketplace. They can then look over the tradespeople's profiles, compare quotes, and hire the best trades person for the project (Palmer-Derrien, 2020) [16].

The crawling mechanism for this platform is the same as for Service Seeking. It has a dedicated electricians classification partition but needs to search for a list of nearby workers based on the input region. Moreover, this platform also has his commended suburbs list for us to traverse.

The ’View More’ button question also appears on this platform. We solved the same problem with the Service Seeking platform.

The new problem appears when the code is running. The connection with HiPages will be refused if there are too many requests in a short time. To address it, we introduced the error handling mechanism loop and time sleep mechanism in the code. Once the error appears again, the code will sleep ten minutes and crawl the next page after sleeping.

3.1.2.2 Worker Information Scraping

In total, 1827 worker profile information was scrapped from all the websites mentioned above for all of Victoria. All the websites had different features and only the common features from all 4 websites were taken for analysis. These common features are as follows:

1) Name - The personal or company name associated with the worker.

2) Address - The address where the worker is situated. The address includes the street name, suburb or area name and the LGA name.

3) License Number - All professional workers have a license number. A license number indicates that they have been verified by the government to do the tasks. For this project, the only interest was to scrape the licenses associated with any electrical task.

4) Description - this free-form text field allows the workers to bring all of the elements of their profile together to express themselves where they can describe their personality and how they can contribute in the best way possible. Every worker has a short description about themselves where they introduce themselves stating their qualifications, experience and skills.

5) Rating - Number of stars each electrician has been given by the customer out of 5.

6) Reviews - Feedback each customer has given the worker.

3.1.3 Crawling and Scraping Challenges

There are many challenges during the process of crawling and Scraping. The first challenge arises in the process of obtaining workers' URLs. We have to get the workers' page link in different platforms with different search mechanisms. Therefore, before we start crawling the platform, we have to find the right way to get the list of electricians we need. Not all platform-accessible information is available in this project. We have tried some well-known trading platforms in Australia, such as Market Place and Service.com. Only the Airtasker, Gumtree, ServiceSeeking and HiPages are available for us. For Airtasker, we need to access the workers' URLs from the tasks page. Nevertheless, for ServiceSeeking and HiPages, the recommended location list by the site itself is necessary for the workers searching.

Moreover, different platforms have different website frames. The way we crawled the workers' URLs needs to be treated differently. Airtasker websites display a list of their workers through infinite scrolling. Some websites display their list by constantly clicking the 'View More' button, such as ServiceSeeking and HiPages.

3.2 Data Processing

The transformation of data into a usable and desired structure is referred to as data processing. A lot of data cleaning was required for this project, as described in Section 3.2.1. We separated the entire dataset into three labels, which are detailed in Section 3.2.2.

3.2.1 Data Cleaning

Most of the data cleaning tasks have been performed using python and Microsoft Excel. The worker data extracted from all four websites was very messy and a lot of cleaning had to be done to get the data ready for analysis. All symbols, punctuation and white spaces were removed from the text data. Not all workers had a rating present, therefore these missing values were marked as 0.

The description column is one of the most important features of our dataset as this is used for Topic Modelling later on. Unfortunately, most of these were not present in the dataset. It was impossible to omit these rows as we would have lost a lot of data. For such cases, each worker review that is mentioned in their webpage was gone through manually. Based on these, we could understand the task that the work has done. For these, we could make up a short description in the dataset. Even after this at most 20 workers had no information present in their profiles and they were brand new. Since we were not able to identify what type of tasks exactly these workers are associated with, these were removed from our dataset.

One more challenge was encountered with the extraction of the license number. Some workers in the websites perform multiple chores and therefore have multiple licenses associated with them. We had to be careful to only extract the license number associated with electrical work. The license numbers, related to electrical work, could be identified as follows:
Numbers starting with REC (e.g., REC 11224) - Registered Electrical Contractors (REC) are required by Energy Safe Victoria to operate in a safe and compliant way. Applicants must have the necessary experience, technical and business knowledge, and proven ability to perform electrical installation work in order to be registered as a REC. [3]
Numbers starting with A (e.g., A14352) - An A grade electrician's license allows the holder to perform all forms of electrical installation work without supervision in Victoria. The holder of this license is not permitted to do electrical installation services for profit or reward For this, the electrician should also hold a Registered Electrical Contractors (REC) license. [3]

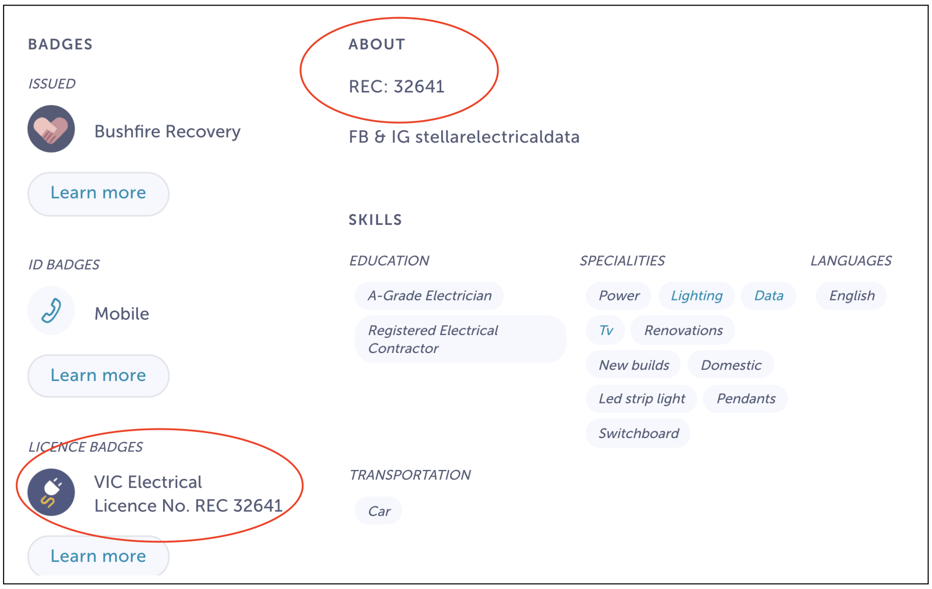
All the data extracted from the four websites was combined into one dataset. But some workers had their information mentioned in two or more websites. These duplicate workers were removed for the final dataset.

3.2.2 Data Filtering

The final dataset had very limited information, so for analysis some extra columns had to be taken. From the Address field, the suburb and postcode information had to be extracted separately for demographic analysis and visualization.
The dataset consists of information about workers associated with all kinds of tasks. The tasks vary from electrical to cleaning and plumbing and many more. We are only interested

in workers that are electricians or those who perform electrical or electronic tasks. To filter out this information in detail, a “license Information” column was made. This column has three fields:

1) Licensed Electricians - Electricians who have a valid license number explicitly mentioned in their website. The licensed can be mentioned separately or in the description column. Figure 1 shows an example:



BADGES

ISSUED

Bushfire Recovery

Learn more

ID BADGES

Mobile

Learn more

LICENCE BADGES

VIC Electrical
Licence No. REC 32641

Learn more

ABOUT

REC: 32641

FB & IG stellarelectricaldata

SKILLS

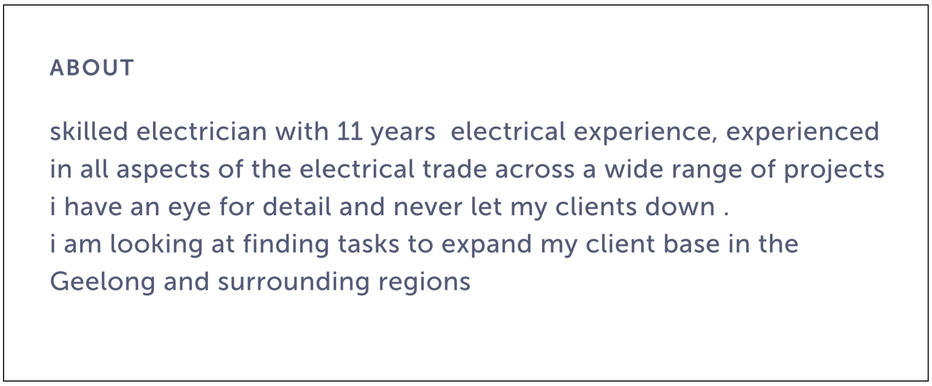
EDUCATION	SPECIALITIES	LANGUAGES
A-Grade Electrician Registered Electrical Contractor	Power, Lighting, Data, Tv, Renovations, New builds, Domestic, Led strip light, Pendants, Switchboard	English

TRANSPORTATION

Car

Figure 1: License Number Indications [17]

2) Electricians - License Number not provided - These workers have mentioned that they are fully licensed in their descriptions but have not mentioned their license numbers anywhere in their profiles. Other keywords like - "Registered", "A-grade", "skilled-electrician", "licensed" were also mentioned in their descriptions which helped to classify them accordingly. Figure 2 shows an example.

