Analysis of semantic search results

In this notebook we will analyze the results from the information retrieval task from profile descriptions of nannies obtained from the site CareGuide: CanadianNanny.ca [1]. So far we have retrieved around 9k+ distinct profile descriptions from all over Canada. The goal is to extract demographic information from the text such as age, gender, nationality and immigration status, and analyze these in order to find systemic inequalities and discrimination. These variables were obtained using information retrieval techniques with two different approaches: a lexical keyword-base model, such as BM25, and semantic search with transformer-generated sentence embeddings.

Data

The profiles were scraped using an R script at multiple points of July, August and September. They were retrieved in order of appearance on the site which is sorted by the last time the user was active on the platform. So far more than 16,000 entries have been scraped from the site, of those around 9,700 are distinct users. To run the information retrieval tasks, duplicates were dropped keeping the latest profile update.

Demographic variables are usually reported by the user in the text-based sections of their profiles, which include the short blurb right under the profile picture, a "Reasons to Hire Me" section, and a longer description in the "About Me" section. We also retrieved the users' name, url, location, reported years of experience, hourly rate, last time active on the site, number of reviews, star rating out of five, bullet points under the "I can work:" subsection (part-time, full-time, summer, etc.), children ages the user has experience with (infant, toddler, newborn, etc.), number of children they can look after, experience with children with medical conditions (diabetes, disability, epilepsy, severe allergies, etc.), transportation requirements (close to transit, has driver's license, etc.), qualifications (first aid, CPR, languages, etc.), and services they can provide (housekeeping, cooking, groceries, swim supervision, etc.).

Table 1: Province statistics

Province/Territory	N (ratio)	Rate (sd)	Years Ex. (sd)	No. Children (sd)	Ratings (No. rev.)
Alberta	1674 (0.172)	19.454 (4.55)	8.376 (6.46)	3.029 (2.06)	3.769 (15)
British Columbia	1592 (0.164)	20.982 (4.18)	7.598 (6.18)	2.777 (1.11)	2.6 (6)
Manitoba	232 (0.024)	16.793 (3.53)	6.703 (5.44)	2.781 (0.94)	NaN (0)
New Brunswick	98 (0.01)	16.443 (2.78)	8.724 (6.96)	3.308 (2.36)	NaN (0)
Newfoundland and Labrador	49 (0.005)	17.49 (2.97)	6.25 (6.74)	3.125 (0.83)	NaN (0)
Northwest Territories	2 (0)	21.1 (8.34)	NaN (NA)	NaN (NA)	NaN (0)
Nova Scotia	198 (0.02)	18.438 (3.49)	9.132 (7.32)	4.2 (4.45)	NaN (0)
Ontario	5085 (0.522)	19.66 (4.86)	8.192 (6.89)	2.986 (2.71)	3.86 (47)
Prince Edward Island	15 (0.002)	16.008 (2.11)	8.6 (8.08)	5 (NA)	NaN (0)
Quebec	466 (0.048)	17.927 (3.78)	7.051 (6.06)	2.703 (1.27)	3 (2)
Saskatchewan	309 (0.032)	16.323 (3.29)	7.193 (6.97)	3.833 (2.37)	NaN (0)
Yukon	11 (0.001)	22.4 (3.37)	7.25 (3.3)	3 (NA)	NaN (0)
National	9733 (1)	19.516 (4.65)	8.017 (6.64)	2.983 (2.35)	3.714 (70)

Exploratory Data Analysis

Table 1 shows average values (expect the first column) by province of specific characteristics reported by users, rates over \$200 CAD hourly were removed, since these are probably a mistake. Ontario is by far the province with the most users with over 50% of these located there, followed by Alberta and British Columbia, which are both well over 1k users and above 15% of the total. At the national we have a mean hourly rate of \$19.51 CAD with a standard deviation of \$4.65, the province with highest rate in average was British Columbia with \$20.98, followed by Ontario at \$19.66; however, both Yukon and Northwest Territories have higher averages rates than any province. The province with the lowest average rate was Prince Edward Island with \$16.01. With regards to the years of experience the national average was around eight, the province with the most experienced nannies was Nova Scotia, while the least experienced was Newfoundland. Only 66 profiles received reviews from clients, and three of those profiles have been deleted since the time they were first retrieved. In total 70 reviews have been filled (counting profiles that received more than one), with a national average rating of 3.71. Users from Ontario are by far the most likely to receive a review and they have an average rating of 3.86, while users from BC had a very low average rating of 2.6.

