USING TOPIC MODELLING TO ANALYSE BUS ROUTE DATA

HS KOEN[1], J CORNELIUS[1] and R OOSTHUIZEN[2]

[1]Council for Scientific and Industrial Research (CSIR), Meiring Naude Road, Brummeria 0001, Pretoria, South Africa
Email: hkoen@csir.co.za; jcornelius1@csir.co.za
[2]Department of Engineering and Technology Management, University of Pretoria, c/o Lynnwood Road and Roper Street, Hatfield, South Africa
Email: rudolph.oosthuizen@up.ac.za

ABSTRACT

The advent of the fourth industrial revolution and the need for connectedness have increased both data availability and quality. This data surge can also be seen in the transport and mobility industry. Anything from onboard global positioning system interfaces to vehicle trackers and wearable technology for passengers and drivers provide access to more data as an untapped source of valuable information and insights to many stakeholders. Topic modelling is traditionally used to structure and interpret text data from a large corpus of documents. In this paper, patterns in bus route data collected over several months by the onboard Global Positioning Systems (GPSs) of buses travelling in Gauteng and the Northwest province are analysed. Since topic modelling is traditionally used on text documents, the bus route coordinates had to be converted into a form readable by the algorithm. This is an ongoing project, but analyses thus far show that the most important terms per topic correspond to key nodes in city centres and points of interest where routes overlap. This information may be used in city planning to optimise the system of bus routes, terminals, and nodes. Organisations may also use this information for business development and job creation.

1. INTRODUCTION

As the world becomes more connected and sensors become more prevalent, the flow and availability of data increases. In recent years, the transport environment has seen an increase in data availability with the advent of GPS devices, onboard sensors, traffic cameras, and even smartphones and smartwatches worn by users while in transit (Zhou, 2021). This paper aims to showcase how machine learning techniques may be used to find interesting patterns in mobility data. Topic modelling was implemented to find patterns in bus route data collected on a locally developed platform over several months in South Africa by bus operators and users of privately-owned buses.

Topic modelling is an unsupervised machine learning technique used in Natural Language Processing (NLP) that automatically classifies text data into different topics to extract semantic information from the data (Eker et al., 2019; Jia et al., 2018). It is called unsupervised because the algorithm does not require data that has been classified or tagged previously. The text data forms part of a collection of documents called the corpus, where each document in the corpus consists of several topics, and each topic consists of several keywords (Agrawal et al., 2018). Automated text processing can cluster, classify, summarise, or categorise many text documents. Several topic modelling algorithms exist, including Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorisation (NMF), Latent Semantic Analysis (LSA), Parallel Latent Dirichlet Allocation (PLDA) and Pachinko

Allocation Model (PAM) (Manthiramoorthi, 2021). Topic modelling's uses include the field of data analysis of social media posts (Nolasco, 2019), emails (Lossio-Ventura, 2021), chats (Chen, 2019), and open-ended survey responses (Finch, 2018). These topic modelling characteristics make it ideal for processing the huge amount of bus route data to detect patterns. Manually processing all the data will be impossible and may not make the latent patterns visible.

LDA, a popular topic modelling algorithm, was chosen for this project. The name implies that the algorithm uses the Dirichlet distribution, and the term "latent" means that topics are hidden and yet to be discovered (Chen et al., 2019). The LDA algorithm assumes that a document contains a set of latent issues, as intended by the specific selection of words chosen by the authors (Kim & Kang, 2018). The algorithm then provides a generative statistical model in which unobserved groups explain the similarity of the data. The model learns the distribution of topics in each document, with their associated word probabilities, to identify major thematic clusters from an extensive corpus of text documents, usually beyond human capacity (Suominen & Toivanen, 2016; Hecking, 2019; Maier et al., 2018). The algorithm processes the document-term matrix to probabilistically form document-topic and topic-word pairs to extract the number of topics defined by the analyst.

The LDA algorithm is called a "bag-of-words" model because the order of the words does not matter. The LDA algorithm's output includes topics present in the corpus of documents, the terms that define a topic, and the allocation of topics per document (Chen et al., 2019; Sethasathien & Prasertsom, 2020). However, domain expert knowledge is still required to analyse, name, and describe each topic.

This paper applies topic modelling through LDA to process GPS data collected on bus routes over a long period and a wide geographic area. The extracted topics may be used to analyse the various routes to understand the mobility patterns of commuters in the region. The next section will discuss the related work in this field, while the process of applying topic modelling will be discussed after that. Section 4 will then present and discuss some key outputs and results of the topic modelling.