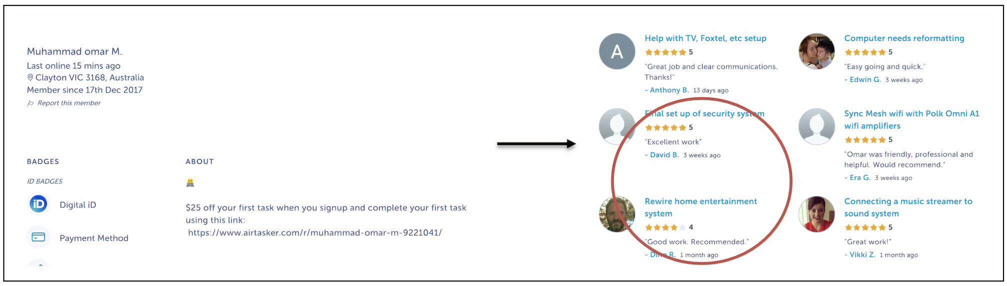


ABOUT

skilled electrician with 11 years electrical experience, experienced
in all aspects of the electrical trade across a wide range of projects
i have an eye for detail and never let my clients down .
i am looking at finding tasks to expand my client base in the
Geelong and surrounding regions

Figure 2: No License provided [18]

3) Not electrician, but performed electric tasks - These are workers who are not electricians. They are mainly associated with other tasks such as cleaning, plumbing etc. But upon reviewing their profiles, they have performed basic handyman tasks like electrical installations and repairs. Instead of removing these cases completely, they have been marked differently for analysis.



Muhammad omar M.
Last online 15 mins ago
Clayton VIC 3168, Australia
Member since 17th Dec 2017
Report this member

BADGES

ID BADGES

Digital iD

Payment Method

ABOUT

$25 off your first task when you signup and complete your first task using this link:
https://www.airtasker.com/r/muhammad-omar-m-9221041/

Help with TV, Foxtel, etc setup
★★★★★ 5
"Great job and clear communications. Thanks!"
- Anthony B. 13 days ago

Computer needs reformatting
★★★★★ 5
"Easy going and quick."
- Edwin G. 3 weeks ago

Final set up of security system
★★★★★ 5
"Excellent work"
- David B. 3 weeks ago

Sync Mesh wifi with Polk Omni A1 wifi amplifiers
★★★★★ 5
"Omar was friendly, professional and helpful. Would recommend."
- Era G. 3 weeks ago

Rewire home entertainment system
★★★★☆ 4
"Good work. Recommended."
- Dino R. 1 month ago

Connecting a music streamer to sound system
★★★★★ 5
"Great work!"
- Vikki Z. 1 month ago

Figure 3: Not electrician, but has done electrical work [19]

	Websites			
License Infomation	Airtasker	Gumtree	HiPages	Service Seeking
Licensed Electricians	85	32	599	171
Electricians - license Number not provided	232	67	NA	176
Not electrician, but performed electric tasks	448	16	NA	NA

Table 3: Data Summary

The above table shows a summary of the data extracted. The most number of licensed electricians were from HiPages with 599 licensed electricians. Airtasker has the most number of non-electricians (who performed electrician tasks) with 448 such workers. Totally we could scrape **1827** workers.

3.3 Analysis Modelling for Worker Data

The analysis and modelling of worker data is done with the help of Latent Dirichlet Allocation and pyLDAvis will be explained in the following sections.

3.3.1 Topic Modelling

Topic modelling is a sort of statistical modelling used to find abstract "themes" in a set of documents. It's is a Unsupervised natural language processing (NLP) technique for identifying repeating patterns of words in a corpus of documents. It can be used to find patterns in a large number of documents, organise big blocks of textual material, and also helps in retrieving information from unstructured text. A topic model clusters documents that are comparable based on the words and expressions that appear most frequently, using patterns such as word frequency and distance between words. These can be then used to find the topics associated with set of words, The figure 4 represents the architecture of topic modelling.

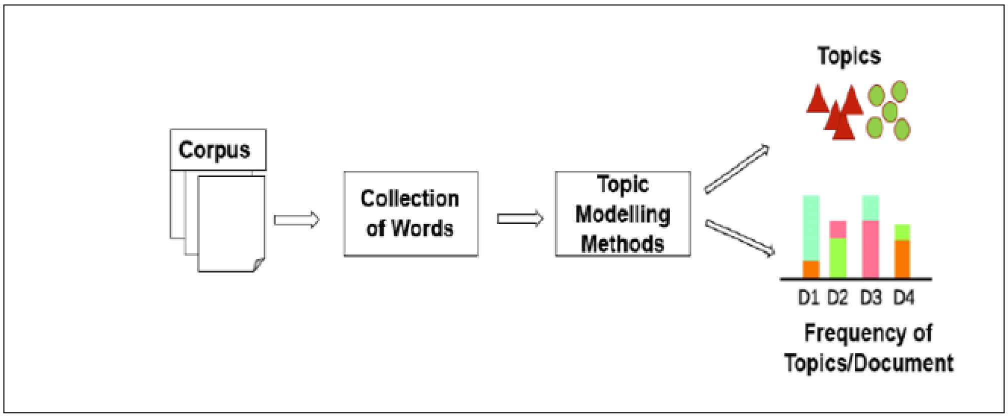


Figure 4: Architecture of Topic Modelling

3.3.2 Latent Dirichlet Allocation

LDA stands for Latent Dirichlet Allocation, it is an unsupervised generative probabilistic model technique used for topic modelling which was introduced in 2003 by Blei et al. It is one of the widely adapted models in the field of Natural language Processing (NLP) for topic modelling. It uses the matrix factorization technique to represent the corpus of document as document-word matrix. After then, the matrix is processed using sampling techniques until the results reach a stable point. The LDA method is applied to scrapped electrician's data from the different websites. The topic modelling is applied to find out the seriousness of the work carried out by each electrician.

3.3.3 Representation and Parameters of LDA Algorithm

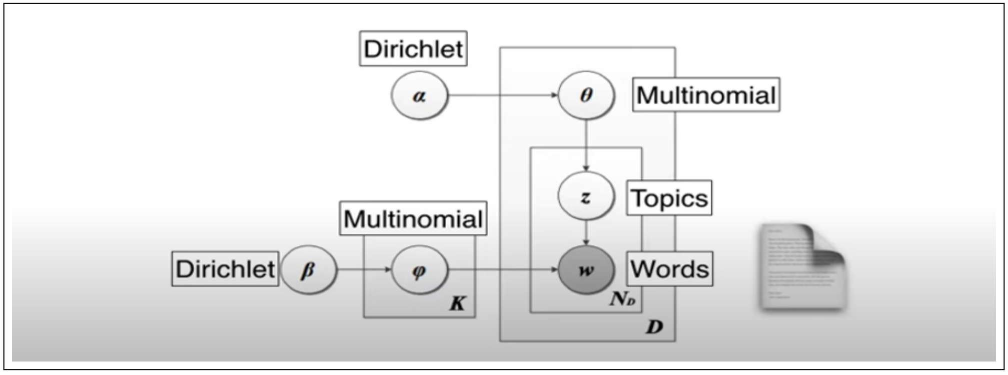


Figure 5: Representation of LDA model

The above figure gives the representation of LDA model and the below table represents the various parameters and notations that are part of LDA model.

Parameters of LDA algorithm	
α	parameter of the Dirichlet prior on the per-document topic distributions
β	parameter of the Dirichlet prior on the per-topic word distributions
θ	follows a multinomial distribution for picking topics
φ	follows a multinomial distribution for picking words
Z	Generates list of topics
W	Genarates list if words for the corresponding topics.

Table 4: LDA Parameters

The joint distribution of the LDA model is given as follows:

$$P(W,Z,\theta,\varphi;\alpha,\beta) = \prod_{j=1}^{M} P(\theta_j;\alpha) \prod_{i=1}^{K} P(\varphi_i;\beta) \prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j) P\left(W_{j,t} \mid \varphi_{Z_{j,t}}\right)$$

The distribution specifies the interdependence's between topic assignment and topic distribution, words in all the documents and their dependency on topics and topic assignments.

