Abstract

The occupational structure describes the distribution of a nation’s population in different occupations. The Standard Occupational Classification system (SOC) 2018 organizes occupations into a hierarchy of groups and subgroups. Each occupational group includes a list of illustrative job titles. However, the job titles are too specific to be used directly for information retrieval from free text. Our work focuses on expanding the SOC taxonomy and finds its mentions in entertainment media (text dialogue) data to understand their portrayal in media. We develop different synonym expansion algorithms to expand the SOC dictionary and evaluate them on movie and TV show subtitles. We find the incidence of different job titles in media content, and examine the correlation between its distribution in media and the occupational structure of the country over time. Our work is the first of its kind to perform a large scale computational study of the lexical representation of professions in Hollywood movie content. We release the enlarged SOC profession taxonomy for further research.

1. Introduction

The occupation or job of an individual defines one’s role in society with respect to the service they provide. Different occupations require different types and levels of skill which is acquired through education and specialized training. It enables the individual to contribute to society and also care for the family in exchange for their work. A profession is any work performed legally for pay or profit. Over time, different kinds of professions have emerged and gained prominence, whereas others have declined. These changes reflect the technological advancement and business innovations in the industry [Dicesare, 1975]. E.g. the eighteenth and early nineteenth century witnessed a decline in agricultural and manual labor jobs due to the industrial revolution and influx of automated machinery. Managerial and white-collar professions took precedence over physical jobs in the 1960s and 1970s [Elvery, 2019]. The distribution of human resources in different occupations is called the occupational structure. The trends of different professions in the occupational structure provide valuable information to business leaders, policymakers and job seekers to understand
the labor market and make decisions. In the U.S., the Bureau of Labor Statistics, which is a unit of the Department of Labor, is responsible for measuring the occupational structure of the country [Goldberg and Moye, 1985]. As part of this effort, they publish several surveys that measure the employment [CES, OES], earnings [CPS] and pricing index [CPI] numbers of the country’s economy. These surveys provide quantitative evidence regarding the occupations that are gaining popularity and the ones that are waning. Is the same trend also reflected in contemporary media? Movies are a reflection of society, both present and past. They are a form of communication of stories whose ideas come from society. Does this mean that they also follow the trend of the occupational structure of society? Are film characters cast more in tertiary business jobs than as farmers? These questions motivate us to understand the portrayal of professions in the media. We want to understand if there exists any relationship between the real world trend in occupations and its representation on the big screen.

Movies have become a powerful ubiquity in modern society. The average person spends 3-4 hours per day watching TV [Rideout, 2016]. As such, they play an important role in shaping public perspective on different issues. It is important to be mindful of stereotypical representation of different occupations shown on film. E.g. Asimow [1999] and Asimow [2000] showed how lawyers are usually presented unfavorably in movies, especially if they are associated with big firms. Dimnik and Felton [2006] studied the portrayal of accountants along five dimensions of stereotypes - eccentric, plodder, dreamer, hero, and villain. Stone and Lee [1990] showed that journalists appearing in prime time, were frequently white males and positively portrayed as articulate, brave and competent. Physicians instead were frequently depicted as greedy, egotistical and uncaring [Flores, 2002]. All such studies have painstakingly examined each movie character cast in the profession of interest. Therefore, it is infeasible to scale this method computationally over a large set of movies. We instead focus on understanding the portrayal of professions in terms of its incidence in popular media. Since character level occupational codings are not available, we use word frequency of the occupation’s job titles as a proxy to represent its prevalence in film. Word frequencies have been previously used to study the trends in education [Moskovkin et al., 2019], culture [Younes and Reips, 2018] and language [Basile et al., 2016]. Brysbaert and New [2009] also argued that frequencies based on television and film subtitles are better than frequencies based on written sources, especially for psycholinguistic research. New et al. [2007] used movie subtitles to approximate word frequencies in human interaction.

The goal of our project is to understand the representation of professions in popular media in terms of the frequency of its job titles. We use the Standard Occupational Classification system [SOC] to get a taxonomy of job titles. However, the taxonomy is not searchable because the job titles are too detailed to appear in movie subtitles. Therefore, we use synonym expansion strategies to create a searchable taxonomy and augment the list of job titles. We use movie subtitles of OpenSubtitles corpus [Lison et al., 2018] to calculate the frequencies of different occupations. We find its trend and compare it with the employment data of the profession obtained from the Occupational Employment Statistics survey [OES]. We have made the taxonomy and related code publicly available 1.

