Job Seeking Application via Text Analytics

1. Executive Summary

This report is regarded as a practical project in the field of job-seeking applications. It mainly focuses on the text analysis area methods and tools such as topic modelling, cosine similarity, as well as NLP (Nature Language Processing) tokenization, stemming and lemmatization. In this paper, we will present a preliminary use case for how to improve the current major job listing platform to offer better recommendations. Moreover, we also include our results analysis, future development and reflections.

2. Introduction

With the advance of technology, available information online has been increasing significantly over the years. While the inflation rates keep increasing after the global pandemic, more people are actively looking for jobs. It brings current job research engines bigger challenges in a higher quality of matching, multiple dimensions of searching conditions and more demand from fresh graduates. Job searching can be an inevitable topic and crucial for other experienced job seekers.

Therefore, we designed three major ways to help to improve the job search experience, which are candidates' resume-based, candidates' learnt courses-based, and general job market trends information based. With these goals, we end the introduction by briefly describing the dataset used in this project.

2.1 Dataset Used

Job Descriptions: This job description dataset is from Kaggle, which contains job descriptions from regions all around the UK. It has around 240k rows and 12 columns. The major data fields that will be used are Job title, Full description, Company and Category etc. The full data files are in Annex A.

SMU MITB Courses: This dataset was scraped off the public information on the SMU website in September 2022. The columns in use are “Course title”, “Course outline” and MITB track information, e.g., “Financial technologies track”, “Analysis track”, “AI track” and so on.

Sample Resume: We managed to retrieve some sample resumes from previous MITB alumni and current MITB students. We process all the text words in the resumes.

3. Methodology

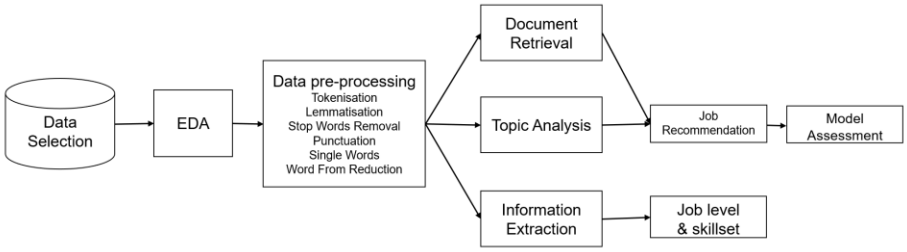
3.1 challenges

In topic analysis, we introduce LDA (Latent Dirichlet allocation) as a typical model that is based on the statistical model. However, there are no definite criteria for quantitative analysis of the quality of extracted topics. Hence the model performance evaluation between cosine similarity and topic modelling is a challenge. The detailed solution is explained in 5.1 and 5.2. The other challenge is in the information extraction task. If we process the job descriptions with stemming and lemmatization at first, then we would lose the relationship between skills and modifiers.

3.2 Solutions and Procedures

We extract 8050 jobs with titles including “data/business/analyst” to improve model accuracy and computation efficiency. We use cosine similarity and topic modelling for resume matching and generate the top 20 recommended jobs for comparison. As these two models are both unsupervised learning, we

introduce a method called Assessment Audit[1] for comparison. Information extraction can generate insights into market trends of job level and skillset for a better understanding. The figure below shows the sequence flow of the implementation.

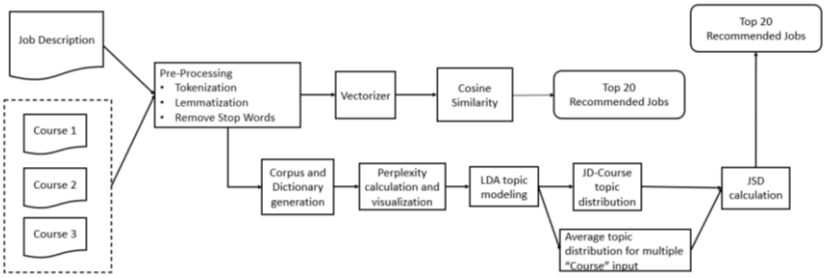


4. Solution Details

4.1 Resume Matching Methodology

We used Scikit-learn's “CountVectorizer” to convert the pre-processed job descriptions and sample resume to a vector of term counts and token counts. After that, we imported the “cosine_similarity” module from the “sklearn. Metrics. pairwise" package. Then we can call “cosine_similarity()” by passing both vectors. Then the system can calculate the cosine similarity. The cosine similarity value range is between [0,100]. By using this functionality, the jobs with the highest cosine similarity would be our matching results for this use case.

