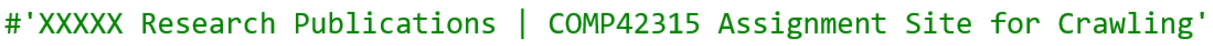
A web-crawling practice on analysing Hubert Shum’s Curriculum Vitae

Introduction

Web-crawling technique has commonly applied on high throughout information platform that numerous data generate in every second. This information retrieval assists tech-rookies to automatically collect, filter, pre-process, and analyse the in-page documents through a systematic and method-followed manner (Kausar et al, 2013). In the following paraphs, a website about Hubert Shum’s publications (2021) has been written in error-free html format, becoming as the question material for figuring out the connections of its in-text information inside.

Question One:

The basic idea of crawling all unique sub-webpages was the concatenation of a global father link and numerous children links before excluding the duplicated information by their link similarity. Because identical layouts were illustrated under each topic, 15 distinctive local links are collected as formatting in identical hyperlink tag and class name, associated with their topic information crawled by similar method on similar positioning span tags, receiving 152-row available samples. Effective observation reduces the difficulty of obtaining information: each original title shapes similarly while only the first phrase is required (Fig.1), which is appropriate for reversed calling that counts letter backwords. Then father link and children link combined together for outside visiting, being stored into a dataframe with corresponding topics. drop_duplicates function was applied by selecting the first-occurred hyperlinks as keys to remove the other duplicates out from the columns, resulting in 110 samples after re-ranking its index sequences.



```
#'XXXXX Research Publications | COMP42315 Assignment Site for Crawling'
```

Fig.1 The Relationship between Original Topic Shape and the Key Factors Inside (XXXX is the key

Question Two:

The solution of Q2 is to rearrange Q1’s results by adding multiple topics and crawling the other paper-related information for being deposited into CSV files.

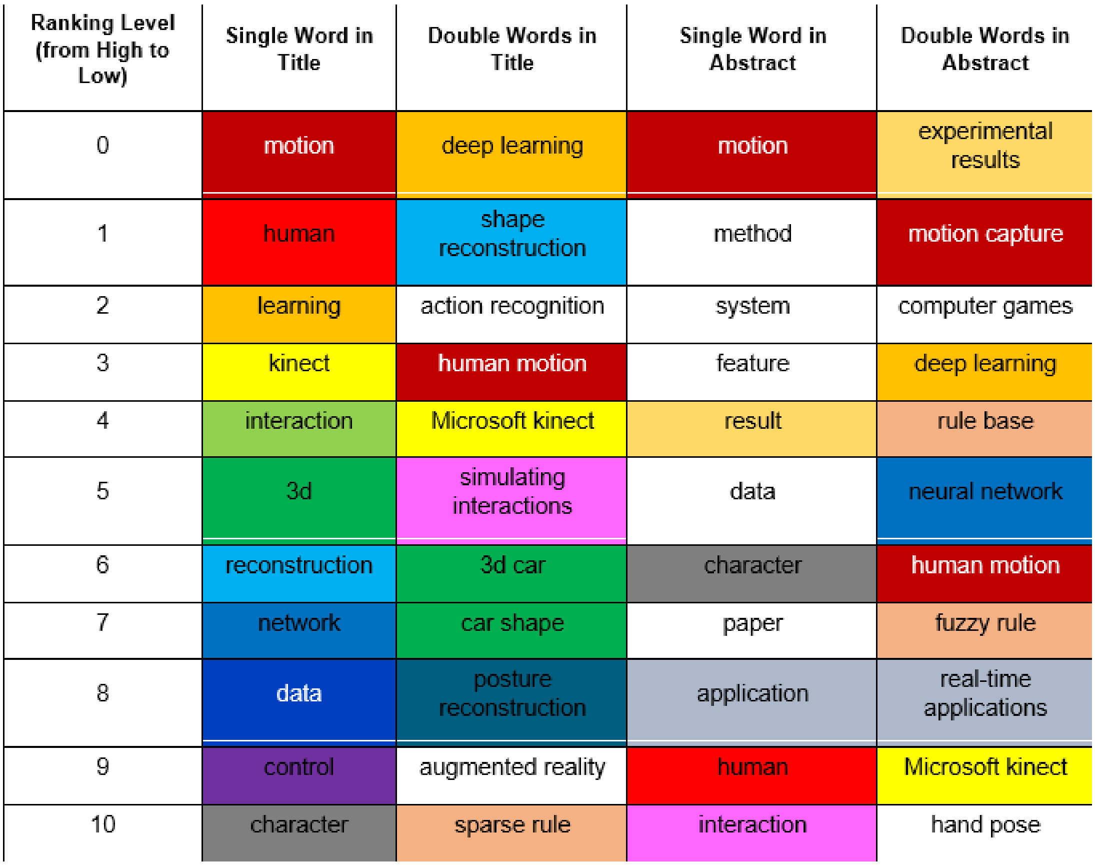
Save a new version of Q1’s answer without dropping any duplicated webpages. Connected it with the old version through merge operation, the second topic of publications was discovered on the view of universal set after setting its key on unique links, then the only third topic was revealed through the same method. All of them were aggregated as the new elements of Q1’s dataframe. The rest information was extracted as additional columns, includes paper’s titles, abstracts, journal names, impact factors, citation numbers, its links for download, DOI details, and the YouTube instruction. Thanks to find_all function that most of the required elements could be derived by counting their ranks after the retrievals of the same attribute values, under a loop search of publication websites. Though there were some none-value records due to occasional missing data, a pair of try and except instructions guarantee the only the meaningful information was reserved. A typical text-processing problem occurred during the separation of impact factor and citation number since they were written into one sentence. Regular expression divided these two features through distinguishing the word difference in both the first three letters of them and the ending hash sign. Furthermore, general observation found that both impact factors and citation number have length limits, which is beneficial for keyword conditions. Finally, the entire dataframe was transformed into a CSV file that containing a matrix of 110 rows and 19 columns.