1. Link shall be placed here upon publication
The rest of the paper is organized as follows - section 2 reviews some past work done in the domain of synonym expansion. Section 3 describes the SOC taxonomy and the OpenSubtitles corpus. Section 4 describes our synonym expansion algorithm and section 5 presents the results of its evaluation. We conclude with the trend analysis in section 6 and discussion of future work in section 7.

2. Related Work

Our motive is to create a searchable taxonomy for professions to study its representation in popular media. Synonym expansion has been applied to expand word dictionaries in previous works. Synonymy is a semantic relationship between words with very similar meanings. WordNet [Miller, 1995] is a lexical database for English, where words are grouped into sets of synonyms called synsets. A word can belong to multiple synsets because it can have multiple meanings. This is called polysemy. E.g. the word *cook* belongs to the synsets *cook.n.01* and *cook.v.01*. The first synset refers to someone who cooks, and the second synset refers to the act of cooking. The letters n (noun) and v (verb) indicate the part of speech tag, and the number indicates how commonly used the synsets are for the given word. Synsets are connected by super and subordinate relations, also called hyperonymy and hyponymy respectively. For e.g. *chef.n.01* is a hyponym of *cook.n.01*, and *skilled_worker.n.01* is a hypernym of *cook.n.01*. Synonym expansion means to find synonyms of a word or set of words, often given some context. For e.g. in the sentence - *The doctor is treating the patient*, a possible synonym of *doctor* is *physician*. Replacing a word with its synonym, while preserving the overall meaning of the sentence, is called lexical substitution. Both synonym expansion and lexical substitution rely on correctly identifying the meaning of the word in its given context. This is called word sense disambiguation, and it plays a vital role in many natural language processing pipelines.

We highlight some related work in these domains. Gong et al. [2005] found synonyms for query expansion. They used the synonym sets of WordNet to expand the terms of a web query and improved the precision and recall of retrieved documents. WordNet synsets have been used to find synonyms of geographical entities [Buscaldi et al., 2005], multi-term expressions [Gong et al., 2006] and general concepts [Zhang et al., 2009] for information retrieval. Sinha and Mihalcea [2009] used WordNet and word sense disambiguation strategies to find candidate synonyms of a word in a sentence, and filtered the list of candidates using Google ngrams for a lexical substitution task in Semeval 2007. They later used a graph based similarity measure to solve the same problem [Sinha and Mihalcea, 2011]. Abdalgader and Skabar [2010] used the union of the synsets of the words of two sentences to develop a semantic sentence similarity measure. Zeng et al. [2012] connected word synonyms in medical texts to their UMLS concepts.

We employ similar WordNet based synonym expansion algorithms to extend the SOC taxonomy. The technical term for such predefined entity lists is gazetteer. Gazetteers are popular in unsupervised named entity recognition tasks, where we do not have enough labeled data to train a supervised model. Nadeau et al. [2006] constructed gazetteers from seed words and heuristically resolved ambiguities between potential named entity classes of a given mention using rule based patterns. WordNet noun hierarchies were used to automate gazetteer construction by Toral and Munoz [2006]. Boteanu et al. [2018]
expanded a shopping taxonomy for efficient product search by matching the product names
to WordNet synsets. They chose the best sense of a product name according to the cosine
similarity of ConceptNet word embeddings [Speer et al., 2017], and took the cartesian
product of concept pairs to generate candidate synonyms. They filtered the list of candidates
using contextual features calculated from the taxonomy structure. Our synonym expansion
algorithm is partly inspired by their approach.

3. Data

Our work studies the representation of professions in media by creating a searchable taxon-
omy for occupations and calculating its frequency in movie subtitles. The Standard Occu-
pational Classification (SOC) system, International Standard Classification of Occupations
and European Skills, Competences, Qualifications and Occupations are some taxonomies of
professions. We use the SOC profession taxonomy, developed and used by the U.S. Bureau
of Labor Statistics, in our work.