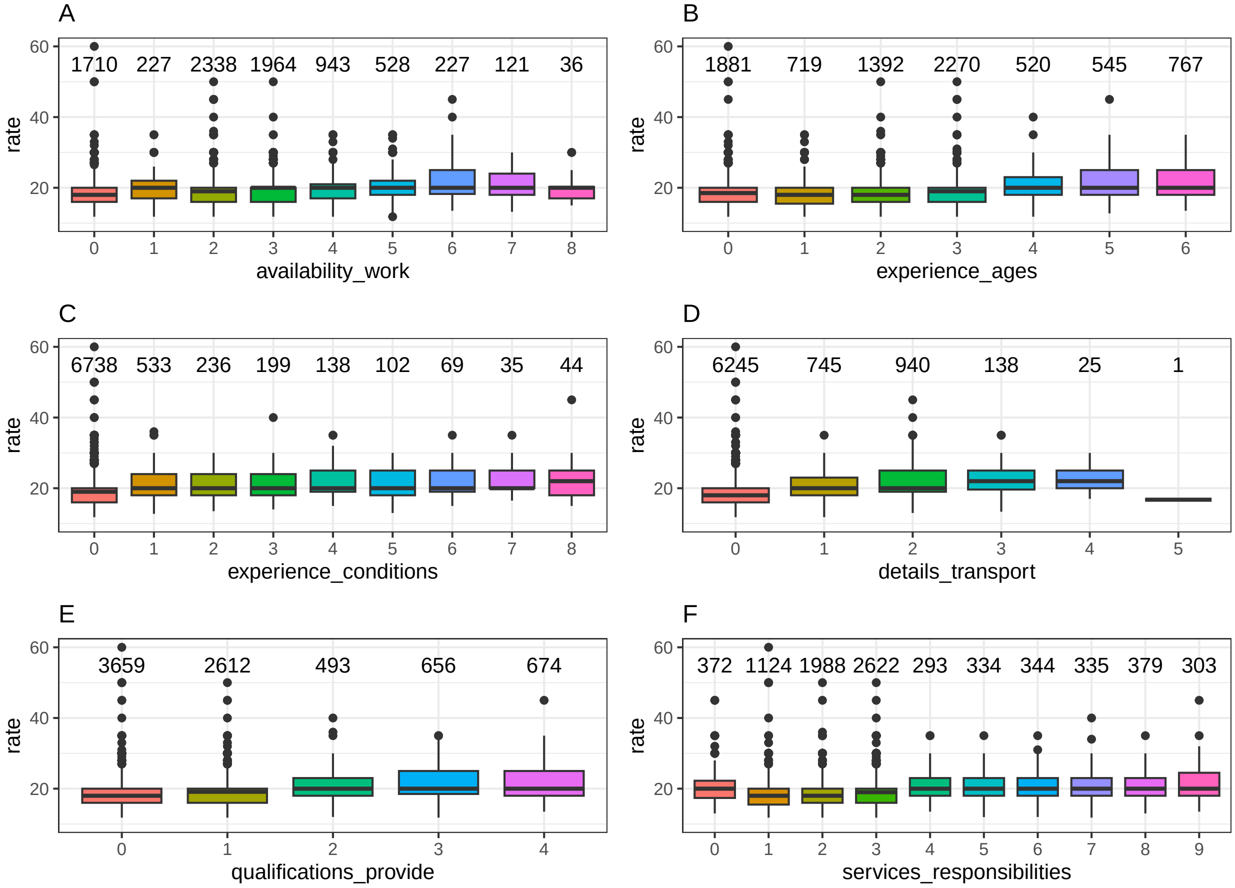

Figure 1: Distribution of requested hourly rates by number of items listed under different subsections.

