Mining the US Technologist data

Abulitibu Tuguluke
DHI Group Inc.
6465 Greenwood Plaza Blvd
Greenwood Village, CO 80111
abtu8803@colorado.edu

1 PROBLEM STATEMENT/MOTIVATION

It is no secret that the new competition for this century is the 'hunt' for talents., "by 2030, more than 85 million jobs could go unfilled because there are not enough skilled people to take them, according to the latest study conducted by Korn Ferry."[1] "The shortage has been amplified by an insufficient number of U.S.-citizen computer science college graduates, restrictive (until recently) and limited H-1B visas for experienced IT professionals to fill the immediate gaps", echoed the TechServe consulting. [2] Dice.com's internal analysis already showed that we are going through technology workers (Technologists) shortage, and that is not due to the 'Great Resignation' but due to market demand. Our report will mine through the US H1B data set along with Dice.com's job and candidate profiles (along with skill sets for each job title), to better understand which technologist job (e.g.: Software developer, Data scientist) are in increasing/decreasing demands, which skill sets (e.g.: Python, C++ etc.) are more popular/unpupular, through time-series. Our goal is to better understand the US technology market and what we can do to help people who try to get into tech market, by providing them with skill-set guidance along with professional training advice.

The motivation for this report came from a conversation with one product manager stating that there are no analysis on technologist supply and demand, even on salary data! The product analysts stated they don't have the data, and the R&D team think it is a analyst job and should not 'waste' time on. I decided to try that as a personal project. As I dive in more, I realized the official data set provided from department of labour is vague and uninformative, most of the job titles are lumped together, and focusing only on salary, hence no detailed information based on skills and locations. I then decided to mine other open source data, while adding relative skill set by mining seperate job description data set.

My philosophy to tackle this problem/project is to use a simple algorithm on multiple complicated sets, and not the other way: some complex algorithms on a nice set instead. Personally, I had some 'fatigue' issues with words like 'Deep' and 'Neural' for now, both at work and school, so I've decided to do some old fashion mining by not using heavy GPU computing with 'sexy' machine learning titles, and just have some miner's fun with my hardhat and shovel.[3]

2 LITERATURE SURVEY

SpaCy's Pipeline component for part-of-speech taggging[4] comes in really handy for the data preparation. What we did was to manually add a 'Tag' button for SMEs to spot the 'tech' words from our training set in HTML, resembling the following table:

	Sentence	Word	Tag	POS	Job description #
0	1	Softwares	Tech	NNPS	description: 1
1	2	SQL	Tech	NNP	description: 1
2	3	Java	Tech	NNPJ	description: 1
3	4	Web	Tech	NNS	description: 1
4	5	Azure	Tech	NNP	description: 1
⋮	⋮	⋮	⋮	⋮	⋮

Once we got enough samples, we apply the Viterbi algo to the future data set and mine the list of words for each job post, we then trim the list to get the overall skill-set for each title, like the following for sample job titles.

Job	Skills
Net Application Developer	Microsoft technologies;Software development;C#;HTML;Quality assurance;ASP.NET;Visual Basic .NET;.NET;Agile.
Android Developer	Software development;Java;Mobile development;Quality assurance;Android development.
⋮	⋮

3 PROPOSED WORK

Shockly, the department of labour data does not give any insite on specific job title but a very general one. The following 3 code has most of the tech jobs:

- Management Occupations 11-0000
- Business and Financial Operations Occupations 13-0000
- Computer and Mathematical Occupations 15-0000

Roughly 35 of them instead of what we proposed of roughly 1400+.

```
['Computer and Mathematical Occupations',
'Computer Occupations, All Other*',
'Mathematical Science Occupations, All Other',
'Computer Occupations',
```

[1] https://www.kornferry.com/insights/this-week-in-leadership/talent-crunch-future-of-work

[2] https://www.techservealliance.org/news/the-state-of-the-technology-talent-shortage/

[3] I will leave that to CSCI 5922-Neural Networks and Deep Learning, this fall.