3.1 SOC

The SOC 2018 taxonomy provides a list of professions and their corresponding job titles.
It is the standard occupational taxonomy used by federal agencies to classify workers into
occupational categories and is regularly revised. It arranges professions into four tiers -
major, minor, broad and detailed groups. There are 23 major groups, 98 minor groups,
459 broad groups, and 867 detailed groups. Each detailed group represents a profession. It
comes with a name, a definition and a set of illustrative job titles. For e.g. the occupation
Chief Executives falls under the major group Management Occupations and it includes the
job titles - Admiral, Chief Financial Officer, Chief Operating Officer etc. Figure 1 shows
the list of major groups and a portion of the Management Occupations major group of
SOC. We counted a total of 6520 unique job titles. We augment this taxonomy to make it
searchable by synonym expansion methods.

3.2 OpenSubtitles

We use movies and TV show subtitles to calculate the frequency of job titles in the media.
We do not employ movie scripts because screenplays of old films often were not found in
public datasets like IMSDb. This would not enable us to find the trend of professions over a long period. Therefore, we use the OpenSubtitles 2018 corpus, which is the largest available dataset of film and TV show subtitles, available from the year 1880 to 2018. We select the years 1950 to 2017 because they contained subtitles of at least 100 movies, giving us sufficient number of samples for each year. We end up with 135,998 movies, whose subtitle files contain approximately 126 million sentences (number of sentences per movie: \( \mu = 932.6, \sigma = 624.8, \text{median} = 812 \)). Most movies contained more than one subtitle file, that are slightly different versions of each other. These different versions were present because some were translated from other languages. We retain the subtitle file which had the most number of sentences so that we can have exactly one subtitle file for each movie.

An example excerpt taken from the subtitle file of the movie *All About Eve (1950)* is shown here -

```
01:54:24: She’s been crying all night and she’s hysterical.
01:54:28: She doesn’t want a doctor and ..
01:54:31: Who is it?
```

We have the timestamp of each utterance, but no information about the speaker or of scene transitions. This limits the context that we can work with for each sentence. Spoken language sentences are also not always grammatically correct and complete, as shown in the second sentence in the above example. The limited context and syntactic inconsistency of subtitles make it challenging to successfully apply standard NLP methods like sequence models or entity taggers. Therefore, we calculate simple handcrafted context features to disambiguate the mention of a job title in the sentence. We build a search index on the OpenSubtitles corpus using the Python WHOOSH package.

4. Methods

We find the frequency of job titles of different professions in movie subtitles. We use the job titles of the SOC taxonomy and augment it using synonym expansion methods. We improve the precision of our dictionary based search by filtering noisy words and using context features. Finally, we define a frequency measure for a profession and use it to find the trend of that profession in movies over time.

4.1 Synonym Expansion

Let \( P \) be a set of job titles belonging to some profession. Let \( Q \) be the augmented set of job titles. \( Q \) is initialized to \( P \). We shall use \( P = \{ \text{Farm Labor Contractor}, \text{Harvesting Contractor} \} \), corresponding to SOC profession Farm Labor Contractor, to illustrate our method. Here \( |P| = 2 \).

4.1.1 Choose Last Word

For all job title \( j \) in \( P \), if \( j \) does not contain any conjunctions, prepositions, parentheses or special characters, has at most 5 words and doesn’t have all letters in capital case, then we take the last word of \( j \) and add it to \( Q \). The described method adds the job title *Contractor*. So \( Q = \{ \text{Farm Labor Contractor}, \text{Harvesting Contractor}, \text{Contractor} \} \). \( |Q| = 3 \).
4.1.2 WordNet-Based

We use WordNet’s synonym and hyponym relations to extend $P$. Words with similar meanings are synonyms of each other. Words with more specific meaning that a given word are its hyponyms.

1. ASSIGN SENSE

We find the sense $s(j)$ of each job title $j \in P$. If $j$ is present in WordNet and has exactly one noun synset, we assign it to $s(j)$. If $j$ is a unigram and has at least one noun synset, we assign the most common noun synset to $s(j)$. If $j$ contains more than one word and is not present in WordNet, we use word sense disambiguation strategy to assign senses to individual words of $j$ [Pedersen et al., 2005]. For e.g. Harvesting Contractor gets the sense - reap.v.01 contractor.n.03. reap.v.01 is the first verb sense of the word reap. contractor.n.03 is the third noun sense of the word contractor.