3.3.4 LDA Algorithm

To actually infer the topics in a corpus, we imagine a generative process whereby the documents are created, so that we may infer, or reverse engineer, it. We imagine the generative process as follows. Documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over all the words. LDA assumes the following generative process for a corpus D consisting of M documents each of length N_i :

1. Choose $\theta_i \sim \operatorname{Dir}(\alpha)$, where $i \in \{1, \ldots, M\}$ and $\operatorname{Dir}(\alpha)$ is a Dirichlet distribution with a symmetric parameter α which typically is sparse $(\alpha < 1)$

2. Choose $\varphi_k \sim \operatorname{Dir}(\beta)$, where $k \in \{1, \ldots, K\}$ and β typically is sparse

3. For each of the word positions i, j, where $i \in \{1, \ldots, M\}$, and $j \in \{1, \ldots, N_i\}$ (a) Choose a topic $z_{i,j} \sim$ Multinomial (θ_i). (b) Choose a word $w_{i,j} \sim$ Multinomial $\left(\varphi_{z_{i,j}}\right)$ [20]

3.4 Framework of LDA algorithm

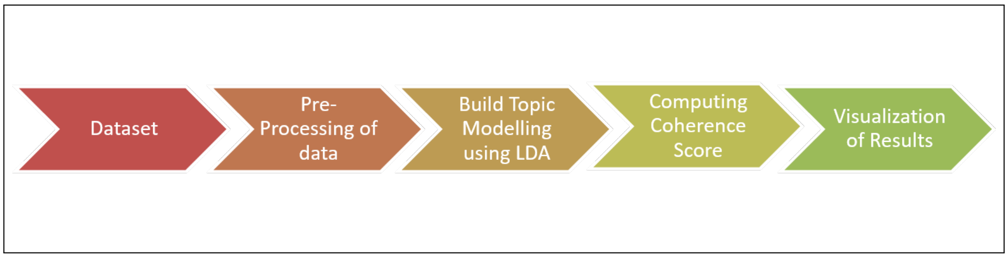


Figure 6: Framework of LDA Topic Modelling

3.4.1 Electrician Dataset

The final dataset hosts around 1827 data about Name, Address, Suburb, License info provided, License info, license number, Description, Rate, Date, website of the electricians who are working in digital platforms like Airtasker, Gumtree, Hipages and Service seek websites. To build topic modelling the worker Description column is used to find out the various topics related to electricians and non-electricians' workers in order to find out the seriousness of the work carried out by each tasker and other columns are used for exploratory data analysis.

3.4.2 Pre-Processing of Data

The natural language toolkit (NLTK) library is utilised to pre-process the data. The individual description of workers is crawled and the following pre-processing steps are applied to clean the data which then can be used for effective modelling of data.

1) Stop words: Removal of English stop words which does not add much information to the text.

2) Case folding: Case folding is done to lowercase the words.

3) Removing punctuation's: The punctuation's are removed from the text for effective processing of data.

4) Text Normalization: **Stemming** is applied to reduce the words to its word stem that affixes to suffixes or to roots words known as lemma. **Lemmatization** is applied to remove the inflectional endings to retain the base or root form of a dictionary word.

5) Tokenization: The sentences are divided into smaller units of words or terms called tokens which then can be stored in a dictionary.

3.4.3 Building Topic Modelling using LDA and Hyper-Parameter Tuning

The LDA model was built using Genism open-source library for performing Unsupervised natural language methods, The following hyper-parameters are chosen after performing the tuning as they yielded the good result for both the data sets. The LDA model is implemented on the scrapped electrician's data of 1826 worker's description, for 50 passes of Gibbs sampling, with number of topics as 4, alpha value set to 0.01 and eta value set to 0.01 with chunk size as 100 through 100 iterations. The model eventually converges to form best model for the given data set. Some of challenges for the model is that there were lot of missing description in the data set, So classified the topic as none, this limited us from achieving more accurate results which is then used in our analysis for later parts.

3.4.4 Model Evaluation Using Coherence Score

The topic coherence scores gives us a measure that helps to differentiate topics that are semantically interpret able topics that are artifacts of statistical inference.

Coherence Score for Different Data set	
LDA model for all workers	0.35
LDA model for non electricians	0.39

Table 5: Model Evaluation

From the above table we can infer that the Latent Dirichlet Allocation model performed better on the Non electricians data set compared to the All worker data set as they have a higher coherence score of 0.39 compared to that of 0.35.

3.5 Analysis for Task Data

We have crawled the task data based on the description of given tasks and applied the LDA model to deeply analyze the data and explore the correlation between the worker's profiles.

3.5.1 Frequency Distribution of Word Counts

When working with a large number of documents, people will want to know how extensive the descriptions are as a whole. Thus, here it is necessary to firstly provide a general word count distribution for all tasks description.

Compared to using a table to summarize the frequency of the word count distribution, we prefer using Python to plot the bar chart demonstrating the distribution of word counts. Additionally, we would provide the mean, median, standard deviation,5% quantile and 95% quantile for the distribution in order to detail the summary.

3.5.2 Distribution of Word Counts for Dominant Topic

For here we would mainly discuss the word counts distribution in dominant topic which is the basis for later word cloud and pyLDAvis visualisation.

3.5.2.1 Visualize the Word Counts Distribution for Each Dominant Topic

In multiple LDA models, each description is composed of multiple topics. Whereas, typically, only one of the topics is dominant. Thus we need to Visualize the dominant topic for each sentence and show the weight of the topic and the keywords in a nicely formatted distribution.

3.5.2.2 Word Cloud for Top n Keyword in Each Topic

A word cloud with the size of the words proportionate to the weight is a pleasing sight, even if people have previously seen the theme keywords in each topic.

3.5.2.3 Plot the Word Cloud for Top n Keywords in Each Dominant Topic

When it comes to the relevance of the keywords in the themes, it matters. In addition, the frequency with which the words appear in the texts is also worth investigating.

In the same chart, plot the word counts and weights of each term.

Keep an eye out for words that appear in numerous themes and whose relative frequency is greater than their weight. Frequently, such phrases turn out to be insignificant. The chart below is the result of re-running the training procedure after adding numerous such phrases to the stop words list at the start.

3.5.3 pyLDAvis for Each Topic

Applying pyLDAvis is the most commonly used and a nice way to Visualize the information contained in a topic model in general.

3.5.4 Task Analysis Summary

To develop the LDA model, we started from scratch by importing, cleaning, and analyzing the task description dataset. Then we learned how to view the outputs of topic models in a variety of ways, including word counts, word cloud, sentence colouring, which intuitively tell you which topic is prominent in each topic. pyLDAvis provide extra information about the topic clustering.

4 Results

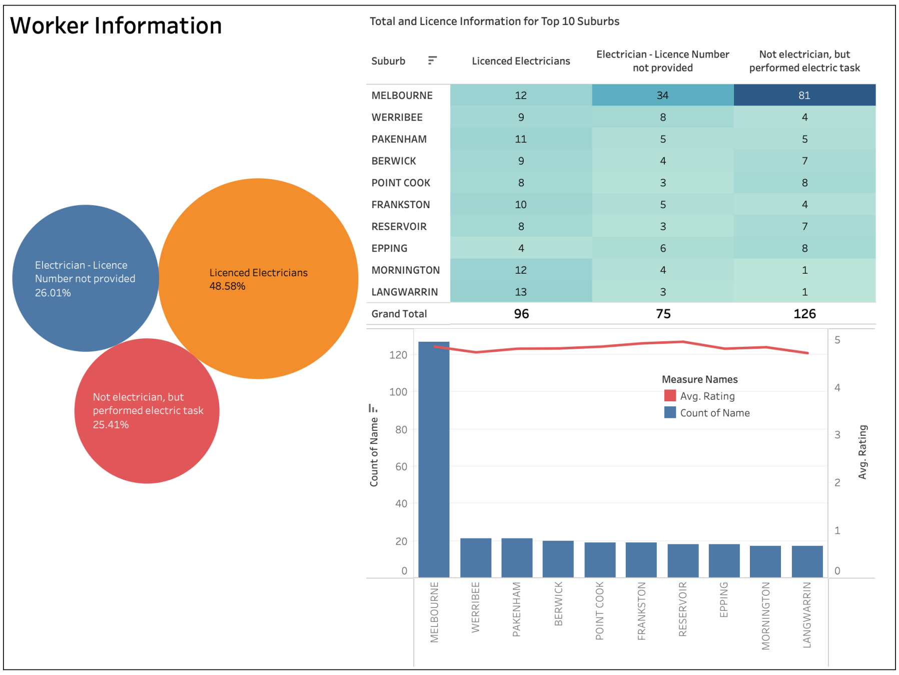
After all the analysis is over, a demonstration of our results is shown in this section. We have used Tableau to demonstrate the results of Worker profiles and demographics, as shown in section 4.1 and python plots to show the results of task analysis of the LDA model as shown in section 4.2.