[4] https://spacy.io/api/tagger

CSCI 5502, Boulder, University of Colorado
© 2022 ACM. 978-x-xxxx-xxxx-x/YY/MM…$15.00
DOI: 10.1145/nnnnnnn.nnnnnnn

```
'Miscellaneous Computer Occupations',
'Computer Occupations, All Other',
'Mathematical Science Occupations',
'Miscellaneous Mathematical Science Occupations',
'Data Scientists and Mathematical Science Occupations, All Other'
...
]
```

Which mean we either has to add a new attribute by clustering the 1400 into 15, or again write a new scrapper it mine glassdoor's data()This should not end up being a scrapper project), which does not guarentee accuracy since all are self-reported, which could be more biased than H1b data. Somehow, we have to reconcile that.

- Data scraping: Screaping HTML data from the web and store them in seperate json file
  EMPLOYER : String, Nominal
  JOB TITLE : String, Nominal
  BASE SALARY: Float, Interval
  LOCATION: String, Nominal
  SUBMIT DATE: (date) String, Ordinal
  START DATE: (date) String, Ordinal
- Data cleaning: The most time consuming process, we need to be able to get rid of the outliers once we spotted them.
- Data preprocessing: Converting the saw data into meaningful use, for numeric data, we need to convert them into float64, date has to be standard date type.
- Data integration: Integrate internal data with mined H1b data by join.
- Data mining and analysis: Clustering and time series.

4 DATA SET

4.1 U.S. Department of Labor data

What better resource to go to than the Department of Labor's statistics, After consulting with a product manger who points to the open API the department provided, I decide to mine their data first. Shockly though, their data does not give any insight on specific job title but a very general one. I've choose to call the data between year 2010 and 2021 so that to better match with H1b's time series, and here's what we get, after thorough analysis of 14027 data entries,with total 34 columns:

There are 1272 job titles from department of labor open data, among them, the only code (15-XXXX) that resemble our definition of technologist is :

- Computer and Mathematical Occupations 15-0000

Unfortunately, only 37 of them instead what we proposed of roughly 1400+ is publicly stored, they are:

- 'Computer and Mathematical Occupations',
- 'Computer and Information Research Scientists',
- 'Computer Systems Analysts', 'Computer Programmers',
- 'Software Developers, Applications',
- 'Software Developers, Systems Software', 'Database Administrators',
- 'Network and Computer Systems Administrators*',
- 'Computer Support Specialists',
- 'Information Security Analysts, Web Developers, and Computer Network Architects',
- 'Computer Occupations, All Other*', 'Actuaries', 'Mathematicians',
- 'Operations Research Analysts', 'Statisticians',
- 'Mathematical Technicians',
- 'Mathematical Science Occupations, All Other',
- 'Computer Occupations', 'Computer and Information Analysts',
- 'Information Security Analysts',
- 'Software Developers and Programmers', 'Web Developers',
- 'Database and Systems Administrators and Network Architects',
- 'Network and Computer Systems Administrators',
- 'Computer Network Architects', 'Computer User Support Specialists',
- 'Computer Network Support Specialists',
- 'Miscellaneous Computer Occupations',
- 'Computer Occupations, All Other',
- 'Mathematical Science Occupations',
- 'Miscellaneous Mathematical Science Occupations',
- 'Database and Network Administrators and Architects',
- 'Database Administrators and Architects',
- 'Software and Web Developers, Programmers, and Testers',
- 'Software Developers and Software Quality Assurance Analysts and Testers',
- 'Web Developers and Digital Interface Designers',
- 'Data Scientists and Mathematical Science Occupations, All Other'.

Table 1: Deparrtment of labour data

	count	unique	top
occ_code	14027	1507	29-2010
occ_group	9731	5	detailed
occ_title	14027	1272	Tour and Travel Guides
group	46	2	major
tot_emp	14027.0	8936.0	11860.0
annual	852	1	True
hourly	62	1	True
year	14027.0	NaN	NaN
area	2658.0	NaN	NaN
area_title	2658	1	U.S.
prim_state	1329	1	US
emp_prse	14027.0	229.0	0.5
mean_prse	14027.0	NaN	NaN
a_mean	14027	6155	*
a_median	14027	5808	#

This dataset does not provide any details on what do these occupations include. We tried some online resources[5] (including DoL's own mapping data from 2018[6]) to try mapping the 1400+ tech job into these 37 categories that yields horribly inaccurate result , which mean we either has to add a new attribute by clustering the 1400 into 15, or again write a new scrapper that mines Glassdoor's data (This should not end up being a scrapper project), even that does not guarantee a good accuracy since all are self-reported, hence more biased than H1b data. Due to the time limit, we have to reconcile that.