2. CANDIDATE UNIGRAMS

For each $j \in P$, if it contains exactly one noun phrase, then the root of the noun phrase becomes a possible candidate unigram $u$ which can be used for synonym expansion. Here, Contractor is a candidate unigram.

3. BEST SENSE

We use WordNet’s Wu-Palmer similarity measure [Pedersen et al., 2004] to find the best sense for each unigram $u$. It measures similarity of two synsets by considering their depth and the depth of the lowest common subsumer.

$$s(u) = \arg \max_{s \in n(u)} \{ \max_{j \in P} \{ wup \text{similarity}(s(j), s) \} \},$$

where $n(u)$ are all noun senses of unigram $u$. If the similarity value - $\max_{j \in P} \{ wup \text{similarity}(s(j), s(u)) \}$ is greater than 0.8, we retain the unigram. In our example, Contractor has four noun senses according to WordNet. The best sense according to the described algorithm is contractor.n.03.

4. EXTEND LIST

We add the words of the synset $s(u)$ and words contained in all hyponyms of $s(u)$ to $Q$ for all unigram $u$ that were retained in the previous step. contractor.n.03 has the hyponym builder.n.03, which contains words - Builder and Constructor. We add them both to $Q$.

The WordNet-Based synonym expansion algorithm outputs $Q = \{ Farm \text{ Labor Contractor, Harvesting Contractor, Contractor, Builder, Constructor} \}$. Here $|Q| = 4$.

4.2 Synonym Filtering

The synonym expansion strategies described above expands the list of job titles. However, not all words in the new list indicate some profession. E.g. there are words like - Father, Gentleman and Washing Machine that get added after synonym expansion. Such words appear because of errors in the ASSIGN SENSE step of the WordNet-Based method (4.1.2), and the heuristic rule used in the CHOOSE LAST WORD method (4.1.1). Using this list directly will decrease precision. We remove noisy words using crowd supervision.
Another source of error is the context in which the job title appears in the sentence. The presence of a job title in a sentence does not always mean that it indicates a profession. E.g. the word cleaner means a profession in the sentence - *He is a cleaner working at this hotel*, but not here - *Boraxo waterless hand cleaner really cleans up for us*. We use contextual features to decide if a given mention of a job title in a sentence is used in the sense of indicating a profession. We do this to further improve the precision of our search.

### 4.2.1 Crowd Filtering

We obtained 11703 job titles after synonym expansion. We enqueued them on MTurk [Buhrmester et al., 2016], and asked annotators whether the most common use of the job title word is to indicate some occupation. We retained only those job titles that were flagged yes in all its annotations. The reduced list contained 7619 job titles, a 16% increase from the original SOC list that contained 6520 job titles.

### 4.2.2 Sense Disambiguation

We use the following 4 features to describe the context of a job title in a sentence - i) part of speech tag of the current, preceding and succeeding word, ii) capitalization, iii) position of the word in the sentence and iv) the preceding and the succeeding word. These features were inspired from work done in unsupervised named entity recognition [Nadeau and Sekine, 2007]. The final dimension of the feature vector is 436.

We randomly sample 100000 sentences from the OpenSubtitles corpus that contained some job titles from our augmented SOC list. We randomly sample an equal number of sentences that do not contain any job titles. The former set makes up the positive examples and the latter set is the negative examples of our training set. We follow a similar approach to create the validation set, which contains 1000 positive and 1000 negative examples. For the negative examples, we randomly choose a phrase in the sentence to construct the feature vector. We train a logistic regression model for the binary classification task of predicting whether the available context indicates a profession. We use SCIKIT-LEARN [Pedregosa et al., 2011] for implementation. We tuned the regularization parameter $C$. We obtained precision = 0.865, recall = 0.961 and F1 = 0.910 on the validation set. We don’t explore sequence models because of the limited context of subtitles (see 3.2) and satisfactory performance obtained using simple logistic regression. We call this the disambiguation model in later sections.