4.1 Tableau Dashboard

To better Visualize our data and results, we create a storyline in Tableau, consisting of four dashboards. People can interact with filters and functions to see the different patterns in different conditions. The share link of our storyline is as follows:
https://public.tableau.com/shared/QJJFHF68Z?:display_count=n&:origin=viz_share_link

4.1.1 Overall Worker Information

Overall, we can see we have extracted 48% of licensed electricians, 26.01% of electricians who do not have a license numbers explicitly present and 25.41% non-electricians.



Total and Licence Information for Top 10 Suburbs

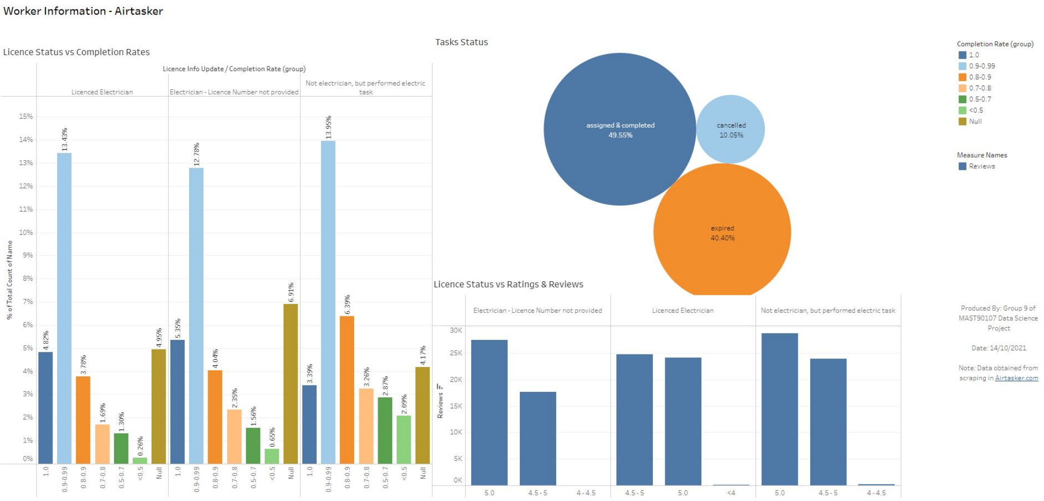
Suburb	Licenced Electricians	Electrician - Licence Number not provided	Not electrician, but performed electric task
MELBOURNE	12	34	81
WERRIBEE	9	8	4
PAKENHAM	11	5	5
BERWICK	9	4	7
POINT COOK	8	3	8
FRANKSTON	10	5	4
RESERVOIR	8	3	7
EPPING	4	6	8
MORNINGTON	12	4	1
LANGWARRIN	13	3	1
Grand Total	96	75	126

Figure 7: Worker Information Dashboard

Melbourne has the most number of workers present, while Langwarrin has the most number of licensed electricians. The average ratings of the workers also remains to be very high across all suburbs of Melbourne.

4.1.2 Worker Information in Airtasker

The second dashboard shows the worker information in Airtasker, which consists of three charts.



Produced By: Group 9 of MAST90107 Data Science Project

Date: 14/10/2021

Note: Data obtained from scraping in Airtasker.com

Figure 8: Airtasker Dashboard

The first chart on the top right corner shows the status among 3400 tasks. Among the 3400 tasks, there is 50% of tasks have been assigned or completed. However, up to 40% of tasks have been expired.

The second graph on the left is the bar chart depending on license status versus completion rate. It is clear that most workers have a completion rate ranging from 90% to 99%, whatever the types of their license status. Comparing the different types of workers, the overall characteristics are similar. The licensed electricians and electricians who did not provide their license have more people on the completion rate of 100% than non-electricians.

The third graph is the bar chart about the number of reviews versus ratings. Interestingly, only the licensed workers have a rating under 4. According to these two bar charts, we can preliminarily know that task posters in Airtasker do not care about whether the workers have the license or not.

4.1.3 Worker Information in Other Three Platforms

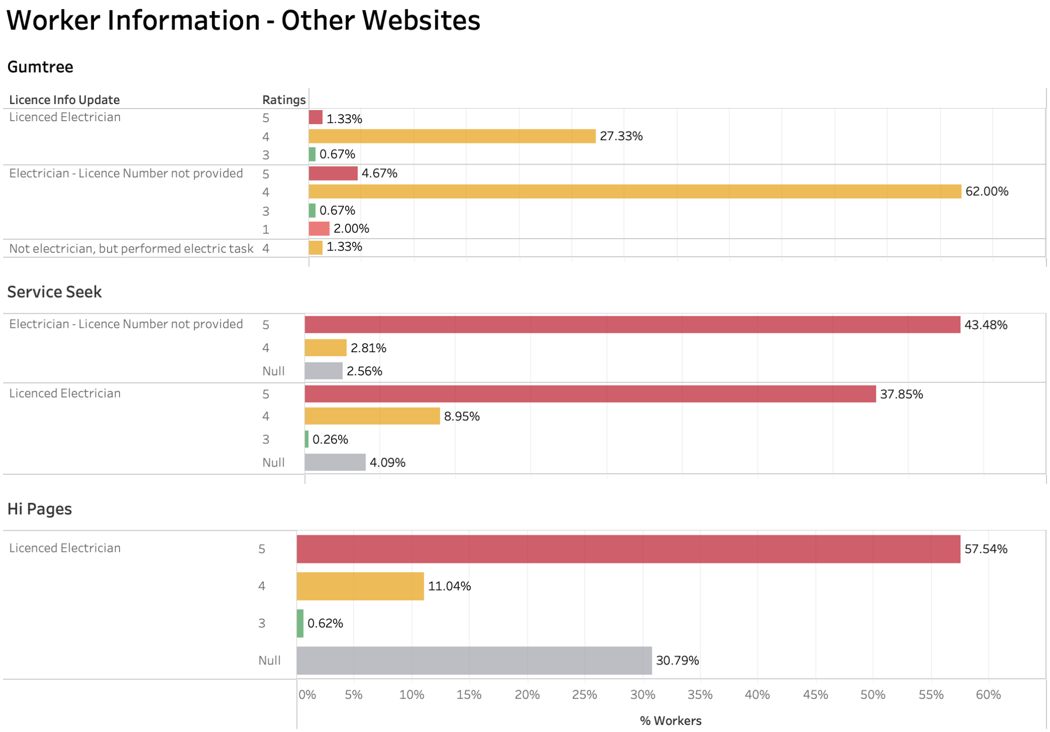


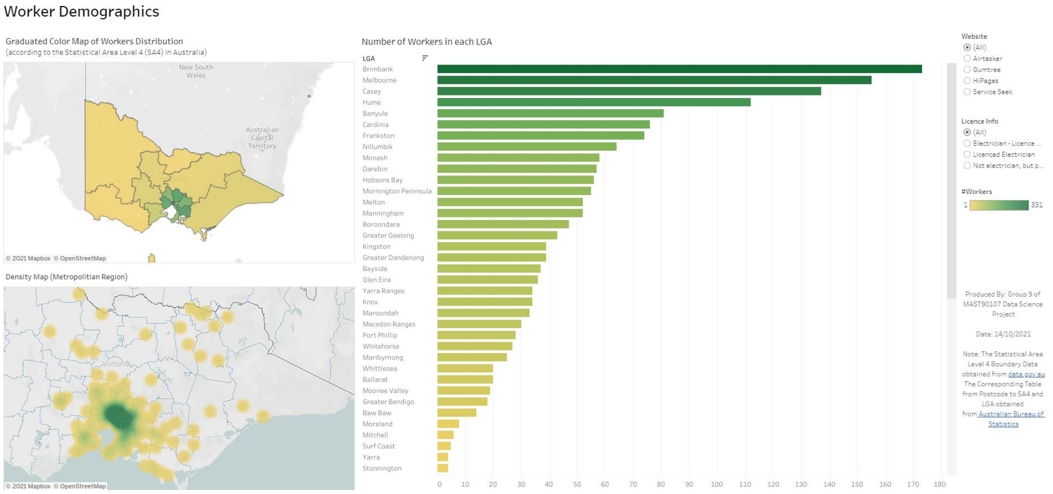
Figure 9: Other Three Platforms Dashboard

The third dashboard is made up of rating data from the other three platforms. HiPages only has licensed electricians. Up to 60% of the companies have a rating of 5. However, there are also 30% of them have no tasks. For the ServiceSeeking, only the licensed workers or the licensed worker but do not provide their number. A large proportion of the companies won the rating of 5. Gumtree has all types of workers. There is no null value, so all the companies had the tasks. Nevertheless, most of them have a rating of 4.