Question Three

Inherits Q2's answers, titles and abstract material are filtered to investigate the 100 most popular words defined as the most frequent in this case. Due to the variation in word length, the importance weights of each word in the two elements differ significantly, which can be analysed separately. In addition, words and two-word phrases employ distinct logic in their research, and eventually these differences formed into a 4*25 matrix design.

The first step was to tokenise the single word in the features to break the sentences into pieces, then excludes the meaningless common words out from the prepared word list through applying stopword dictionary, which generates the basic materials of further analysis. Subjects, prepositions, and special symbols are the objects that stopword concerned since only nouns and adjectives make sense. Chopped words should be lowercased that otherwise they would be considered differently from their uppercase brothers. Then the two-word phrases were selected by n-gram model when its n equals 2, and being ranked for their occurrence inside the documents. Nevertheless, the steps of choosing word were more complicated that research materials should be lemmatised into verbs and nouns because adjectives or adverbs contain less information. Additionally, even in the perspectives of verbs, use of tenses would confuse the algorithm by classifying them into different blocks of ranking.

Overall, the selected words showed an intensive similarity at 77.5% that over three quarter of words appeared multiple times whether on titles or abstracts in terms of single words or word groups (Fig.2)



Ranking Level (from High to Low)	Single Word in Title	Double Words in Title	Single Word in Abstract	Double Words in Abstract
0	motion	deep learning	motion	experimental results
1	human	shape reconstruction	method	motion capture
2	learning	action recognition	system	computer games
3	kinect	human motion	feature	deep learning
4	interaction	Microsoft kinect	result	rule base
5	3d	simulating interactions	data	neural network
6	reconstruction	3d car	character	human motion
7	network	car shape	paper	fuzzy rule
8	data	posture reconstruction	application	real-time applications
9	control	augmented reality	human	Microsoft kinect
10	character	sparse rule	interaction	hand pose

Fig.2 Word Similarity in Both Titles and Abstracts in Terms of Single Word and Two-Word Phrases (Similar words painted into the same colours)

Question Four

A series of processes could answer Q4. Count each publications' topic occurrences primarily and then select the Top Three topics of occurrence as the most popular topic in academic. Owing to replicated topics showing as '2nd topic' or '3rd topic' in Q2's dataframe, the frame slicing skill was employed to find the common of conditions in each column. Fig.3 depicted the sum of occurrences under each topic segment, that 'Character Animation', 'Motion Analysis', and 'Interaction Modelling' were preferable welcomed.