As mentioned before, each user describes on their profiles the services they are willing or qualified to provide. One thing worth exploring is how the requested hourly rate varies depending on the number of qualifications the nannies listed. Figure 1 shows the distribution of hourly rates by the number of items users wrote under those subsections. Above each boxplot, the number of users per category is shown. We can see that, except for the subsection of “responsibilities” under “Services,” those who didn’t have items listed under that subsection had the lowest median rate. However, the difference between distributions isn’t significant enough to draw any conclusions just yet. Figure 2 shows the difference in rate distribution for users who reported being qualified to provide Cardiopulmonary resuscitation (CPR). We can see that those who can provide CPR tend to request a higher hourly rate on average.

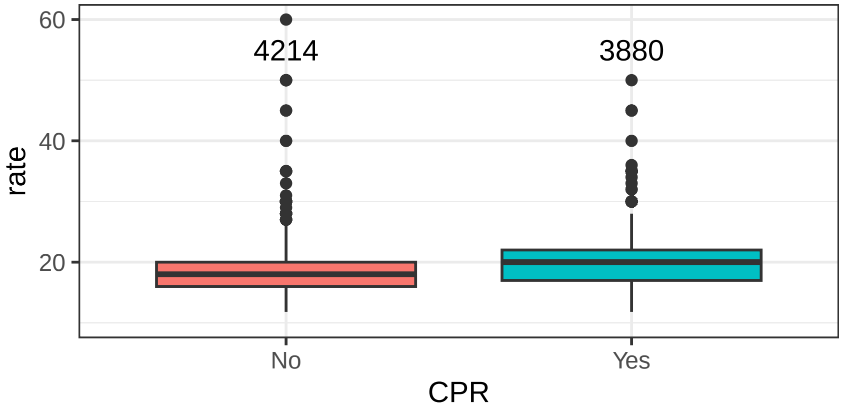

Figure 2: Distribution of hourly rates based on whether the user can provide CPR.

Exploratory Text Analysis

The profile descriptions of the users were highly similar to one another; all of them talked overwhelmingly positively about themselves, which isn't surprising since they are trying to establish trust with their possible employers. In a more in-depth text analysis, it was found that the users who had a negative sentiment score wrote a very short description and included phrases such as "don't hesitate to. . . " and most lexicons classify the word "hesitate" as negative. In Figure 3, we can see the most common words that appeared in the "About Me" section of the profiles. Unsurprisingly, the most common words were "children," "care," and "nanny."

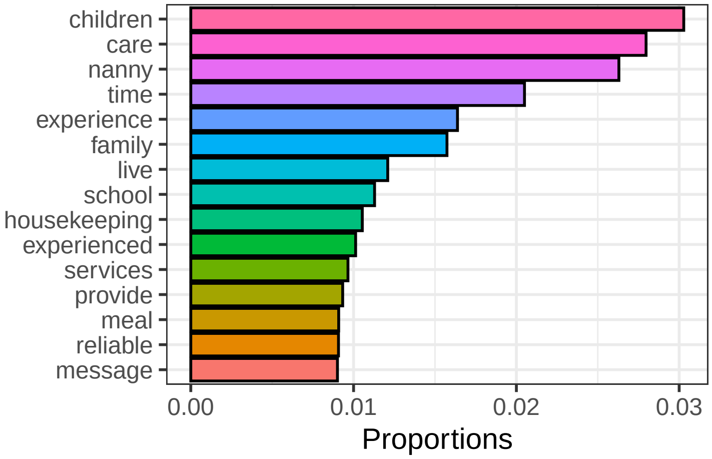

Figure 3: Proportion of appearance of words in the profile description (removing stop words).