4.2 Course Matching Methodology



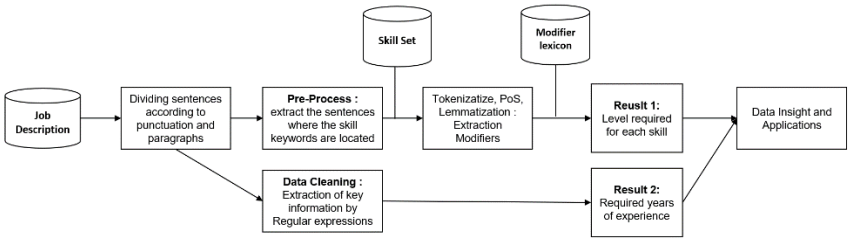
Cosine Similarity: To match their courses with the job market as resume matching.

Topic Analysis: We vectorize course instructions and job descriptions. To distinguish the importance of different words in courses and jobs, we use the TF-IDF technique to model all courses and jobs and transform them into vectors (Vayansky, I., & Kumar, S. A., 2020). The above text vectors are used for topic modelling with LDA model. We conduct LDA topic modelling for all courses and jobs, and different topics can provide different knowledge from different dimensions. Here we settle on topic number 6, and it will be explained in section 5.2. This may effectively help subsequent similarity calculation and recommendation. It could well reduce vector dimensionality to save computation time and cost. For each proposed course, we use Jensen–Shannon Divergence (JSD) to calculate the distance and finally

[1] Assessment audit allows the laboratory to understand how well it is performing when compared to a benchmark or standard.

recommend the 20 jobs with the highest similarity. (Menéndez, M. L., Pardo, J. A., Pardo, L., & Pardo, M. C., 1997). It is calculated as follows: $JS(P_1||P_2) = \frac{1}{2}KL\left(P_1||\frac{P_1+P_2}{2}\right)+\frac{1}{2}KL(P_2||\frac{P_1+P_2}{2})$, refers to Annex G.

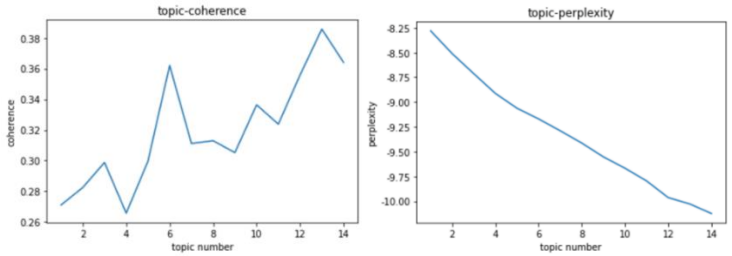
4.3 Information Extraction



We intend to use text analysis to extract this information and convert it into structured data. First, we divide the resume description into sentences based on punctuation and paragraphs. Secondly, we extract sentences with skill keywords. Thirdly, we lexically classify the sentences based on PoS(part of speech) and extract the modifiers (Adjective, Adjective, Noun) in the sentences. Finally, we use a self-labelled dictionary of modifiers to determine the level of the skill. The different modifiers correspond to 4 different levels Basic, Experienced, Proficient, and Expert. In addition, we also extract the number of years of work experience required in the job description by regular expression.

5. Experiments

5.1 Effect of different topic numbers



We conducted extensive experiments to search for the best value of the topic number. The right figure shows that the coherence of the model reaches the local optimum when the topic number is 6. Based on the effects of the two metrics, we finally set the topic number as 6, considering that 13 topic words would increase the vector dimensions, which increases the computation cost and time for subsequent similarity calculation. This is a trade-off value. (Mikolov, T., Chen, K., Corrado, G., & Dean, J. ,2013)

5.2 Model Comparison: Cosine Similarity VS Topic Analysis Methods - Assessment Audit

To compare and quantify the recommended job list's performance between the two methods: cosine similarity and topic modelling with Jensen–Shannon divergence, we design a human assessment with one question: "Do you think this job fits the course you took before?". The answer values are (1) is related. (0) is hard to say, and (-1) is not related. Annex E is the assessment template.