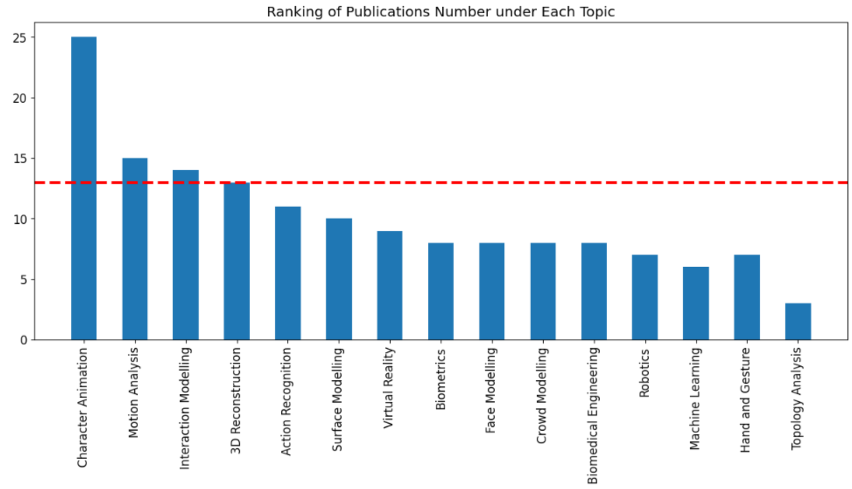


Fig.3 The Sum of Occurrences of Each Topic

The research scale was shrunken, it is essential to elaborate who were the top three authors have written the most papers. Counter function summarised the desired results in terms of dictionary retrieval that only four names were revealed. An interesting finding was that both ‘Hubert P. H. Shum’ and ‘Edmond S. L. Ho’ were the top-three authors in all chosen topics. Therefore, the co-author relationships were built around these four people under three topic themes through counting the other authors’ occurrences. Because participants who only collaborated once have less academic influence, these participants (usually students) were removed from the lists. Fig.4-6 illustrated that ‘Hubert P. H. Shum’ has participated the most article writing in ‘Character Animation’, usually with ‘Taku Komura’ and ‘Edmond S. L. Ho’, but rare interactions occurred between these two co-authors.