4.1.4 Worker Demographics

We extracted the workers from different platforms. Different platforms may use different geographical expressions. In order to unify them, we use two different frames. One is the Statistical Area Level 4 (SA4) in the ABS structure, and another is the Local Government Area (lga) in the non-ABS structure.

We have the postcode of every worker. According to the correspondence table between postcode and SA4 main code, we assigned each task to their related SA4 area. Similarly, we did the same thing with the LGA allocation process.



Worker Demographics

Graduated Color Map of Workers Distribution
(according to the Statistical Area Level 4 (SA4) in Australia)

Density Map (Metropolitan Region)

Number of Workers in each LGA

Website
● (All)
○ Airtasker
○ Gumtree
○ HiPages
○ Service Seek

Licence Info
● (All)
○ Electrician - Licence ...
○ Licenced Electrician
○ Not electrician, but p...

Produced By: Group 9 of MAST90107 Data Science Project

Date: 14/10/2021

Note: The Statistical Area Level 4 Boundary Data obtained from data.gov.au The Corresponding Table from Postcode to SA4 and LGA obtained from Australian Bureau of Statistics

Figure 10: Worker Demographics

We use three charts to show workers demographics. The graduated Color Map use the SA4 boundary. The darker the colour, the more workers we have in that area.

Density Map is a heating map. We took each worker's postcode and represented them as the points on the map. Then, we can get the density of points.

The right bar chart is the order of the number of workers in different lga areas. In the filtering part, we can check the distribution map on different websites and workers' types.

4.2 Task Analysis within Topic Modelling Results

Based on the LDA model's topic modeling and task analysis, we use the multiple LDA model within four dominant topics for analysis and visualization. Also, these four dominant topics, to a large extent, represent the whole task analysis within the LDA model. The comparison between the four dominant topics helps us better understand the model as well.

4.2.1 Visualization of LDA Model with all Electricians

Here we consider the dataset of all electricians extracted from all four websites.

4.2.1.1 Distribution of Word Counts

Figure 11 shows the Word count and the importance of topic keywords.

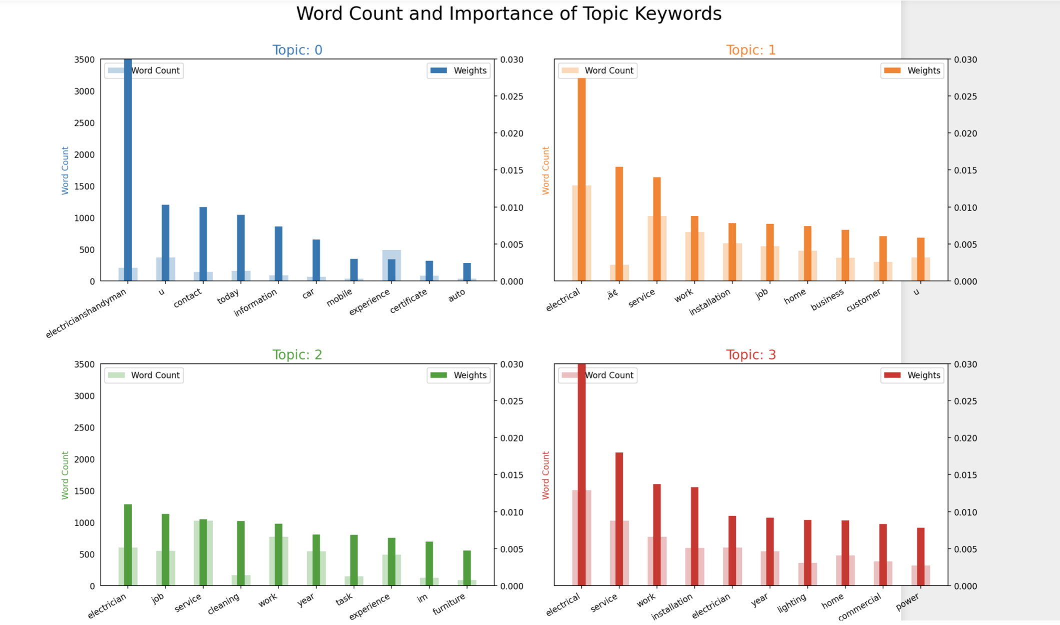


Figure 11: Distribution of word counts for all electricians

From these 4 topics, we could see that the words ’electrical’, ’electrician’ are the main keywords among these topics. Then the next important word based on word counts is the ’service’ which occupied a large weight in topic1,2,3. The last useful keyword extracted from distributions is the ‘installation’ which is the top5 keyword in both topic1 and topic 3. Thus, the keyword that I summarized from this distribution of word counts is ’electrical’, ’electrician’, ’service’, ’installation’.

4.2.1.2 Word Cloud Analysis

Figure 12 shows the Plotting of Word Cloud in LDA model.

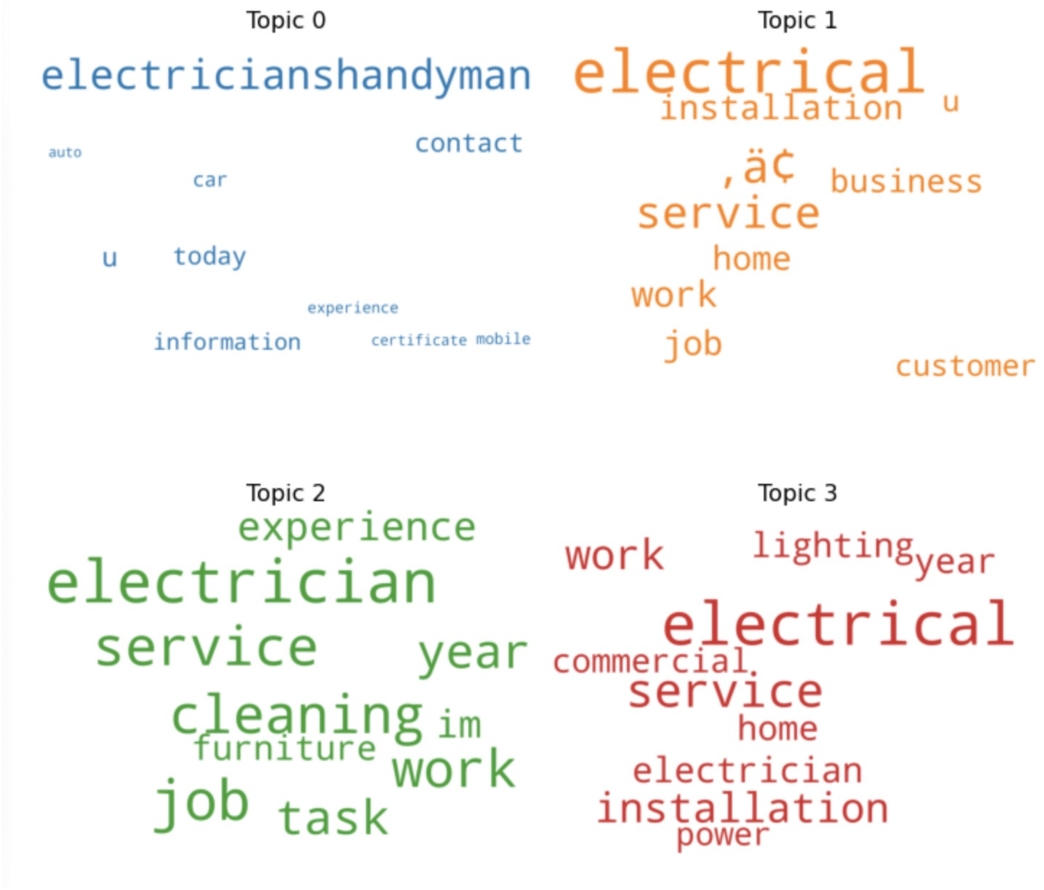


Figure 12: Word Cloud for all Electrician

From Figure 12, we can directly see that the most impressive words are 'electrical' and 'electrician', and 'service'. Moreover, we could summarize the keywords 'cleaning', 'furniture' in topic2 and 'installation', 'lighting' in topic3. The result that we obtained is the same as the figure above since both graphs are plotted based on the word counts in each sentence. Figure 13 shows the Tableau dashboard Word cloud

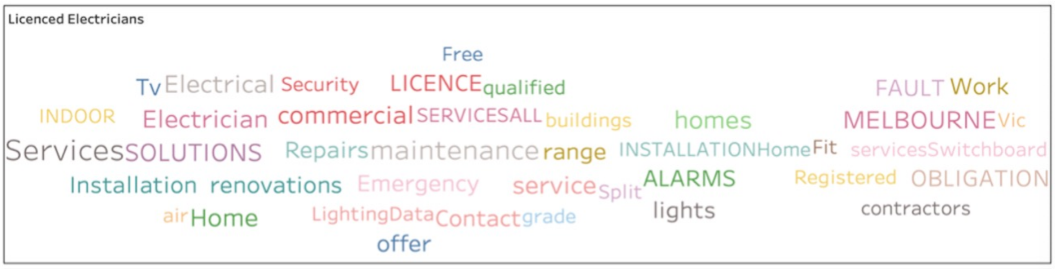


Figure 13: Tableau Word Cloud for all licensed Electrician

From the word cloud made by the tableau dashboard, we could summarize the keywords for licensed electricians are ’electrical’, ’maintenance’, ’installation’, ’light’, ’services’,’repairs’.