We are also interested in how words are related to each other; that is, what are the most common words that come before a specific word. In Figure 4, we can see an arrangement of words in a network, or "graph." This was done using the igraph and ggraph libraries [2-4]. Some of the most common bigram examples are "qualified/passionate/professional nanny," "pet care," "meal preparation," and "primary school." We can also see that the network detects the city and province the nannies are based in when they mention it in their descriptions, as well as, their hourly rate.

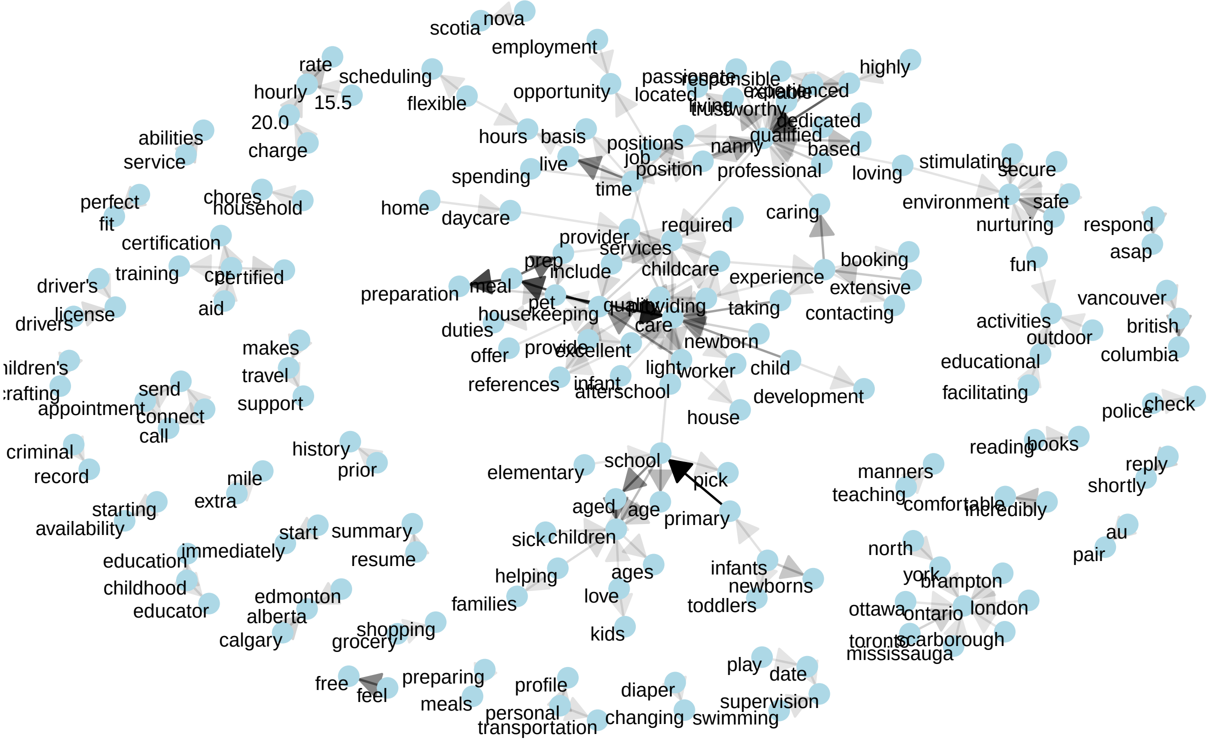

Figure 4: Graph representation of bigrams in the profile descriptions (removing stop words).

Information Retrieval Approaches

Information Retrieval (IR) is the process of obtaining any type of media based on user information needs. The resulting IR system is often called a search engine [5]. The IR task that we consider returns a list of ordered sentences, taken from profile descriptions, based on a query. The IR architecture uses the vector representations of queries and sentences, which are then ordered based on a similarity function like cosine or dot product.