[5] https://occupationdata.github.io/apst_mapping.pdf
[6] https://www.bls.gov/soc/2018/home.htm, only 20+ were able to matched, that leaves 1380+ unmatched still

4.2 Scrapped H1b data

- H1B data scraping from from open data online.[7] The data set is mainly from the United States Department of Labor (DOL) on how many H1-B petitions were filed and approved, with detailed information such as Employer, the title of job, city/state locations, along with base salary, submission and acceptance dates.
- Skill set (Internal data)
- US Bureau of Labor Statistics [8]
- Data mined and stored on my machine (Prcessor: 2.6 GHz 6-Core Intel Core i7. Memory: 32 GB 2667 MHz DDR4. Graphics:AMD Radeon Pro 5500M 4 GB Intel UHD Graphics 630 1536 MB).

5 EVALUATION METHODS

There are many ways how we can evaluate the results. Clustering Performance Evaluation Metrics like Silhouette coefficient is a good start when applying clustering algorithm, for time series, all the error evaluations. Besides, comparing time-series analysis data within each job we can compute the increaing or decreasing sides of each job along with correlations. Finally ,we are looking forward to apply all the relavent techniques we are currently learning in the class.

As for salary data, percentile based on title, location, and or starting time is definitely the first thing to try, 99% and 1% are definitely outliers for that, and balancing what percentile is something I need to be careful about. I am not thinking of 'elimination' since all data counts for that column, I may if it is way imbalanced. Hence we need separate analyses for both individuals and employers. For skill sets, I am considering density distance, which is yet to explore. Such as:

5.1 Elbow method for clustering: TBD

"Using the "elbow" or "knee of a curve" as a cutoff point is a common heuristic in mathematical optimization to choose a point where diminishing returns are no longer worth the additional cost. In clustering, this means one should choose a number of clusters so that adding another cluster doesn't give much better modeling of the data."[9]

6 TOOLS

- Beautifulsoup for mining.
- Python library for all the computing and analysis.
    Spacy
    Sklearn
    Tableau for visualization
- JSON for data store.

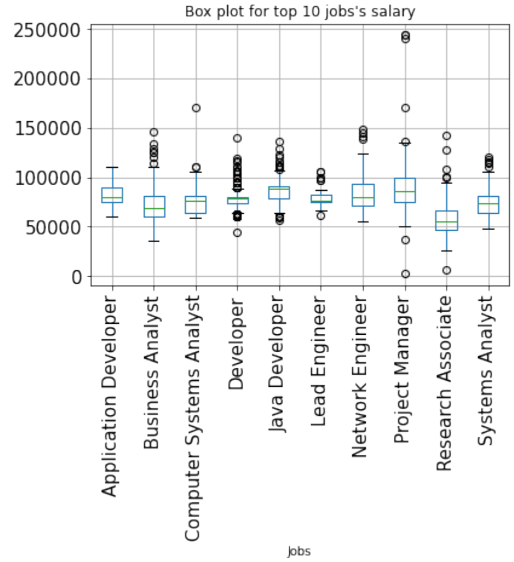

Figure 1: Top 10 top job title with the highest pay boxplot.

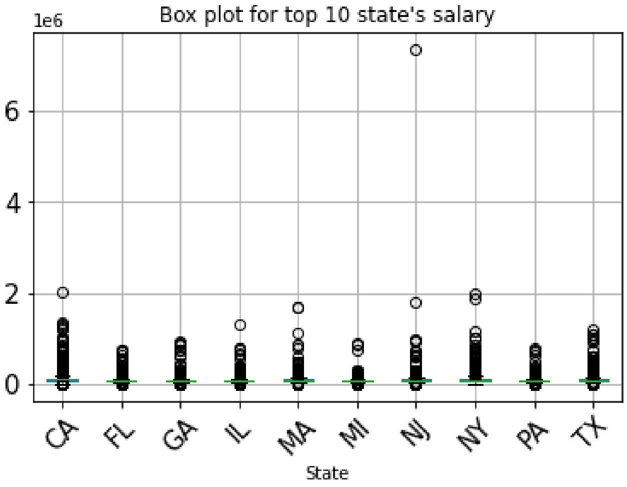

Figure 2: Top 10 top state with the highest pay boxplot.