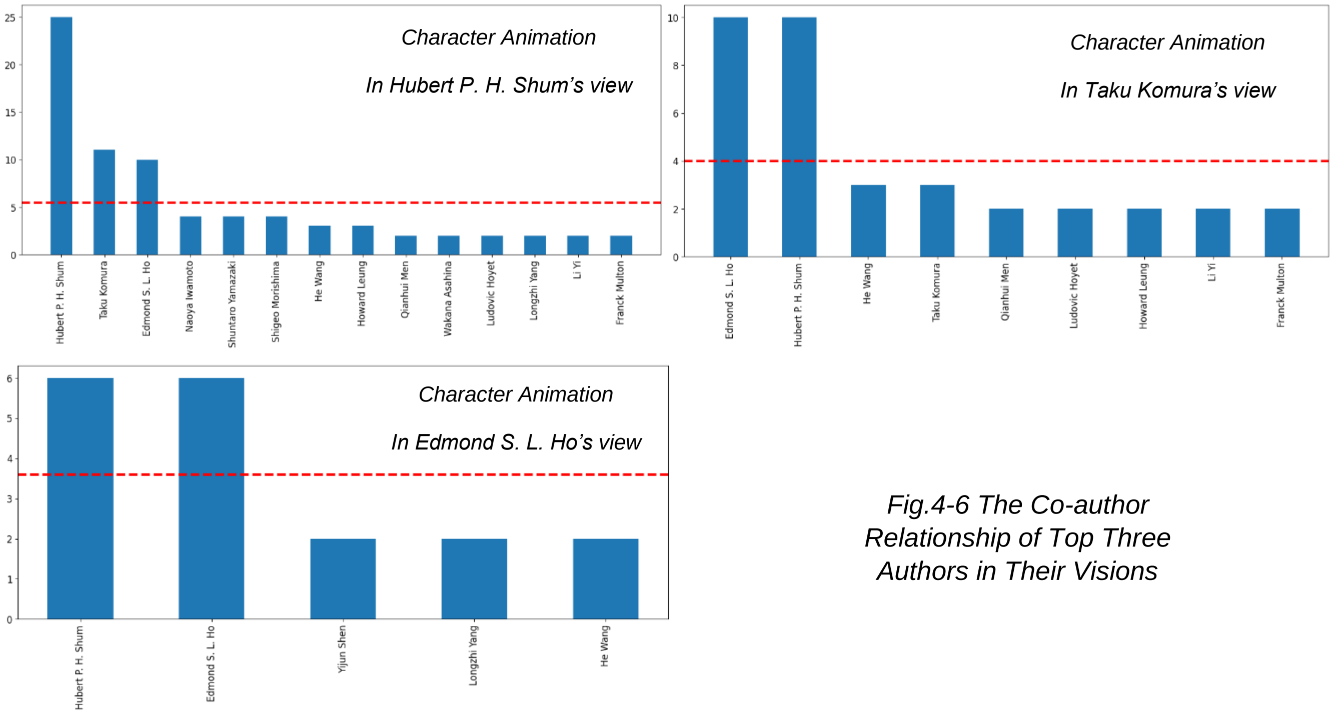


Fig.4-6 The Co-author Relationship of Top Three Authors in Their Visions

Two other topics followed the same pattern that Hubert participated into all co-authors’ projects (Fig.7-12). The only difference was that the number of articles under those two topics was much smaller than the above topic’ as less attentions from academics. It was a tendency that less co-authors’ names appear that the sum number of collaboration fell, implying there was a finite number of cooperation in fixed workgroups, which is reasonable that it is difficult for project team members to accomplish multiple other papers in a short time.