6. Results and Analysis

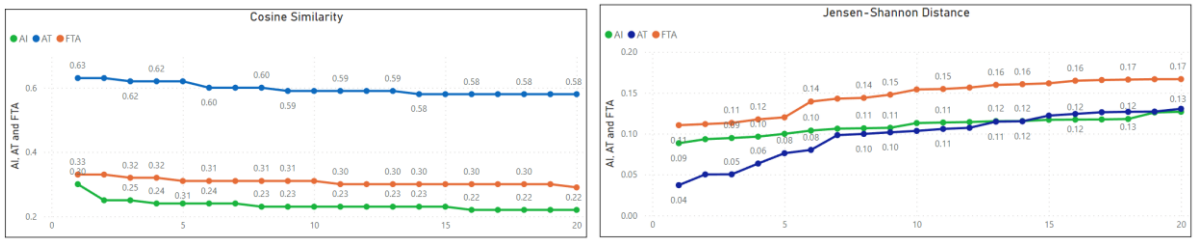
6.1 Resume Matching

Using the preprocessed text data, results show the top 20 recommended jobs by the cosine similarity score. Annex D contains a sample of our results for resume matching.

According to the results, the most relevant and matching jobs are all IT Jobs. It is reasonable because the sample resume is from a python applications developer. And the users can specifically obtain the similarity score of their resume and the job they are interested in. We also recommend the top 20 matching jobs to the users. It can help the users to have a higher probability of obtaining the offer for compatibility. Meanwhile, the users can also improve their resume to be more suitable for the job by the similarity score. (Mikolov, T., Chen, K., Corrado, G., & Dean, J. ,2013)

6.2 Matching Course to Jobs

To evaluate the model performance, we assume three MITB students are using our system. Each of them is from the Analytics track, Financial Technology & Analytics and Artificial Intelligence track. We input their core courses and get recommended jobs for each student. Annex B contains the core courses and recommended jobs result using cosine similarity. Annex C contains the core course and recommended job results using topic modelling and Jensen- Shannon Distance.



The cosine similarity score of the Analytics Track is higher than other tracks. Therefore, the performance of this Job matching system is better when matching the Analytics Track courses. The smaller the Jensen-Shannon Distance, the more similar the course and job descriptions are. The performance of this Job matching system is better when matching the Analytics Track and Artificial Intelligence courses.

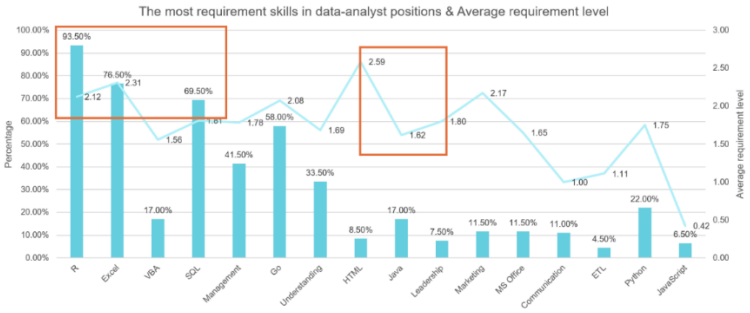
To compare and quantify the recommended job list's performance between the two methods, we invite four MITB students to generate the top 20 recommended jobs using two methods based on each student's learned courses and ask them to complete the human assessment. The Human Assessment result is in ANNEX H. The average score of cosine similarity is 6.75, which is higher than topic modelling. The performance of the cosine similarity method is better than the Topic modeling-JSD method.

6.3 Job Skills Extraction

Title	R	Excel	VBA	SQL	Management	Go	Understanding
Core Java Engineer London, Big Data	3	3		1			
Exposure Management & Reinsurance Analyst	2	3		2	2		
IT Business Systems Developer (C, NET, SQL)	2			1	1		
Core Java Engineer London, Big Data	3	3		1			
Data Analyst (SAS / SQL)	1	1	1	1		1	1
Service Desk Analyst Banking, Windows, Microsoft, AD,VMWare	3	3			3	3	3
QA Analyst (Automation)	2			2			
QA Analyst (Automation)	2			2			
Senior Business Analyst	2	2			1		
Teradata Consultant (Senior)	1	3		2	2	2	1
Reporting Analyst SQL Financial Services	2	1	1	3	3	1	
QA Analyst (Automation)	2			2			

Our text-processing algorithm converts job description information from a database into a structured skills table. When users view the job information, they can see the skill labels we extracted, with different colours representing different labels.