2. RELATED WORK

Gholampour et al. (2020) apply LDA based topic modelling to find abnormal traffic patterns using speed camera data. Topic models are applied to the data to extract monthly and annual traffic patterns. According to the authors, this method can be applied to urban traffic with a successful detection rate of 99% of unusual conditions (Gholampour et al., 2020). The authors use the LDA algorithm to detect anomalous events by comparing topic proportion vectors for different documents. A significant difference between the two vectors could signify an anomalous event (Gholampour, et al., 2020). The "hour in day" and "Location ID" were included in the word format to locate the topic corresponding to the hour of the day and the location for a specific topic. This is similar to what was done with the GPS coordinates in the bus route data. The authors state that topics do not need to have "clear human-readable meanings" but that cases can be related to certain situations (Gholampour et al., 2020).

Tang et al. (2018) visually analysed traffic patterns to help shop owners choose suitable locations for setting up new businesses. Usually, this would be performed using population statistics, person flow calculations, etcetera. All these approaches take time and require some effort, and the authors state that it is difficult to obtain the necessary data to conduct the tests. Topic modelling combined with the traffic volume information help in choosing

optimal zones. The distribution and semantics of topics from time, space, and points of interest are visualised in three different views, including the LDA, general traffic flow, and attribute views (Tang et al., 2018). The LDA view displays the spatial and temporal distributions of topics integrated with the semantic information from the points of interest. Merging the topics with traffic features helps discover the best possible areas to minimise the search range. It visualises traffic flow through these key areas as a graph, and the attribute view, in turn, is developed to display numerous spatiotemporal traffic attributes (Tang et al., 2018).

Chu et al. (2014) convert location data to street names and then perform LDA topic modelling for semantic analysis. The authors concatenate the grid index and time and form words that maintain spatiotemporal information. However, Tang et al. (2018) extract routes from the records. Each route corresponds to a document, and all routes together form the corpus of documents. The GPS coordinates are converted to words for the LDA processing. The location of data points can be mapped to a grid to simplify the problem and reduce the range of words. The authors perform tests where they use between four and ten topics, and their results show that the best performance is achieved when five topics are chosen (Tang et al., 2018). The LDA-based topic modelling infers the probability that a topic belongs to a route or the probability that a word belongs to a topic between the routes and the topics and between the topics and the words (Tang et al. 2018). Key nodes are defined as a "point visit" by considering the number of GPS coordinates in a grid cell, semantic correlation, and specific times of the day.

Tang (2021) prepares trip data for the LDA the same way as for text data. GPS coordinates are rounded to the third decimal to reduce the size of the dataset. After applying topic modelling, the words are converted back to GPS coordinates and visualised (the top 50 terms per topic) on the map. The model reveals a diverse set of route types that separate the data into distinct groupings.

3. APPLYING TOPIC MODELLING TO BUS ROUTE DATA

The data is collected from an in-house platform for shared awareness and integration that uses web and mobile technology, called Cmore (Cmore, 2021). The bus route data is logged by roughly 190 passengers using the Cmore app on their smartphones and is collected by the CSIR's Smart Mobility group. The passengers travel on several buses and a number of different routes, but this information was not available in the dataset used.

The GPS data is logged automatically on a user's phone in the background, and a data point is logged every few seconds. The dataset contains many data points, which is enough for the LDA algorithm to extract topics. Each bus route datapoint contains a timestamp and a client ID associated with the GPS coordinates. This metadata is used to organise the GPS coordinates by route. These routes are stored as lists and written to a comma-separated (CSV) file and serve as the input to the topic modelling process illustrated in Figure 1. Kindly note that the abbreviation "Coord" in Figure 1 refers to "coordinates".

The topic modelling process loads the CSV files (the GPS coordinates and route data). After that, a Massachusetts Institute of Technology (MIT) word list containing 10,000 words is loaded; words from this list are then concatenated into a larger wordlist containing 1.5 million words (Price). The route data is then filtered to remove routes that contain less than 1,000 GPS coordinate pairs. Any repeated GPS coordinate pairs in a route are removed to reduce noise from stationary buses or buses idling at traffic lights. A unique list

of GPS coordinates is then constructed and each GPS coordinate pair in this list is further assigned to a word in the concatenated MIT word list. This is done to construct a dictionary that maps GPS coordinates to words and, for later use, an inverse dictionary that maps words to GPS coordinates.