BM25Okapi

BM25 stands for "Best Matching 25" (i.e the 25th iteration of the function), and Okapi makes reference to the first IR system that used it. This is a bag-of-words model in which query and document vectors are based on unigram word counts, and each word is consider independently of its position. This function is similar to TF-IDF (term frequency - inverse document frequency) weighting, but it adds two parameters: k, which adjusts the balance between term frequency and IDF, and b, which controls the importance of document length normalization [5]. The BM25 score of a query Q containing keywords $q_1, ..., q_n$ and a document D is:

$$\text{score}(D, Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \times \frac{tf(q_i, D) \ \cdot \ (k+1)}{tf(q_1, D) + k \times \left(1 - b + \frac{b\ |D|}{|D|_{\text{avg}}}\right)}$$

Where $tf(q_i, D)$ is the number of times the keyword q_i appears on document D, $|D|$ is the length of the document and $|D|_{\text{avg}}$ is the average document length. The inverse document frequency $\text{IDF}(q_i)$ is often computed as:

$$\mathrm{IDF}(q_i) = \log\left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1\right)$$

Where N is the total number of documents and $n(q_i)$ is the number of documents containing the keyword q_i. To implement this method we are using the python package `rank_bm25` which sets $k = 1.5$ and $b = 0.75$ by default [6]. It also sets keywords with negative IDF to the average of the non-negative IDFs multiply by 0.25. In various IR tasks is common to remove "stop words" from the text, because these words carry little semantic knowledge and they often hurt the process. However, it's not strictly necessary since the IDF downweights these stop words.

Table 2: Users that requiere visa sponsorship retrieved by IR models.

Model	Query	Threshold	Total Users Retrieved	True Positives
BM25	Do you need sponsorship?	5.0	248	84
	immigration visa sponsor sponsorship	0.0	213	183
multi-qa-mpnet	Do you need sponsorship?	0.4	198	87
	immigration visa sponsor sponsorship	0.4	97	84

The fatal flaw of the BM25 scoring function is that the query's keywords have to appear in the documents that it's searching through. So if the author used a synonym or expressed it in another way the model has no way of knowing that that is a relevant document. An additional disadvantage is that it doesn't take into account how close the keywords are together in the document, in our application we are dividing the profile descriptions by sentence, so the proximity between words shouldn't be an important issue.

Sentence Embeddings

Our second approach was to compute vector representations of sentences, or embeddings using transformer models. We used the `SentenceTransformers` package in python, which was developed for semantic search tasks just like this one, for a full explanation of the models used in this package we refer the users to the original Sentence-BERT paper [7]. Basically these models work by "masking" a word from the sentence and using the context both before and after this word (bidirectional) to predict it. Given that BERT (Bidirectional Encoder Representations from Transformers) is quite large and therefore computationally expensive, we didn't use it to encode our sentences. In contrast, we mainly used the `multi-qa-mpnet-base` bi-encoder model, which was specifically trained on 215 million question-answer pairs from sources such as yahoo answers, StackExchange and Google and Bing search queries, making it a perfect match for our use case. Once the embeddings of both the query and sentences are computed we can use the cosine or dot product similarity score to ranked them.

The main advantage of this approach is that we don't need to have the query keywords appear on the sentences, the model should retrieve sentences that don't contain any of the words but have synonyms or similar ones. This could be both an advantage and disadvantage, however, because the model might select certain sentences that although are similar, aren't really relevant for that specific query. For example, when feeding the query "Do you need visa sponsorship?", the model would selected sentences such as: "I have a driver's license", one can claim that a visa and a driver's license are similar but the latter isn't really relevant for our goal.

Information retrieved

We were interested in retrieving demographic variables of the user, like age and nationality, as well as, their immigration status in Canada and if they require sponsorship. All these variables were retrieved using both approaches described above. For all of these the top 1k sentences were obtained then select the one with the maximum score per distinct profile. For the age variable, we used regular expression to extract numbers within the text, then we removed those younger than 14 and older than 100 years and those with reported years of experience + 10 higher than their age, since these are most likely mistakes. For the countries we used a list of country names and their adjectival and demonymic forms retrieved from wikipedia [8]. For the sponsorship arbitrary score thresholds were selected, then manually label the resulting sentences.