4.2.1.3 pyLDAvis Analysis

Based on the strong function of the python tool pyLDAvis, thus we could apply it for better summarizing and understanding these four dominant topics.

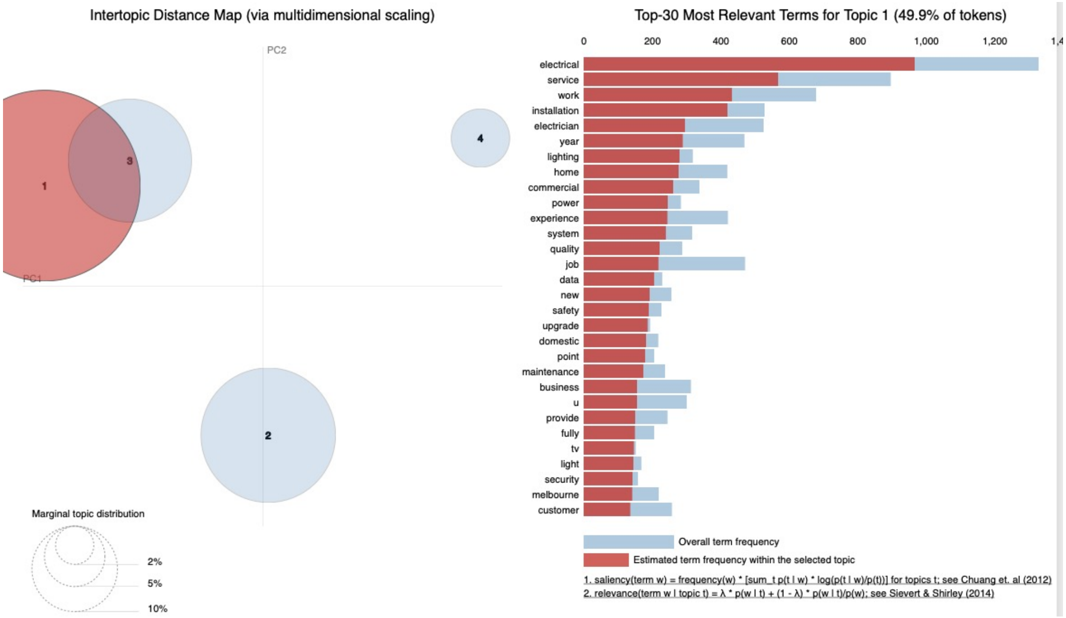


Figure 14: pyLDAvis for all Electricians

From the graph, number 1 largest circle represents topic 3 in previous dominant topics. Thus, from the right sidebar chart we could see that the 4 keywords we have summarized before are ’electrical’, ’electrician’, ’service’ and ’installation’ are in top5. Also, ’lighting’ could be a part of keywords although it does not occur in other topics, topic 3 has occupied half of all the texts, thus ‘lighting’ has a relatively huge weight during the whole texts.

The second circle represents topic 2. In this topic, we could see that it mainly focuses on the ’cleaning’, ’removal’, ’furniture’. These are not typical electrical work, thus here we find out these ‘electricians’ who only do the handymen works, most of them did not get a formal license, or basically, they are not electricians but have been filtered into electricians in websites. And this topic occupied 24.7%, which closed to result of the proportion of nonelectricians (464/1827 = 25.4%) shown in the table above. Since we have already filtered out these non electricians, thus we would analyze them in the next part.

Circle 3 and circle 4 represent topic 1 and topic 0 respectively. Topic 1 is quite similar to topic3 which has the same keywords, and topic 0 has a little occupation that we could neglect.

4.2.2 Visualization of LDA Model with Non Electrician

Here we consider the data set of non electricians extracted from all 4 websites.

4.2.2.1 Distribution of Word Counts

The Figure 15 presents the word counts and importance of keywords.

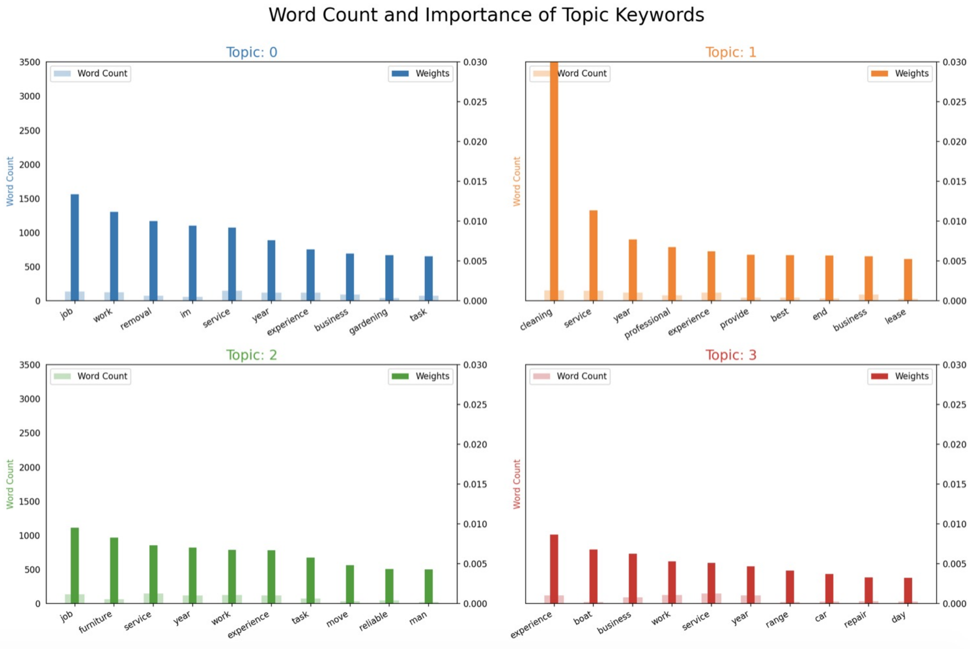


Figure 15: Distribution of Word Counts for Non Electrician

From these 4 graphs, in topic0, ignoring the meaningless words ’job’, ’work’, we could extract the word ’removal’ from the topic. In topic1, we choose the ’cleaning’ with the highest word counts. In topic2, we choose ’furniture’ as the keywords and in topic 3, only the word ’repair’ is useful. Thus, we summarized the keywords are ’removal’, ’cleaning’, ’furniture’, and ’repair’.

4.2.2.2 Word Cloud Analysis

The figure 16 shows the Word Cloud for non-electricians in LDA model.