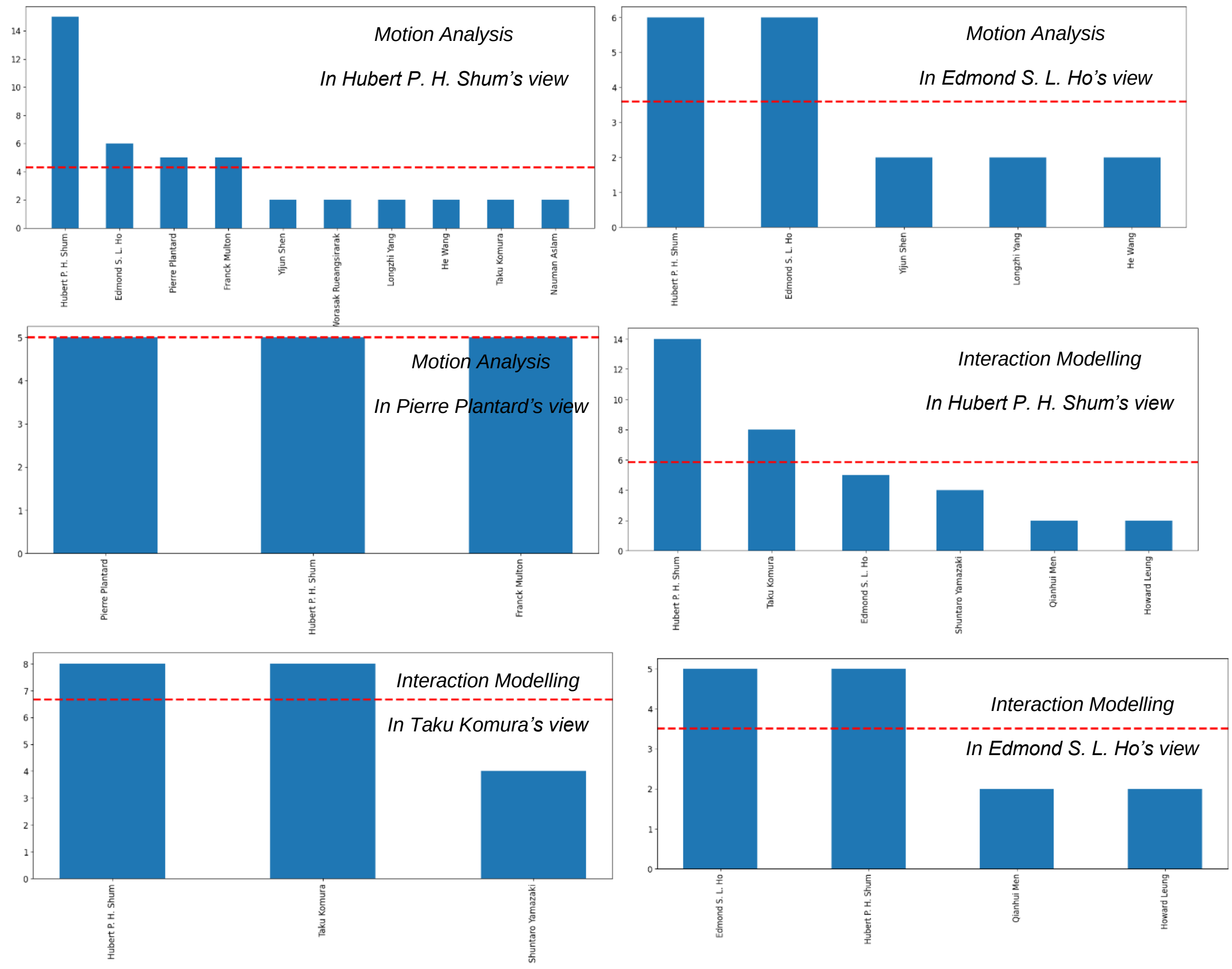
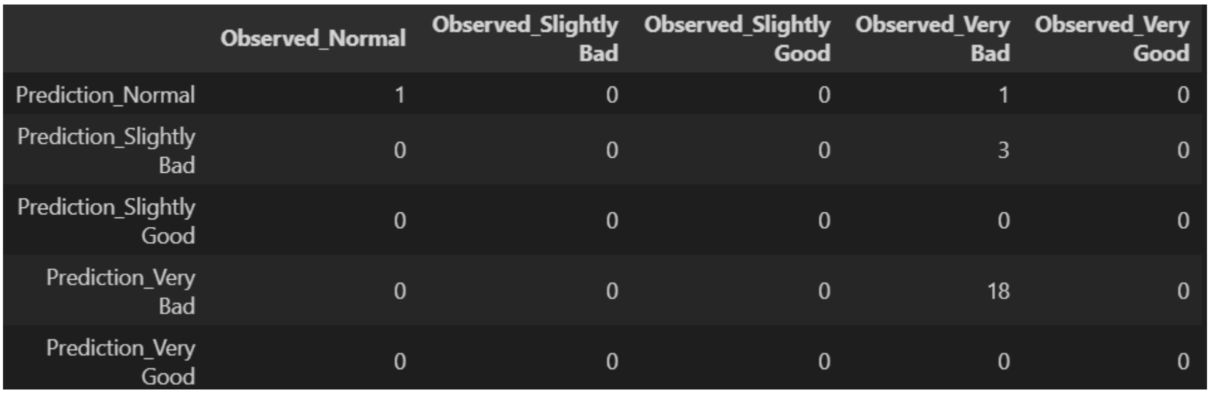