### 4.3 Frequency Measure

We use the new dictionaries obtained via synonym expansion, and the contextual features to search for relevant sentences. A sentence is relevant if and only if it explicitly contains a job title and its sense in the sentence refers to the profession. Let $T(P, y)$ be the number of relevant sentences that contain mentions of job titles of profession $P$ in the movie subtitles of year $y$. Then the normalized frequency measure of $P$ in year $y$ is $F(P, y) = \frac{T(P, y)}{\sum_{P'} T(P', y)}$, where $P' \in$ all SOC professions. To construct the frequency measure of a major group $M$ which includes multiple professions, we take the union of the relevant sentences for each profession $P \in M$ and calculate $T(M, y)$ and then $F(M, y)$ similarly.
5. Experiments

Job titles like *Acupuncturist*, *Deputy Officer* and *Clocksmith* always mean a profession irrespective of the context in which they are used, whereas job titles like *Cleaner* and *Cook* can mean different things. See section 4.2 for an example of how *Cleaner* can be used in different senses. We evaluate the precision of our different methods on the task of correctly identifying relevant sentences containing such polysemous job titles. We use the number of senses of a job title as the measure of its polysemy.

<table>
<thead>
<tr>
<th>Dictionary</th>
<th>Filter</th>
<th>Context</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>-</td>
<td>OR</td>
<td>0.695</td>
<td>0.933</td>
<td>0.796</td>
</tr>
<tr>
<td>Soc</td>
<td>-</td>
<td>-</td>
<td><strong>0.909</strong></td>
<td>0.264</td>
<td><strong>0.409</strong></td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>AND</td>
<td>0.921</td>
<td>0.250</td>
<td>0.393</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>OR</td>
<td>0.695</td>
<td>0.948</td>
<td>0.802</td>
</tr>
<tr>
<td></td>
<td>YES</td>
<td>-</td>
<td>0.808</td>
<td>0.728</td>
<td>0.766</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>AND</td>
<td>0.825</td>
<td>0.683</td>
<td>0.747</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>OR</td>
<td>0.690</td>
<td>0.979</td>
<td>0.809</td>
</tr>
<tr>
<td>Soc-LW</td>
<td>YES</td>
<td>AND</td>
<td><strong>0.862</strong></td>
<td>0.581</td>
<td>0.694</td>
</tr>
<tr>
<td></td>
<td>YES</td>
<td>OR</td>
<td>0.695</td>
<td>0.971</td>
<td><strong>0.810</strong></td>
</tr>
<tr>
<td>Soc-WN</td>
<td>YES</td>
<td>-</td>
<td>0.673</td>
<td>1.000</td>
<td>0.805</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>AND</td>
<td>0.695</td>
<td>0.933</td>
<td>0.796</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>OR</td>
<td>0.673</td>
<td>1.000</td>
<td>0.805</td>
</tr>
<tr>
<td></td>
<td>YES</td>
<td>AND</td>
<td><strong>0.768</strong></td>
<td>0.769</td>
<td>0.769</td>
</tr>
<tr>
<td></td>
<td>YES</td>
<td>OR</td>
<td>0.691</td>
<td>0.984</td>
<td><strong>0.812</strong></td>
</tr>
</tbody>
</table>

Table 1: Precision, Recall and F1 values of different methods on the test set

We choose job titles that have more than one noun sense according to WordNet and if they occur in at least 50 subtitle sentences in our corpus. For each chosen job title, we randomly sample 25 sentences containing it and enqueue them on MTurk. Annotators answer yes for sentences if and only if the contained job title indicates some profession. Note that sentences can implicitly imply some profession without the explicit use of a job title. E.g. “Charlie works in the chocolate factory” implies that the person Charlie is a chocolate factory worker, but should be flagged no according to our definition of a relevant sentence. We collect at least three annotations for each sentence. We retain those sentences which had complete agreement (all yes or all no). There were 9487 sentences in total - 6387 relevant and 3100 non-relevant sentences. We use this as the test set for our evaluation. See Table 1 for the results. Soc-LW and Soc-WN are the augmented SOC dictionaries obtained via the Choose Last Word and WordNet-Based methods respectively. A yes in the Filter column means that we used crowd supervision to filter the dictionary. If the Context column is empty, context features were not used to decide if the sentence is relevant. If it contains and, a sentence is relevant if it contains words from the dictionary and they are used to indicate the profession according to the disambiguation model. If it contains or, a sentence is relevant if either condition mentioned in the prior sentence is true.