The route data is then traversed, and each GPS coordinate is replaced with a word using the GPS coordinate-to-word dictionary. This procedure is performed since the topic modelling using LDA requires words rather than GPS coordinates. Subsequently, this converted route data is then vectorised and passed to the Sklearn LDA function in Python to determine topics based on this new word list. The number of topics was chosen to be 10, 15, and 20 based on the initial visual analysis in QGIS (Quantum Geographic Information System). QGIS is an open-source geographic information system that can create, edit, view, and analyse digital maps and location data (QGIS, 2021).

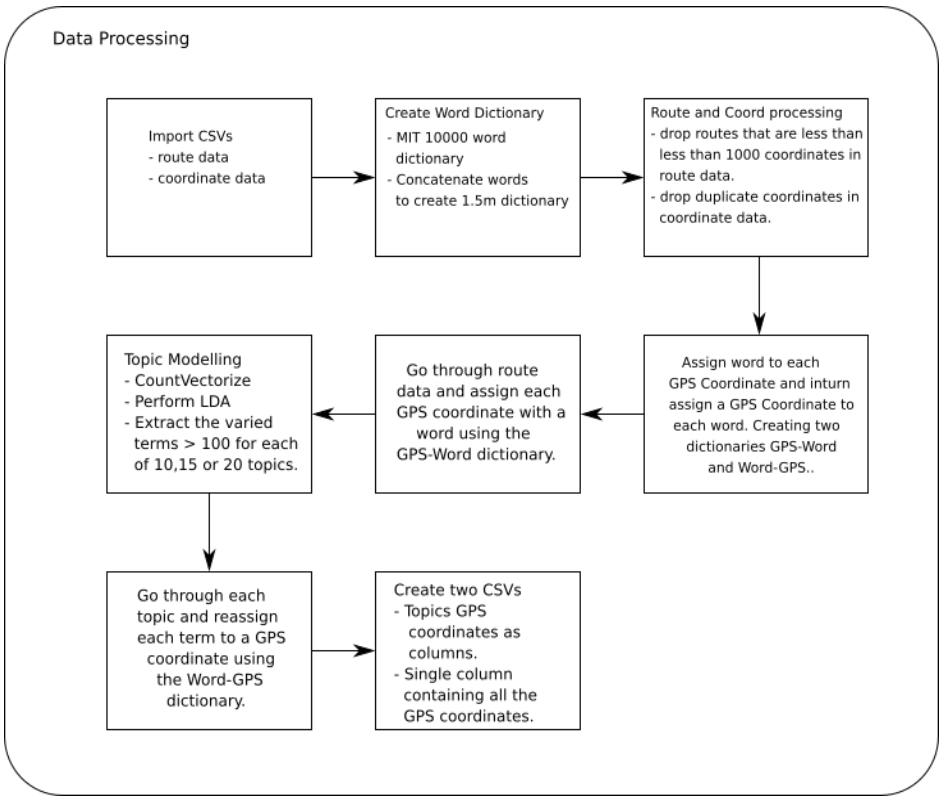

Figure 1: Bus route data conversion

The number of terms used in each topic exceeds 100 until repeat values are detected in the Document-Term matrix (DTM). The DTM is a matrix that maps all the terms to all the documents in the corpus, and it describes the frequency of terms that appear in each document. Each matrix row represents a document, and each column represents a word frequency. Each topic is iterated through, and each concatenated word is replaced by GPS coordinates using the word-to-GPS coordinate dictionary. The GPS coordinate data are then plotted in QGIS.

Unfortunately, choosing the number of topics is not an exact science in topic modelling. The most suitable number of topics must be chosen by evaluating the interpretability of the different visualisations. The modelling outputs must make sense to the researchers and improve their understanding of the data. When the number of topics approaches the number of data points, there will be a term for almost all data points, but that is futile as no patterns will arise; you will merely see all the data points again. In this case, topic values of

10, 15, and 20 were visualised. It was decided to use 20 topics as it yielded the most clusters of topic terms across the map with reasonable interpretations.