Fig.7-12 The Co-author Relationship of Top Three Authors in Their Visions

Question Five

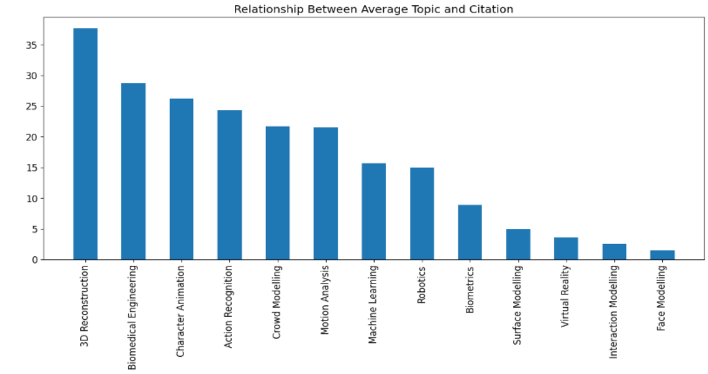
This task aims at figuring out the connections among citation numbers and the other relevant features displayed in Q2's dataframe, while could be measured by KNN model, a typical classifier labelling data through neighbours' label from a certain distance, could explains the classification accuracy through feature analysis. According to Onodera and Yoshikane's research (2015), the importance of researched fields, first author's achievements (which is the author named at the first place), and reputation of journal published had highly significant influences on the overall citations. Therefore, the first-shown topic, first author, second author (In consideration of academic experts sharing their positions to students), and journal details were selected as explanatory variables. Owning to the categorical output assumption of KNN model, citation numbers was split into five subsets from 'Very Bad' to 'Very Good', with equal distance at 25 since the maximum number founded in Q2 was 125. Then all explanatory and response data were separated in 75% and 25% in training and test sets individually for accuracy validation, which is beneficial to examine the model's performance in clean dataset. During the model construction, it is empirical to set K to the root of sample size to avoid noise from over-short distance, which was 10 in this case proofed by for loop experiment. The accuracy result of the entire dataset indicated that 74.1% features were correctly classified, but mainly concentrating on ‘Very Bad’ class since its domination on occurrences. This biased prediction disclosed that the KNN model was overfitting when zero values appeared inside the other features of the confusion matrix (Fig.13). Furthermore, it is impossible for KNN classification to distinguish the feature importance as it weights them equivalently inside a hyper-dimensional space (Thirumuruganathan, 2010).



	Observed_Normal	Observed_Slightly Bad	Observed_Slightly Good	Observed_Very Bad	Observed_Very Good
Prediction_Normal	1	0	0	1	0
Prediction_Slightly Bad	0	0	0	3	0
Prediction_Slightly Good	0	0	0	0	0
Prediction_Very Bad	0	0	0	18	0
Prediction_Very Good	0	0	0	0	0

Fig.13 Confusion Matrix of KNN Model

Hence how the selected features directly influence on citation numbers would be presented into one-by-one visualisations. Group_by function has summed up the citation number under each feature, while the counts of feature’s item would become as the divisor for alleviate the overweighting problem when higher citations happened in higher occurrences. In the one-by-one visualisations between topic-citations and journal-citations, ‘3D Reconstruction’ and ‘IEEE Transactions on Cybernetics’ were quoted the most, with an over 15% gap from the next best option (Fig.14-15). And the graphs of first and second author have found some fruitful academic stars when students only cooperated once were excluded for coincidence, while the sophisticated authors only achieved their scores at around 40 (Fig.16-17).