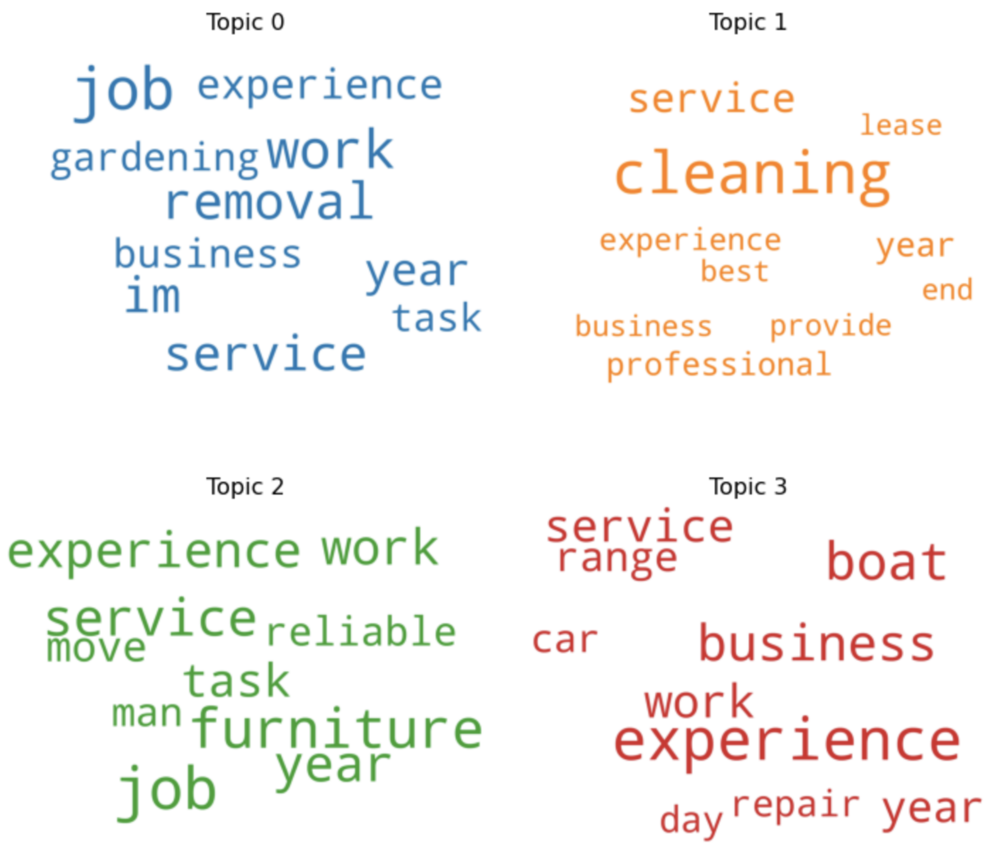


Figure 16: Word Cloud for Non Electricians

From the graph, the keyword which appeared in each topic is ’service’. For topic 0, the most impressive word is removal and topic 1 is cleaning. In topic 2, we could see ‘furniture is the biggest word and in topic 3 we could extract the useful keyword ’repair’. Thus, we summarize the keywords are ’service’, ’removal’, ’cleaning’, ’furniture’ and ’repair’ in these word clouds.

The figure 17 shows the Tableau Word Cloud for non-electricians.

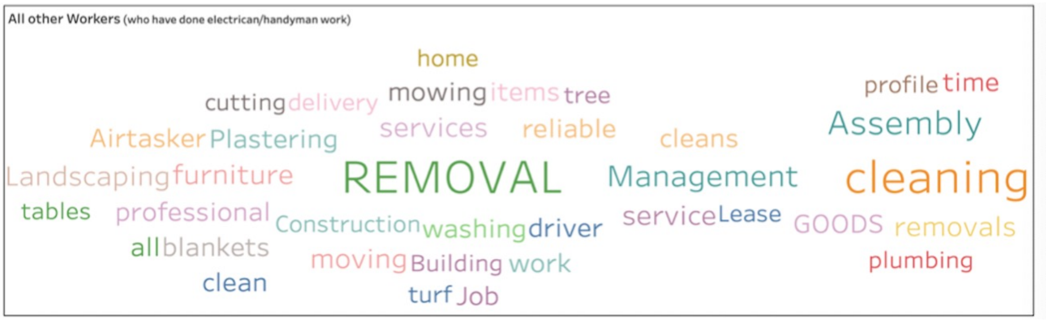


Figure 17: Tableau Word Cloud for Non Electricians

For the non electricians in tableau dashboard, the keywords are ’removal’,’cleaning’,’furniture’,’washing’.

4.2.2.3 pyLDAvis Analysis

The figure 18 shows pyLDAvis to the given non-electrician data set.

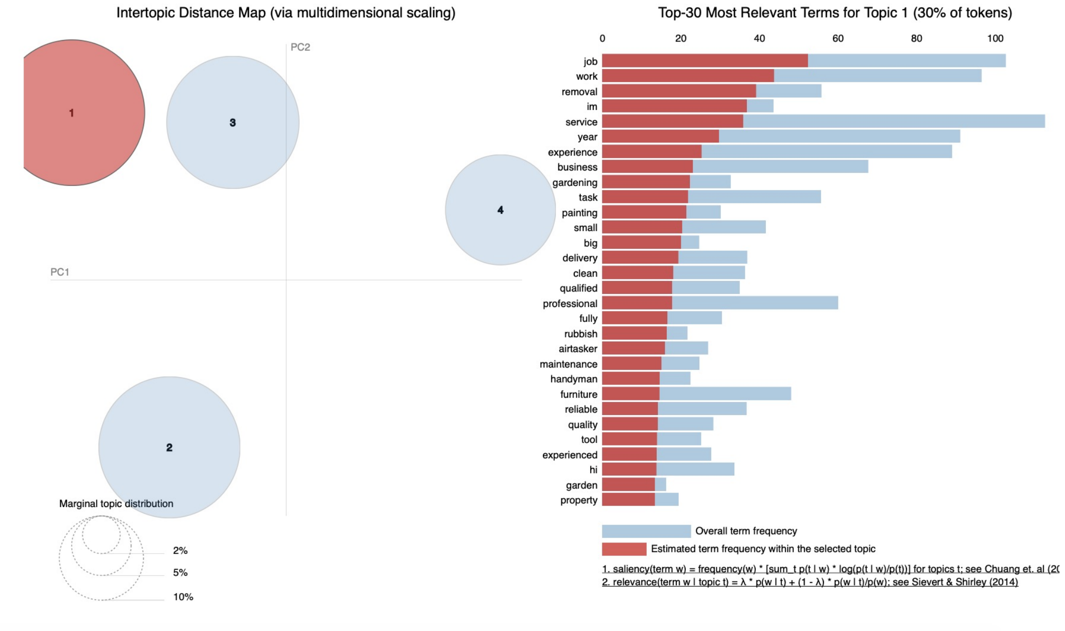

Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 1 (30% of tokens)
0
20
40
60
80
100
job
work
removal
im
service
year
experience
business
gardening
task
painting
small
big
delivery
clean
qualified
professional
fully
rubbish
airtasker
maintenance
handyman
furniture
reliable
quality
tool
experienced
hi
garden
property
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (20
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 18: pyLDAvis for Non-Electricians

From Figure 18, we can see that three circles 123 have a similar size about 28% of tokens. Thus, we would like to extract words from every 3 topics. Circle1 represents topic 0, circle2 represents topic1 and circle3 represents 3. Thus as the previous word counts showed and the right sidebar chart demonstrated, we conclude that the keywords as ’removal’, ’cleaning’, ’furniture’. The circle4 has only 17.1% tokens and the useful word keyword has less proportion in topic 3 thus we didn’t consider it as keywords during the pyLDAvis. Therefore, we get the keywords ’removal’, ’cleaning’, ’furniture’, from pyLDAvis analysis.

4.2.3 Conclusion with the Visualization of Topic Modelling Results

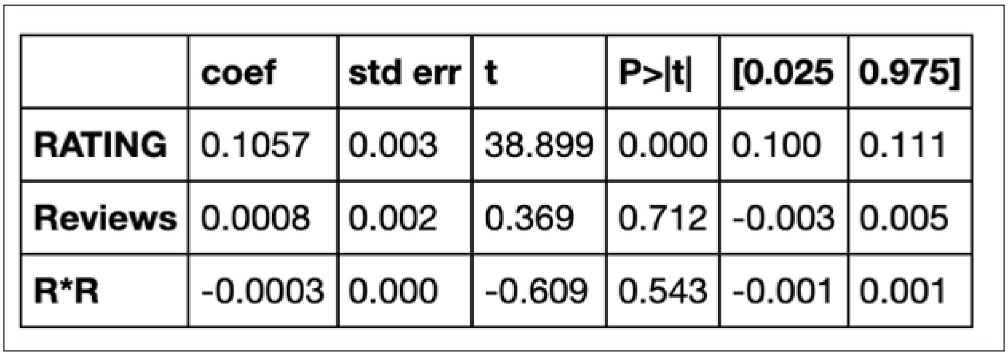
To conclude the results that we obtained above, for all electricians we crawled in 4 websites, we summarize the keywords as ’electrical’, ’electrician’, ’service’ [U+FF0C] ’ installation’, and ’lighting’. We also mention that a quarter of these data, in their keywords are some handymen work such as cleaning and repair. Thus, we analyze these parts of the data using the non electrician (who do not have an official license and not an electrician but have been filtered into the electrician category caused by their review and task contains words related to electrician.

From this part of the data, the keywords commonly used are ’removal’, ’cleaning’, ’furniture’. Thus from the analysis, we have given, most of the electricians(74.6%) doing the electrical work ‘installation’ lighting’ from their description and review, and few of them (25.4%)doing simply handymen work focuses on ’removal’, ’cleaning’, ’furniture’.

And here, based on the keywords in description analysis, we recommend the customer to query the true electricians with formal license and provide a license number, if they need some help in electrical service or device. Since the non electrician which has been filtered into electrician actually cannot do any jobs with the electrical work, it also is represented in their profile. If the requirement is a basic handymen job, it is free to choose, but for safety, we strongly recommend people to choose the licensed electrician with provided licensed number for query electrical work.