4. RESULTS AND EVALUATION

Spatial patterns in bus route data was collected over 17 months in the Gauteng and Northwest provinces of South Africa. The GPS coordinate data and map metadata were visualised to extract topic features. Map metadata could be bus stops, traffic lights, routes, and areas of interest. The routes with a distance greater than 1,000 coordinate pairs were plotted using QGIS, as seen in Figure 2. The routes are shown in black and follow roads. The geographic area of study was the area covered by this bus company which, as can be seen from Figure 2, is the northern most part of South Africa.

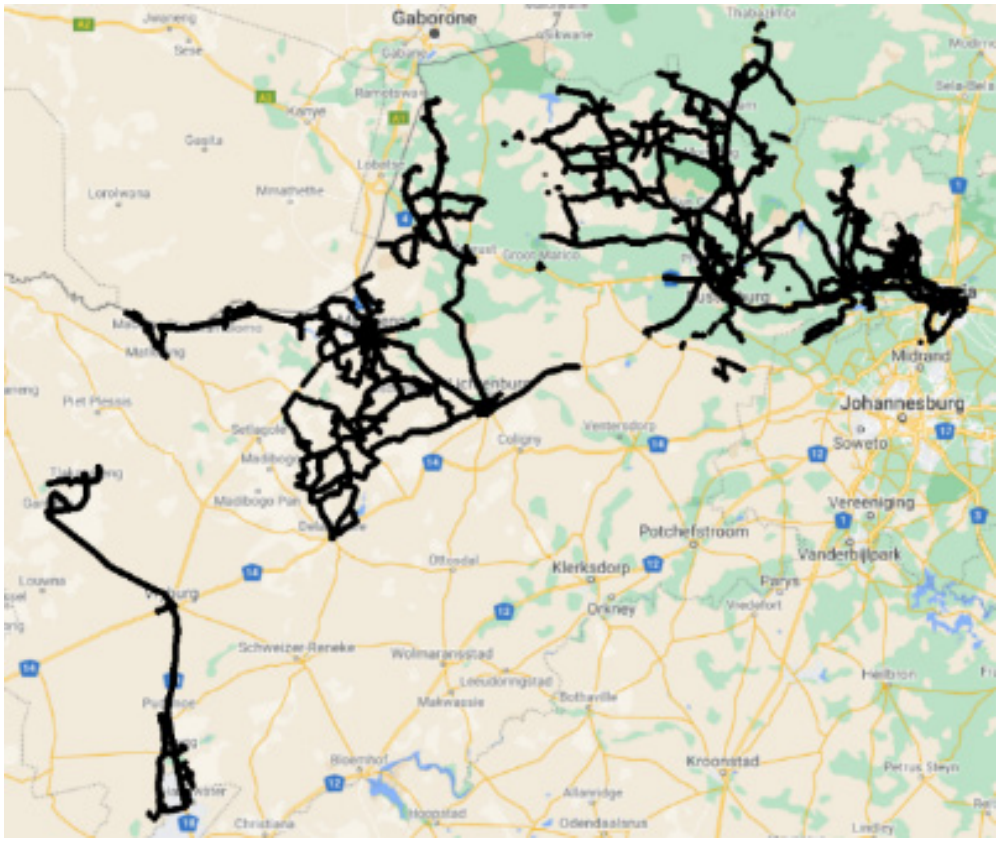

Figure 2: Bus routes greater than 1,000 coordinate pairs in length

The result of the LDA topic modelling and the geographical distribution of the 20 topics is shown in Figure 3. Most points appear on a road, while some occur in groups indicating routes. Others occur in areas indicating popular bus stops, and a few occur in isolation. These plots and digital map layers can be used to find connections to landmarks, geographical features, and traffic/routes to infer useful information. The number of topics, and the number of terms per topic, make it difficult to distinguish patterns at first glance.

All 20 topics will not be shown in this paper for brevity, only the most interesting cases. Figure 4 illustrates Topic 0, which clusters around small towns and settlements such as Moubane, Mmatau, and Bapong. This could indicate that buses either collect people from their homes to go to work, or drop them off at home after work, or both.