[7] https://h1bdata.info/index.php.
[8] https://www.bls.gov/developers/api$_$python.htm
[9] wikipedia

Table 2: Scrapped data as off July 15th

	Dice_job_title	EMPLOYER	JOB TITLE	BASE SALARY	LOCATION	SUBMIT DATE	START DATE	Job_Title
count	2495079	2495052	2467304	2467304	2467304	2467304	2467304	2495079

7 MILESTONES COMPLETED

Date	Work Plan	Status
Week 1 - 6: 16 May - 26 June	Research and data scrapping while class works.	Complete
Week 7: 27 June - 3 July	Submit Project Part 1: Project ProposalsAssignment.	Complete
Week 8: 4 July - 10 July	Data cleaning	Complete
Week 9: 11 July - 17 July	Submit Project Part 2: Proposal PaperAssignment	Complete
Week 10: 18 July - 24 July	EDA and Classification Algorithms.	semi-Complete
Week 10: 18 July - 24 July	Analyzing Department of Labor API data	Complete

8 MILESTONES TO-DO

Date	Work Plan	Status
Week 10: 18 July - 24 July	EDA and Classification Algorithms.	semi-Complete
Week 11: 25 July - 31 July	Submit Project Part 3: Progress ReportAssignment.	Compelete
Week 12: 1 August - 7 August	Continue writing the project, and building interactive demo, build an interactive Tableau dashboard.	To be Compeleted
(Final)Week 13: 1 8 August - 10 August	Submit Peer Evaluation Form, Project Final Report(with codes), and Project PresentationAssignment(Video).	To be Compeleted

9 RESULTS SO FAR

Although, the discovery does not mean DoL's data are completely worthless, there are some good insights that confirm our believe that tech talents are in rising demand. Take Figure 3 for example: The top 2 lines represent 'computer occupation' and 'software developer' professions, their growing rates are even higher than the rest of technologists, it is a clear indication of higher demand for such labor, as opposed to the Farming jobs, which stays relatively the same, in some cases, even declined, as shown in Figure 4. If we look at the salary data for technologist in Figure 5, we noticed the significant salary increase over the year, take into account the inflation and supply, it only amply the demand side of the tech market, which forces the employers to hand out competative salary. As oppose farming job market shown in Figure 6, where is did not even cover the inflation over a decade.

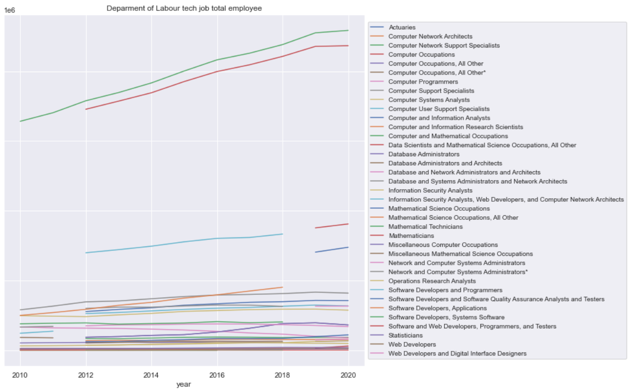

Figure 3: Deparment of Labour tech jobs number of employee.

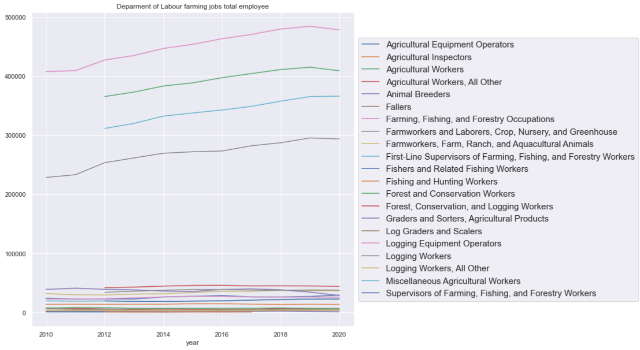

Figure 4: Deparment of Labour farming jobs number of employee.

All of these give us more reasons to find our own data, as an immigrant worker myself, the second 'gold mine' to dig is H1b, which will be the main focus of the next section.