4.3 Correlations Between the License Status with the Ratings and Reviews

The Correlations test for Licensed status versus ratings and reviews are conducted to find out whether they have any relation with the Licensed status.



	coef	std err	t	P>|t|	[0.025	0.975]
RATING	0.1057	0.003	38.899	0.000	0.100	0.111
Reviews	0.0008	0.002	0.369	0.712	-0.003	0.005
R*R	-0.0003	0.000	-0.609	0.543	-0.001	0.001

Figure 19: Correlations Between the license Status with the Ratings and Reviews

From the summary of the test we can infer that, Since the p-value of the RATING is actually 0, we can reject the null hypothesis that the coefficient with RATING is 0. Hence, one can infer that license is linearly related with the rating, i.e. that licensed workers are highly rated.

On the other hand, since p-values of the reviews and the product of reviews and Ratings are higher, we accept the null hypothesis. This means that license does not linearly depend on the number of provided Reviews.

5 Challenges

As all projects go, we faced a number of significant challenges overall. Some are listed below:

1) Understanding the business problem: Working on the business problem statement is one of the key objectives of a Data Scientist. The most of the time, it is a team, not the Data Scientist, who comes up with it. We were unsure how to start when we first received this project. The client was more concerned about the outcome. To transform this into a Data Science problem, we had to put our brains together.

   The client just wanted us to scrape the data and compare the worker data and demographics. But we, as a team turned this into a Data Science problem by applying some modelling techniques and used visualizations methods to understand the data better and show our results in a proper matter.

2) Multiple Data Sources: The main task of this project was to scrape data from different websites. But all websites are designed in a different way. Instead of just running or making one code that would run for all the websites, we had to design a separate code for all the four websites. Some were easy, but websites like Service Seeking was very difficult to scrape due to its privacy reasons. Some required timers to scrape data.

3) Missing Data: There many data points that were missing after scrapping the data from multiple data sources which are essential for carrying out our analysis. In our case the license number and description of were missing for certain workers.

4) Not Enough Features: Although we could scrape approximately 1800 worker details, we couldn't get many features from all websites. Airtasker had a lot of features but Gumtree had not much features. Therefore, when we went to combine the data, we had to exclude many features because they were not common. This problem could have been solved if we could explore more websites to get more data. More data could have resulted us to conduct more tests and get better results.

5) Data Security: The clients were keen to get data from the other sources like Facebook marketplace and Instagram's pages, but due to the data protection and encryption meth-

ods followed by these websites, the scrapping methods we used couldn't succeed, thus limiting the number of data we can gather.

6 Conclusion & Future Work

We have successfully created an end-to-end project staring from data collection to visualization. We have successfully identified licensed and unlicensed electricians on all four platforms and have clearly distinguished them. Each set has a different set of analysis and results. License Numbers seem to be the most important aspect to determine whether the worker is validated or not. But according to our results, people do not seem to care about the license numbers at all, as long as the ratings are good.

Seeing the data on hand, 475 electricians can be promoted by ESV to update their license details. Looking at the results, we can see that HiPages have the most licensed electricians and Airtasker has the most workers who do all kinds of tasks. Looking at the demographics, we see that the HiPages workers are spread all over Victoria. This shows that this website has workers all over Victoria and its a trusted website. Also, this website only allows licensed electricians to work, which ensures the workers and the poster's safety.

One of the main tasks was to compare the licensed electricians with that of the client data. Most of the licenses were already present, the ones not present, ESV can register them and give them authentic licenses. The comparison data is as follows:

	License Numbers	
Website	Present	Not Present
HiPages	527	72
Airtasker	73	20
GumTree	26	5
Service Seeking	151	20

Table 6: Comparison with ESV Data

Apart from the 475 electricians, ESV can also update the 117 electrician's licenses into their database as these licenses are not present originally. ESV will use our results to promote more licensed workers in websites ensuring safety and security for all.

Coming to non-electricians, our results also show that cleaners and plumbers are the main tasks that also do small electrical chores. ESV should promote more workers to update their license details on Airtasker. Non-licensed workers are an issue for safety and security. Recently, an electrician on Airtasker was fined with $10,000 [21] as he posed to work as a qualified electrician. From our other analysis we also see that people actually do not care about the license status of the worker for their tasks. That is a wrong analogy, the websites should not allow unlicensed workers to work at all.

There are two future directions for this project: (1) explore other websites to gather more electrician data and monitor their license status. (2) Find a more efficient way to automate the scrapping of data so that a new code is not required for every website.

7 Team Contribution

The below table shows the contribution for group 9

Name	**Roles**
Chen Zhou	Task Analysis and Python Visualization
Meghna Panda	Web Scrapping, Data Processing and Tableau visualization
Vignesh Lakshminarayanan	Team Lead, Topic Modelling, Correlation analysis
Xinyu Mao	Web Crawling, Tableau Visualization

References

[1] S. Costello, "Airtasker admits guidelines were breached by sparky's botched job," Apr 2021. [Online]. Available: https://9now.nine.com.au/a-current-affair/airtasker-admits-fault-after-aussie-unqualified-sparkys-botched-job/c4448fbb-a601-4327-917d-264681bc79c2

[2] A. Robertson, “About esv.” [Online]. Available: https://esv.vic.gov.au/about-esv/corporate-information/about-esv/

[3] R. Anitra, “Electrician’s license (a/a/e).” [Online]. Available: https://esv.vic.gov.au/licensing-coes/electrical-licences/electricians-licence/

[4] D. M. Thomas and S. Mathur, “Data analysis by web scraping using python,” in *2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)*. IEEE, 2019, pp. 450–454.

[5] M. E. Asikri, S. Krit, and H. Chaib, “Using web scraping in a knowledge environment to build ontologies using python and scrapy,” *European Journal of Molecular & Clinical Medicine*, vol. 7, no. 3, pp. 433–442, 2020.

[6] A. V. Saurkar, K. G. Pathare, and S. A. Gode, “An overview on web scraping techniques and tools,” *International Journal on Future Revolution in Computer Science & Communication Engineering*, vol. 4, no. 4, pp. 363–367, 2018.

[7] C. Steven, “Web scraping wikipedia using python and beautifulsoup,” 11 2019.

[8] J. Boyd-Graber, D. Mimno, and D. Newman, “Care and feeding of topic models: Problems, diagnostics, and improvements,” *Handbook of mixed membership models and their applications*, vol. 225255, 2014.

[9] T. Hofmann, “Probabilistic latent semantic analysis,” *arXiv preprint arXiv:1301.6705*, 2013.

[10] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh, “On smoothing and inference for topic models,” *arXiv preprint arXiv:1205.2662*, 2012.

[11] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, “Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey,” *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15 169–15 211, 2019.

[12] “Airtasker about.” [Online]. Available: https://www.airtasker.com/about/

[13] “Gumtree about.” [Online]. Available: https://www.gumtree.com/info/life/about-us/

[14] "Basics - what is gumtree?" [Online]. Available: https://help.gumtree.com.au/AU/articles/en_US/KB_Article/what-is-Gumtree-AU?c=PKB%3ABasics&vcategory2=What_is_Gumtree&vgroup1=PKB&ab=20

[15] M. Mason, "Site drives a turnabout in job seeking," Sep 2012. [Online]. Available: https://www.smh.com.au/business/small-business/site-drives-a-turnabout-in-job-seeking-20120902-258ck.html

[16] S. Palmer-Derrien, L. Maskiell, and SmartCompany, "Hipages becomes latest aussie startup to ipo on the asx," Nov 2020. [Online]. Available: https://www.smartcompany.com.au/startupsmart/news/hipages-ipo-asx/

[17] Airtasker-profile. [Online]. Available: https://www.airtasker.com/users/lachlan-b-4132247/

[18] Airtasker-profile. [Online]. Available: https://www.airtasker.com/users/alec-t-18429333/

[19] Airtasker-profile. [Online]. Available: https://www.airtasker.com/users/muhammad-omar-m-9221041/

[20] Wikipedia contributors, "Latent dirichlet allocation — Wikipedia, the free encyclopedia," 2021, [Online; accessed 27-October-2021]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Latent_Dirichlet_allocation&oldid=1049678663

[21] Airtasker-article. [Online]. Available: https://www.worksafe.qld.gov.au/news-and-events/newsletters/esafe-newsletters/esafe-editions/esafe-electrical/2019-bulletins/$100,000-fine-for-airtasker-fake-electrician/

8 Appendix

All of our work has been updated into a Git Repository. The link is attached below:
https://github.com/lakshminaray/Capstone-Project-.git
The meeting logs can be accessed via GitHub **Meeting mins** folder.