6.4 Job skills analysis - Top 10 skills for each position

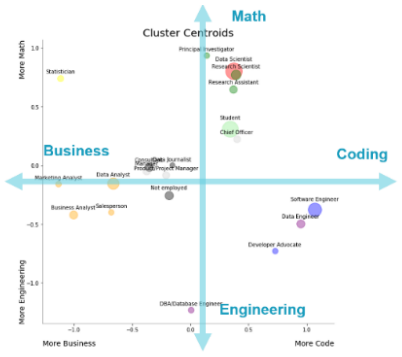


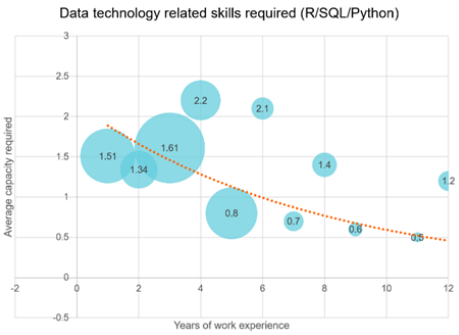
Insight 1: This graph represents the frequency of each skill in the job descriptions of all data analysts. From this graph, we can see a high percentage of R, Excel, and SQL. This means that these skills are the basic skills that need to be mastered for graduates looking for data analysts.

Insight 2: Programming skills such as go, java, HTML, and python appear infrequently, but their average requirements are higher than ordinary skills, so these skills need intensive practice for graduates.

6.5 Job skills analysis - Career comparison based on skill analysis

To make it clear for graduates to find a suitable job for them, we have provided career comparison charts for graduates. For example, for MITB students who want to find a data-related job. We use 4 dimensions to differentiate the available data-based jobs. These four dimensions are determined based on the skills required for each position.





6.6 Job skills analysis - Changes in career skills requirements

The average requirement for technical skills decreases as the number of years of experience increases, as many people begin to transition into managers, requiring more soft and managerial skills. And technical requirements tend to be more important for users in the early stages of career development.

7. Discussions and Gap Analysis

The LDA model has the potential to explore more effective dimensionality reduction methods. The topic distribution of the LDA model tends to bias towards words with high frequency, resulting in most words that can represent the topic being overwhelmed by a small number of words with high frequency. This may reduce the topic representation ability, and thus we can explore weighting methods to build a more effective LDA model.

8. Future Work and Conclusion

To improve our project, we can do a sentiment analysis based on those comments from colleagues in the recommended companies. Secondly, by "Word2vec" and "Glove", we would build a model on the actual linguistic relationship of the words such as the real meaning of the words, similarity with other words etc. Lastly, the application would be more useful and efficient if our UI interface can update automatically whenever the MITB courses or job descriptions change in the future.

In conclusion, our application can help users to apply for their jobs more efficiently.

9. References

Vayansky, I., & Kumar, S. A. (2020). A review of topic modelling methods. *Information Systems*, 94, 101582.

Jing, L. P., Huang, H. K., & Shi, H. B. (2002, November). Improved feature selection approach TFIDF in text mining. In *Proceedings. International Conference on Machine Learning and Cybernetics* (Vol. 2, pp. 944-946). IEEE.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modelling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11), 15169-15211.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.

Menéndez, M. L., Pardo, J. A., Pardo, L., & Pardo, M. C. (1997). The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2), 307-318.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint* arXiv:1301.3781.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12, 2825-2830.

10. Appendix

ANNEX A: Job Descriptions

Id	A unique identifier for each job ad
Title	A free-text field supplied by the job advertiser as the title of the job ads. It's the job title or role in general.
Full Description	The full text of the job ads is provided by the job advertisers. Whenever we see some values like '***s', these are values stripped from the description to ensure that no salary information appears within the descriptions.
Company	The name of the employer is supplied by the job advertisers.
Category	Which of the 30 standard job categories fits into it.