From table 1, we observe that the original Soc dictionary has high precision (0.909) but low recall (0.264). This happens because its job titles are very specific. Less than 7% of Soc job titles are unigrams, but they occur in more than 90% of relevant sentences. Soc contains job titles like Family Physician and Mechanical Engineer, but the words Physician and Engineer are absent. Recall improves significantly with the new dictionaries Soc-LW (+0.464) and Soc-WN (+0.736), but we lose precision. This occurs because of the presence of noisy words and wrong context (see section 4.2.2 for example). The filtered dictionaries (Filter = yes) significantly improve precision but still have lower F1 scores. Combining the results of dictionary look-up with context features give the best overall F1 scores for all three dictionaries. Combination using disjunction (Context = or) showed improved gains in F1 scores than conjunction (Context = and), but had lower precision. We use filtered Soc-WN dictionary combined with context features via disjunction as the model to find relevant sentences for calculating frequency measures. All tests for significance were conducted using permutation test, at $\alpha = 0.05$ and Bonferroni correction $n = 16$ [Sedgwick, 2012].

6. Analysis

We use the Soc-WN dictionary combined with contextual features to find relevant sentences for each profession. We use this to calculate the normalized frequency $F(M, Y)$ for major profession group $M$ and for all years $Y$ (see 4.3). $F(M) = (F(M, y))_y$ where $y \in$ all years, gives us the frequency time series of $M$. We analyze the trend of $F(M)$ over time and against the employment numbers of $M$. 
6.1 Movie Trend

We calculate the normalized frequency measure $F(M, Y)$ for all 23 major profession groups of SOC and for each year $Y$ from 1950 to 2017. We calculate Spearman’s rank correlation coefficient for the time series $F(M)$. Correlation coefficient was significantly positive for 6 profession groups ($\alpha = 0.05$). Business and Finance; Art, Design, Entertainment, Sports and Media and Community and Social Service profession groups had the highest positive correlation. Correlation coefficient was significantly negative for 6 other profession groups. Farming, Fishing and Forestry; Construction and Extraction and Installation, Maintenance and Repair profession groups had the highest negative correlation. Figure 2a shows the trend of frequency measure $F(M)$ across time for these profession groups. A similar analysis of individual professions revealed some interesting trends. Word frequency of professions like Therapists, Programmers and Detectives increased over time, whereas those of Operators, Clerks and Laborers decreased over time.

6.2 Movie vs. Employment Trend

We compare the frequency measures $F(M)$ of major profession group $M$ against its employment numbers $E(M)$. $E(M, Y)$ is the fraction of population employed in some profession of $M$ in year $Y$. $E(M) = (E(M, y))_y$ is the resulting employment time series of $M$, and $E(Y) = (E(m, Y))_m$ where $m \in$ all major profession groups, defines the occupational structure in year $Y$. We calculate Spearman’s rank correlation coefficient between $F(M)$ and $E(M)$. It was significantly positively correlated for 9 major groups ($\alpha = 0.05$). Both the employment share and movie frequency of major profession groups Arts, Design, Entertainment, Sports, and Media; Business and Finance and Computer and Mathematics occupations increased over time, whereas both decreased for major profession groups Production, Installation, Maintenance, and Repair and Construction and Extraction occupations. Frequency and employment was significantly negatively correlated for Community and Social Service and Healthcare Support occupations. The frequency of their job titles decreased over time in movie subtitles but their employment share increased. Figure 2b shows the trend of frequency and employment for the above three cases.

7. Conclusion

In this paper, we have studied the representation of professions in media by calculating the frequency of job titles in film subtitles. We use the SOC taxonomy and make it searchable using synonym expansion strategies based on WordNet. We use the augmented taxonomy to search relevant sentences and improve the precision using contextual features. We finally calculate the frequency values using the retrieved sentences and compare them across time and against employment share. Most of the major profession groups showed a significant positive correlation with their employment trend. Managerial and business occupations occur more in subtitles than construction and labor jobs. The general trend seems to indicate an increase in the frequency of white-collar job titles than blue-collar job titles. Our synonym expansion method can be generalized to augment and create other similar taxonomies. Future work includes applying the same methods to the SIC taxonomy for industry and business names.
Representation of Professions in Media

Acknowledgments

Omitted for blind review.

References


VM Moskovkin, TV Saprykina, EV Pupynina, VA Belenko, and IA Shumakova. Examination of trends in education with the google books ngram viewer. 2019.