10 PART 1 PROPOSAL FEEDBACKS FROM CLASSMATES AND MY THOUGHTS

Pratik Prasanna Raghavendra

Interesting topic, Abulitibu. Just curious, what would outlier data look like in your case? Also, how do you detect and eliminate them, since for example, really high income may simply be an outlier for individuals, but an important point for other employers?
My reply: As for salary data, percentile based on title, location, and or starting time is definitely the first thing to try, 99% and 1% are definitely outliers for that, and balancing what percentile is something I need to be careful about. I am not thinking of 'elimination' since all data counts for that column, I may if it is way

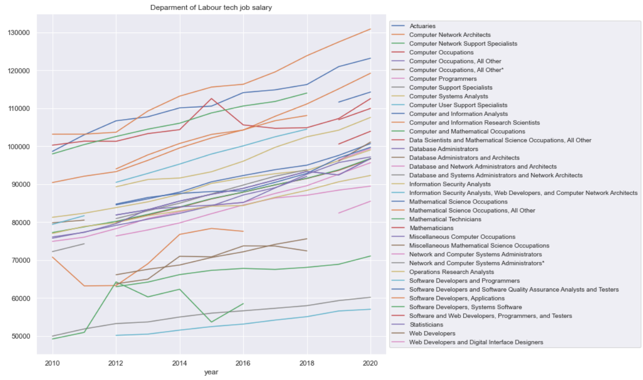

Figure 5: Deparment of Labour tech job salary.

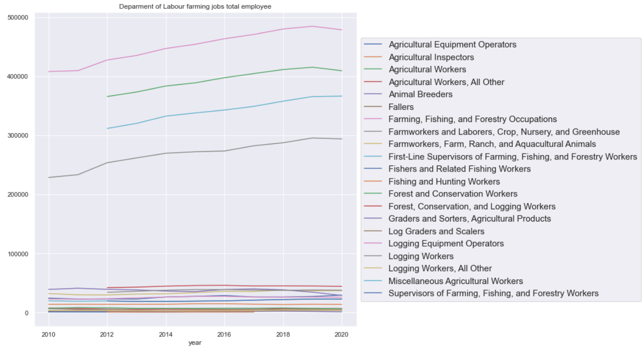

Figure 6: Deparment of Labour farming jobs salary.

imbalanced. Hence we need separate analyses for both individuals and employers. For skill sets, I am thinking density distance, which is yet to be explored.

Rohit Kharat

I find this a unique topic and interesting to analyze in today's scenario. I liked that you are using web scraping, which I enjoy; using the BeautifulSoup library for web mining is fun. I am not familiar with the SpaCy library, something you can shed a light on. I also liked the idea of using location information for your analysis. I would recommend geolocation API something which I used earlier in my projects. I would also recommend doing feature reduction or getting the important features in your proposed work for focusing on only the crucial attributes. My reply: Seems like I need to reduce a lot of features. Too repetitive. The latest data shows:

	count	unique	top	freq
Dice_job_title	603967	1107	Systems Analyst	35605
EMPLOYER	603967	66538	TATA CONSULTANCY SERVICES LIMITED	36885
JOB TITLE	603967	1655	SYSTEMS ANALYST	35605
BASE SALARY	603967	37791	60,000	23256
LOCATION	603967	10453	NEW YORK, NY	34957
SUBMIT DATE	603967	2747	03/16/2018	2377
START DATE	603967	2897	10/01/2020	17065
Job_Title	603967	1107	Systems Analyst	35605
relatedSkills	603967	1101	Software development;Systems analysis;SQL;Qual...	35605

PCA is always my go to algorithm when I redecide to reduce the information's dimensions. For Geo, I think for now I will just use a really good map, since I still don't have demand data from job sites.

Abhinav Gupta

This looks like an unusual topic! Curious to see what your data cleaning and preprocessing would look like. As your job skill data is internal, I don't know how it looks like, but I have a couple of questions. First question is that technologies are evolving and higher demand will always be for the new tech. As that tech, perhaps, is not even in the market yet, how would you predict the demand? Second question is if you would be considering countries and degrees as visa approval ratio varies highly on these factors. Final question is that as per uscis.gov - "The H-1B temporary visa program has been exploited and abused by employers primarily seeking to fill entry-level positions and reduce overall business costs", so wouldn't your dataset be corrupted too?
My reply: For small or foreign companies, it seems sto be so, but not for big and advanced tech companies, Google filing counts for top 10, yet the salary is not below the market. It all depend on what we choose. I started as CPT then OPT, later H1b, yes, at first, I was the cheap labor, but once I accumulated enough skill, my salary is above the market while still on H1b, took almost 3 years. But when it comes to 'Corrupted' data, yes we do have that, and that's the purpose of introducing data mining, isn't it?