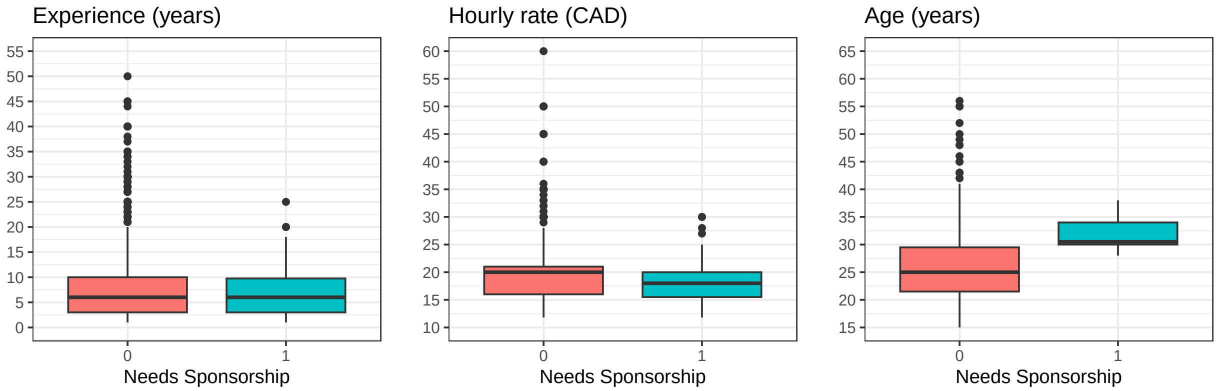

Figure 5: Distribution of years of experience and hourly rate of users based on visa requierements.

Results

Immigration status

All three variables were extracted with both approaches and two different queries: a keyword-based query and a question. With regards to retrieving users who needed sponsorship table 2 shows the queries and thresholds used, as well as, the number of users and true positives (users who actually required sponsorship as reviewed manually afterwards) retrieved. The BM25 model, with the keyword-based query, retrieved the most true positives from the texts; even though, the bi-encoder model had a better ratio of true positives, it's unclear if this holds when reducing the threshold to allow the retrieval of more users. Therefore, we consider that the BM25 model is the most appropriate for this specific task.

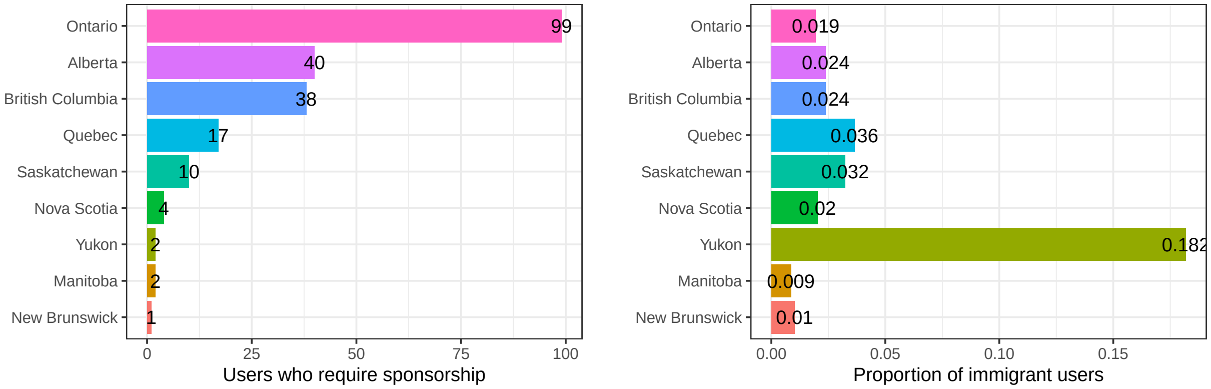

Figure 6: Number and proportion of users who require sponsorship by province or territory.