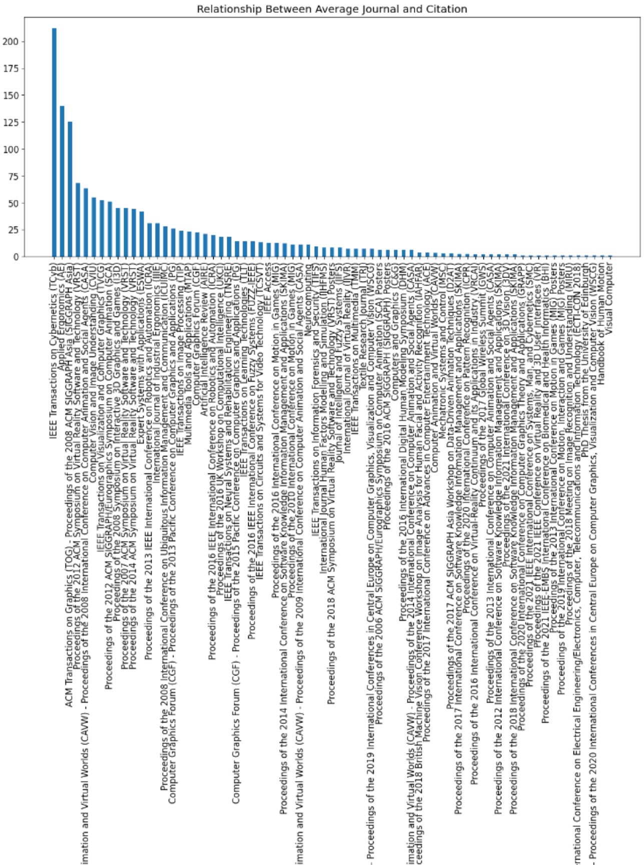


Fig.14-15 Relationship between Average Topic/Journal Weight and Citation

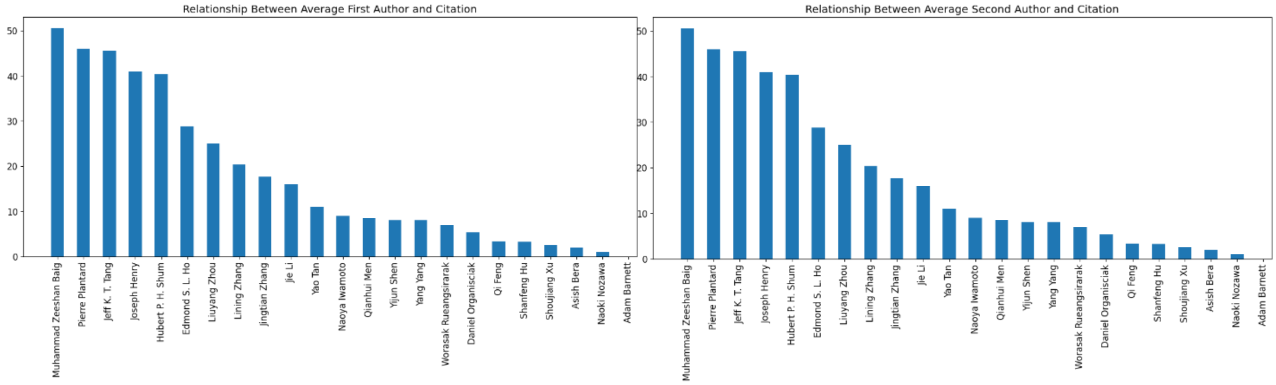


Fig.16-17 Relationship between Average First/Second Author Weight and Citation

Conclusion

In conclusion, 3D reconstruction and IEEE Journal might be most influential factors compared with other rivalries in their separated groups, while the four academics on writing the most papers were somewhere in the upper middle of influential importance, suggesting that high productivity was not the only method of increasing citations. However, the importance deviation could not be measured through the methods mentioned that multiple regression might optimise this solution in further analysis.

Reference

Kausar, M.A, Dhaka, V.S., and Singh, S.K. (2013) 'Web Crawler: A Review', *International Journal of Computer Applications,* 63(2), pp. 31-36.

Onodera, N. and Yoshikane, F. (2015) 'Factors Affecting Citation Rates of Research Articles', *Journal of the Association for Information Science and Technology,* 66(4), p. 739–764.

Shum, H. (2021) *COMP42315 Assignment Site for Crawling.* Available at: https://community.dur.ac.uk/hubert.shum/comp42315/ (Accessed: 16 Feburary 2022)

Thirumuruganathan, S. (2010) *A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm.* Available at: https://saravananthirumuruganathan.wordpress.com/2010/05/17/a-detailed-introduction-to-k-nearest-neighbor-knn-algorithm/ (Accessed: 17 Feburary 2022)