ANNEX B: Matching Course to Jobs by using Cosine Similarity: Top 20 Recommended jobs

	Analytics track	Financial Technology & Analytics	Artificial Intelligence track
Core Course	Data Analytics Lab Python Programming & Data Analysis Applied Statistical Analysis with R Data Science for Business*	Digital Banking & Trends Fintech Innovations & Startups* Data Science in Financial Services* Python Programming & DataAnalysis	Algorithm Design & Implementation Introduction to Artificial Intelligence Applied Machine Learning AI Planning & Decision Making
Job List	Learning and Development Business Partner Senior Finance Business Partner Business Intelligence Analyst …	Data Analyst – SQL / ETL Data Supply Chain Administration Data Quality Manager …	Data Management Consultant Business Intelligence (BI) Consultants Advisory Financial Analyst …

ANNEX C: Matching Course to Jobs by using Topic Modeling: Top 20 Recommended jobs

Track	Artificial Intelligence	Analytics	Financial Technology & Analytics
Core Course	Algorithm Design & Implementation Introduction to Artificial Intelligence Applied Machine Learning AI Planning & Decision Making	Data Analytics Lab Python Programming & Data Analysis Applied Statistical Analysis with R Data Science for Business	Digital Banking & Trends Fintech Innovations & Startups Data Science in Financial Services Python Programming & Data Analysis
Job List	Project Manager Business Change Financial Control Analyst ….	Health Research Analyst Commercial Finance Analyst BI Data Input Administrator …	Finance Analyst Demand Analyst Data Management Consultant …

	Unnamed: 0	Id	Title	FullDescription	.ocationRaw	LocationNormalized	ContractType	ContractTime	Company	Category	SalaryRaw	SalaryNormalized	SourceName	FullDescription_Raw	score
7676	113035	69731518	Software Development Process Analyst Bank, Ag...	software development process engineer banking ...	London	London	full_time	permanent	Information Technology Services	IT Jobs	From 45,000 to 55,000 per year	50000	planetrecruit.com	Software Development Process Engineer Banking...	0.46
2563	43323	68496296	Lead Technical Business Analyst	client capita uks leading business process out...	London City EC1A2	Central London	NaN	permanent	aap3	IT Jobs	50000.00 - 55000.00 GBP Annual	52500	jobserve.com	Our client Capita is the UK's leading Business...	0.46
3130	52222	68665951	Development Technical Lead/Project Manager – g...	new opportunity experienced development techni...	Woking Surrey South East	Woking	NaN	permanent	RWI	IT Jobs	From 35,000 to 45,000 per annum plus benefits	40000	totaljobs.com	A new opportunity for an experienced Developme...	0.45
7545	111551	69687078	Business Intelligence Developer	business intelligence developer microsoft sql ...	The City,London	London	NaN	permanent	Code IT Recruitment Ltd	IT Jobs	40,000-55,000	47500	jobsite.co.uk	Business Intelligence Developer (Microsoft SQL...	0.45
3642	57404	68685314	Business Analyst	valueworks medium sized award winning high gro...	Wigan Lancashire North West	UK	NaN	contract	Valueworks Limited	IT Jobs	134 per day flexible benefits package.	32160	totaljobs.com	Valueworks is a medium sized, award winning h...	0.44
5268	79665	69019049	Java Analyst Developer	job title java analyst developer salary locati...	Oxford, Oxfordshire	Oxford	NaN	permanent	CBSbutler	IT Jobs	40000 - 46000/annum Benefits	43000	cv-library.co.uk	Job Title: Java Analyst Developer Salary: Up t...	0.44
4216	63719	68704904	Business Analyst	business analyst required join medium sized aw...	Wigan Lancashire North West	UK	NaN	contract	Recruitment Genius	IT Jobs	35000 per annum	35000	totaljobs.com	A Business Analyst is required to join a mediu...	0.43
6997	104377	69572312	Java Analyst Programmer	java analyst programmer cuffley potters bar k ...	Hertfordshire Cuffley EN6 4	Cuffley	NaN	permanent	Absolute Solution	IT Jobs	35000.00 - 40000.00 GBP Annual	37500	jobserve.com	Java Analyst Programmer Cuffley, Potters Bar, ...	0.42
537	13285	66600493	New Business Development Telesales Training ...	established client specialist college central ...	London	London	NaN	permanent	Morgan Jones Recruitment Consultants	Sales Jobs	24000 - 33000/annum 24K plus Circa 8k OTE bonus	28500	cv-library.co.uk	Our established client is a Specialist College...	0.42
5563	82760	69042083	Java Analyst Programmer	java analyst programmer cuffley potters bar k ...	Potters Bar Hertfordshire South East	UK	NaN	permanent	Absolute Solution Limited	IT Jobs	From 35,000 to 40,000 per annum 35-40k plus bens	37500	cwjobs.co.uk	Java Analyst Programmer Cuffley, Potters Bar, ...	0.42