As we can see in figure 5 users who need to be sponsored are willing to charge a slightly lower rate than those who doesn't. The distribution of reported years of experience of nannies who require sponsorship is much less spread out than those who doesn't. Figure 6 shows the number and proportion of users requesting sponsorship by province, unsurprisingly Ontario has the most foreign users, however, the Yukon territory has by far the highest proportion, where 2 out of 11 users where found to be looking for sponsorship.

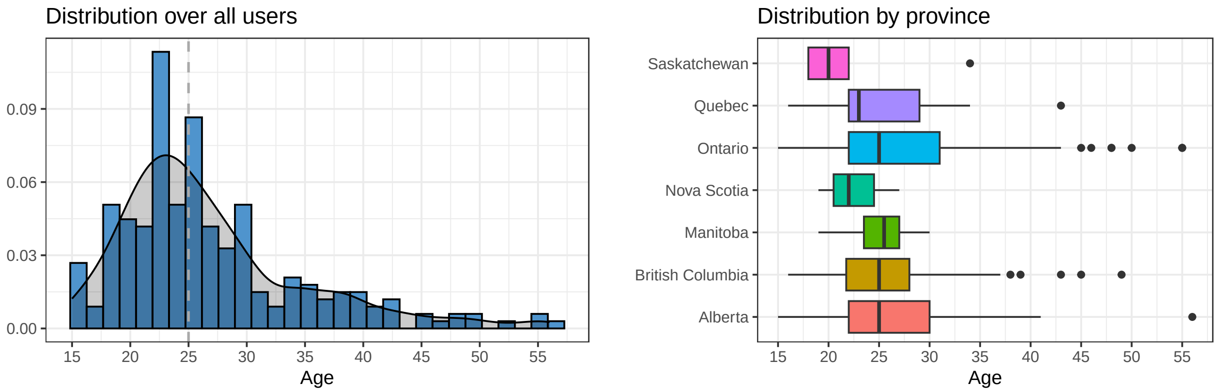

Figure 7: Age distribution of users and by province (selecting those with at least 2 users with valid ages).

Users' age

For this task we used the queries: "How old are you?" and "I am years old". There was no need to select a threshold when retrieving users' age, since out of the top sentences we just extract the numbers and performed the subsequent transformations mentioned before. The BM25 with the keyword-based query was able to retrieve 237 users with apparent valid ages and 223 with the question based one, while the bi-encoder performed better with the question query it only manage to extract 195 users with valid ages. Important to note that the entries haven't been double checked manually.

Figure 7 shows the distribution of age over all users and by province. The median over the whole population is 25, and the provinces' median ranges from 20 to 25 years old. It's clear that the distributions of the total users and for each provinces are right-skewed. There were only 6 users requesting to be sponsored with valid ages, in figure 5 we can see that there's a substantial difference between the ages of this group and locals. While users who are able to work legally in Canada have a median age very similar to the total at 25 years, nannies who want to immigrate to Canada are on average 32 years old; however, more data is needed to claim that there's a significant difference.

Nationalities

The two queries we tested with in order to retrieve users' country or nationality were: "I am from" and "Where are you from?". However, almost every word in these queries are included in the stop words lexicon that we are using before feeding the queries into the BM25 model, so it's not able to retrieve any valid sentence (future iterations will not remove stop words from the text so a proper comparison will be made). The transformer model is able to retrieve 243 users with valid nationalities using the keyword-based query.

The majority of foreign users (78) are originally from the Philippines, followed by Mexico with 21 users. Figure 8 shows the differences in rate distribution and reported years of experience by country of origin. Nannies from the Philippines are willing to take a substantially lower pay than others, with a median hourly rate of less than $16 CAD and some users even going down to $12 per hour. Users from Mexico, Japan, Colombia or Brazil have an hourly rate much closer to the global median at $20 an hour.