Varun Manjunath

An interesting topic. But is often the case that H1b workers are paid less as H1b is a nonimmigrant visa and is temporary in nature. So the point where you perform analysis on the dataset to check whether H1b workers are paid less is not very useful. Rather a more interesting trend would be to check which location hires the most H1b candidates. Dataset description with a bit more depth would be great in your case. Also, are you planning an inner join query to combine the datasets? Would suggest you remove certain outliers in the combined dataset before performing K-Means clustering.
My reply: left join mostly, inner join give few results.

Danielle Aras

This is a very interesting and relevant topic. I agree that data cleaning is an important part of your project as you are merging different data sources with different features. I would be curious to know more details about the data sets like the size and date range. Technology is a fast moving field so the most current data will be the most useful for job seekers; looking at which jobs are trending up or down is a great idea.

I agree with other posters that the H1B visa data will introduce some challenges to your analysis. Given the trend of employers looking to take advantage of the system to under-pay workers, the popularity of H1B workers in a particular job may not actually correspond with a genuine demand that is not met by domestic workers. An interesting "spin-off" question would be creating an algorithm to detect companies abusing the system based on the availability of candidates in their area and H1B salaries.
My reply: I like this idea and will add it to discovery session.

REFERENCES

[1] Aurélie Breidenbach, Caroline Mahlow, and Andreas Schreiber. 2021. Implicit Gender Bias in Computer Science – A Qualitative Study. (July 2021). http://arxiv.org/abs/2107.01624 arXiv:2107.01624 [cs].

[2] Jens-Joris Decorte, Jeroen Van Hautte, Thomas Demeester, and Chris Develder. 2021. JobBERT: Understanding Job Titles through Skills. (Sept. 2021). http://arxiv.org/abs/2109.09605 arXiv:2109.09605 [cs].

[3] Nicolas Gutierrez and Manuela Wiesinger-Widi. 2016. AUGURY: A time-series based application for the analysis and forecasting of system and network performance metrics. In *2016 18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*. 351–358. DOI: https://doi.org/10.1109/SYNASC.2016.062 arXiv:1607.08344 [cs].

[4] Bin Ji, Shasha Li, Jie Yu, Jun Ma, and Huijun Liu. 2022. Win-Win Cooperation: Bundling Sequence and Span Models for Named Entity Recognition. (July 2022). http://arxiv.org/abs/2207.03300 arXiv:2207.03300 [cs].

[5] Zachary A. Pardos and Andrew Joo Hun Nam. 2018. A Map of Knowledge. (Nov. 2018). http://arxiv.org/abs/1811.07974 arXiv:1811.07974 [cs].

[6] Nina Smirnova and Philipp Mayr. 2022. Evaluation of Embedding Models for Automatic Extraction and Classification of Acknowledged Entities in Scientific Documents. (June 2022). http://arxiv.org/abs/2206.10939 arXiv:2206.10939 [cs].

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. (Dec. 2017). http://arxiv.org/abs/1706.03762 arXiv:1706.03762 [cs].

[8] Xinyi Wang, Zitao Wang, Weijian Sun, and Wei Hu. 2022. Enhancing Document-level Relation Extraction by Entity Knowledge Injection. (July 2022). http://arxiv.org/abs/2207.11433 arXiv:2207.11433 [cs].

[9] Ziyang Wang, Wei Wei, Chenwei Xu, Jun Xu, and Xian-Ling Mao. 2022. Person-job fit estimation from candidate profile and related recruitment history with co-attention neural networks. (June 2022). http://arxiv.org/abs/2206.09116 arXiv:2206.09116 [cs].

[10] Chuhui Xue, Jiaxing Huang, Shijian Lu, Changhu Wang, and Song Bai. 2022. Contextual Text Block Detection towards Scene Text Understanding. (July 2022). http://arxiv.org/abs/2207.12955 arXiv:2207.12955 [cs].

[11] Xingzhi Zhou and Nevin L. Zhang. 2022. Deep Clustering with Features from Self-Supervised Pretraining. (July 2022). http://arxiv.org/abs/2207.13364 arXiv:2207.13364 [cs].