ANNEX D: Resume Matching Results

ANNEX E: Human Assessment

"Do you think this job fits the course you took before?"	Job 1	Job 2	Job 3	Job 4	…	Sum
Cosine Similarity						
Topic modeling-JSD						

ANNEX F: Topic words of all topics

	topic1-keyword	topic1-weight	topic2-keyword	topic2-weight	topic3-keyword	topic3-weight	topic4-keyword	topic4-weight	topic5-keyword	topic5-weight	topic6-keyword	topic6-weight
word1	SQL	0.005	sales	0.007	hr	0.016	test	0.006	data	0.004	marketing	0.005
word2	database	0.005	planner	0.006	generalist	0.005	marketing	0.005	hr	0.004	data	0.004
word3	repair	0.004	media	0.005	sales	0.004	digital	0.005	eCommerce	0.003	analyst	0.003
word4	server	0.004	jobs	0.005	partner	0.004	Europe	0.004	SQL	0.003	analysis	0.003
word5	data	0.004	digital	0.005	bp	0.003	sales	0.004	analyst	0.003	travel	0.003
word6	dynamics	0.003	subscription	0.004	er	0.003	agencies	0.004	graduate	0.002	SAS	0.003
word7	partners	0.003	data	0.004	reward	0.003	publishing	0.004	social	0.002	financial	0.002
word8	support	0.003	manager	0.003	people	0.003	new	0.003	requirements	0.002	sap	0.002
word9	systems	0.003	marketing	0.003	managers	0.002	testing	0.003	systems	0.002	direct	0.002
word10	service	0.003	LinkedIn	0.003	senior	0.002	European	0.003	developer	0.002	pelican	0.002
word11	ax	0.003	campaign	0.003	employee	0.002	adviser	0.003	process	0.002	market	0.002
word12	marketing	0.003	twitter	0.003	organisation	0.002	agency	0.003	support	0.002	insight	0.002
word13	sales	0.003	check	0.003	management	0.002	development	0.003	business	0.002	us	0.002
word14	career	0.003	graduate	0.003	business	0.002	marketmacildowie	0.002	software	0.002	finance	0.002
word15	desk	0.002	discuss	0.003	manager	0.002	tester	0.002	looking	0.002	position	0.002

ANNEX G: Course-Job Matching Example

Course input: “Digital Banking & Trends”, “Big Data: Tools & Techniques”, “Digital Transformation in Retail Banking Technology”, “Corporate Banking & Blockchain”, “Financial Markets Systems & Technology”		
Id	Job Title	Similarity
66744213	Data Manager	0.0733926
66773714	Refinery Planning Scheduling Optimisation Analyst	0.081499912
66925936	Head of Business Improvement and IT (Charity) to ****k	0.088282973
66983449	Senior Commercial Analyst	0.091433465

67099444	Financial Analyst	0.11493504
67804467	Credit Analysts at the Junior and Senior Levels	0.11630287
68242001	Financial Analyst	0.120930567
68685977	Energy Data Analyst, BMSi British Gas	0.121933259
68688901	Business Analyst	0.122746982
68693523	PRICING ANALYST	0.124150343
68800808	Head of Business Improvement and IT (Charity) to ****k	0.124151528
68824229	Business Continuity Coordinator (BC / DR Practices)	0.125807598
68824859	Business Insight Manager Field Based	0.128315464
69010492	Health Records Analyst	0.130164757
69018852	Business Relationship Manager	0.133935839
69089107	Insurance Business Change Project and Program Manager	0.134002596
69591297	Analyst	0.134064198
69599951	Graduate Financial Analyst	0.135287702
69669491	Insurance/Actuary Analysts	0.136219621
69805559	Business Process Improvement Lead Edinburgh or London	0.136880413

ANNEX H: Human Assessment for Recommended Jobs to compare methods performance

	GSF	LSY	XSY	ZZY	Average
Cosine Similarity	12	4	12	-1	6.75
Topic modeling-JSD	3	1	8	12	6