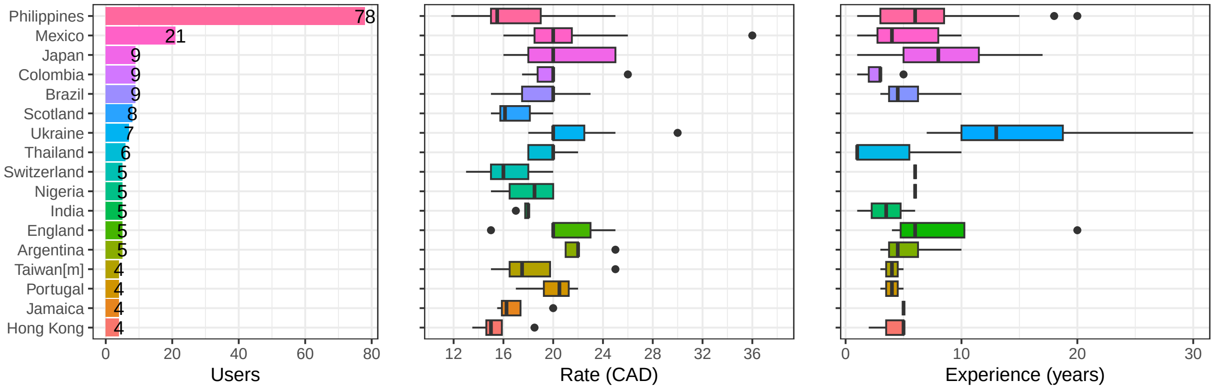

Figure 8: Age distribution of users and by province (selecting those with at least 2 users with valid ages).

Conclusion

In this report we analyzed profile descriptions of nannies across Canada and extracted demographic variables from the text descriptions using information retrieval techniques with two different approaches: a lexical keyword-based algorithm (BM25) and calculating sentence embeddings with transformer models. Both approaches were able to extract meaningful insights from the data, but the BM25 function was able to retrieve more information with respect to the immigration status and age variables. This shows that deterministic models are still highly valuable in certain IR tasks.

We were able to extract 183 users, using BM25 scoring, not authorized to work in Canada who are asking for visa sponsorship, we didn't take into account immigrants who mentioned having a work visa, we also didn't make a distinction between users asking for sponsorship from outside or inside the country. It's clear that these nannies are willing to charge a lower hourly rate than those with a regular immigration status and their age was significantly higher than the average age of the local nannies. There weren't substantial differences with regards to their experience or qualifications. Around 230 users mentioned being born in a foreign country, the vast majority of these were from the Philippines, which is in accordance with current immigration trends in Canada. In future iterations of the project we would like to look at analyze all immigrants as a whole and compare them to the local workforce, as well as, analyzing users by gender and by their profile picture with image processing techniques.

References

[1] CareGuide: CanadianNanny.ca. Retrieved July 3, 2023 from: https://canadiannanny.ca/nannies/ontario.

[2] Silge, J., &; Robinson, D. (2017). Text mining with R: A tidy approach. O'Reilly.

[3] Pedersen T (2022). ggraph: An Implementation of Grammar of Graphics for Graphs and Networks. https://ggraph.data-imaginist.com, https://github.com/thomasp85/ggraph.

[4] Csardi G, Nepusz T (2006). "The igraph software package for complex network research." InterJournal, Complex Systems, 1695. https://igraph.org.

[5] Jurafsky, D., & Martin, J. H. (2022). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Pearson.

[6] Dorian Brown. (2020). Rank-BM25: A Collection of BM25 Algorithms in Python.

[7] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

[8] List of adjectival and demonymic forms for countries and nations. Wikipedia. Retrieved September 8th, 2023 from: https://en.wikipedia.org/wiki/List_of_adjectival_and_demonymic_forms_for_countries_and_nations.