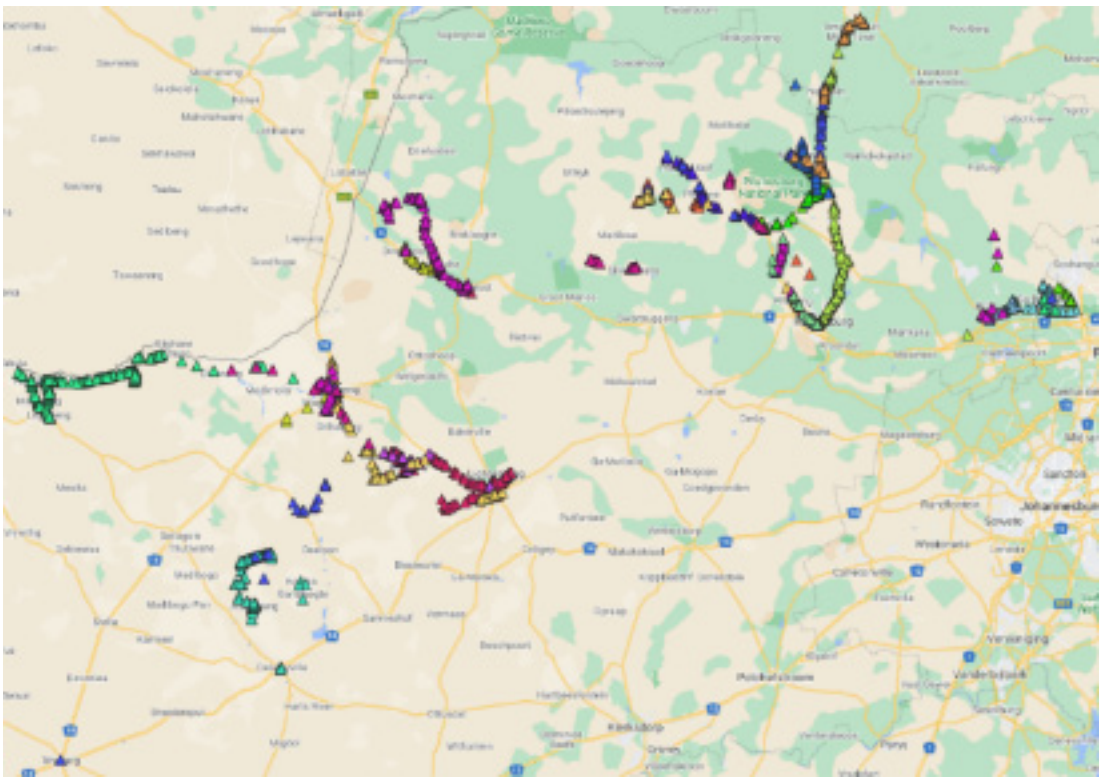

Figure 3: Twenty topics displayed geographically

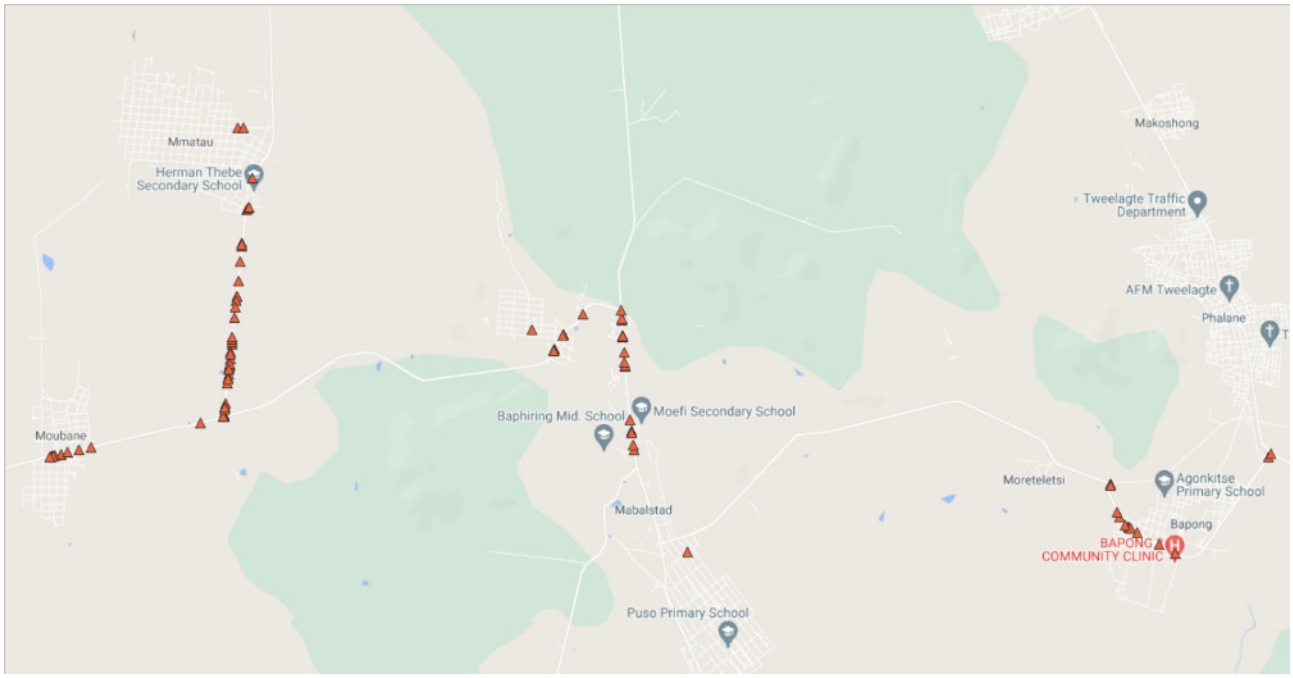

Figure 4: Topic 0

Topic 1, shown in Figure 5, is concentrated north of Pretoria in Ga-Rankuwa and Rosslyn. Most of the term clustering occurs in a new part of Ga-Rankuwa, the large roads connecting Ga-Rankuwa to Rosslyn (Molotlegi road) and the M21 through Ga-Rankuwa (Lucas Mangope drive and Kgware road).

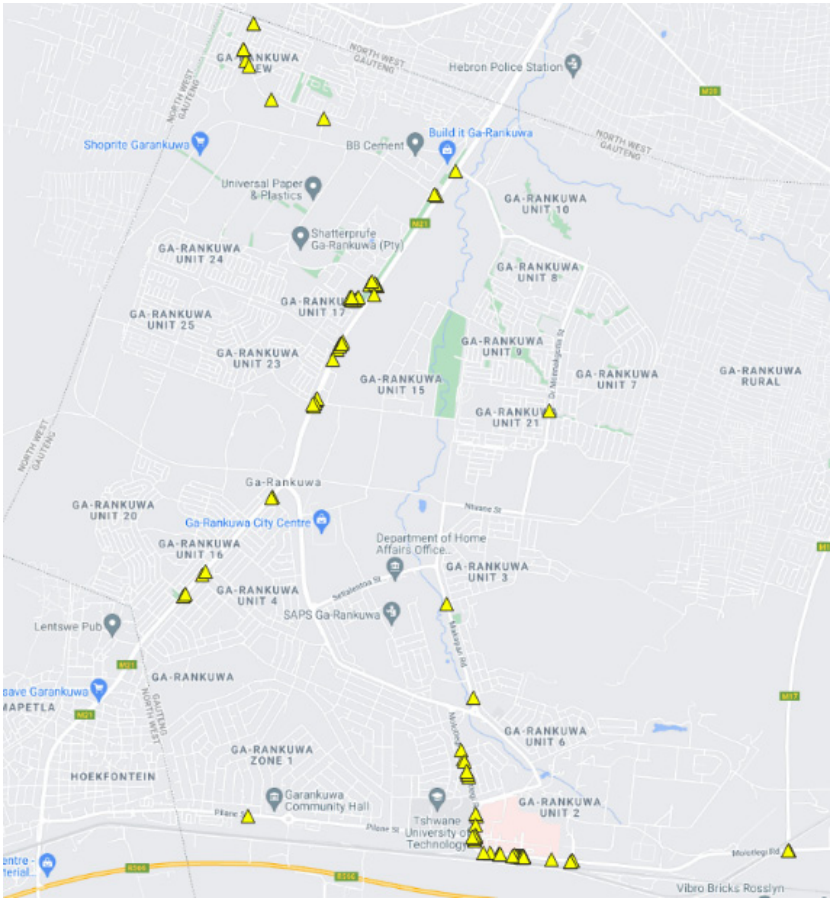

Figure 5: Topic 1 - Ga-Rankuwa

Topic 4 in Figure 6 focusses on Lichtenburg. There is a clustering of topic terms in nearby settlements, Itsoseng and Bodibe, and on the road connecting these settlements to Lichtenburg. This could imply that people who live in these settlements work in Lichtenburg. In Lichtenburg, the topic terms form an oval around the city centre. According to the terms in Topic 4, public transport prefers to drive around Lichtenburg rather than going through the town. To someone not familiar with the town, it might make more sense, just by looking at the map, to drive through the town rather than around it.

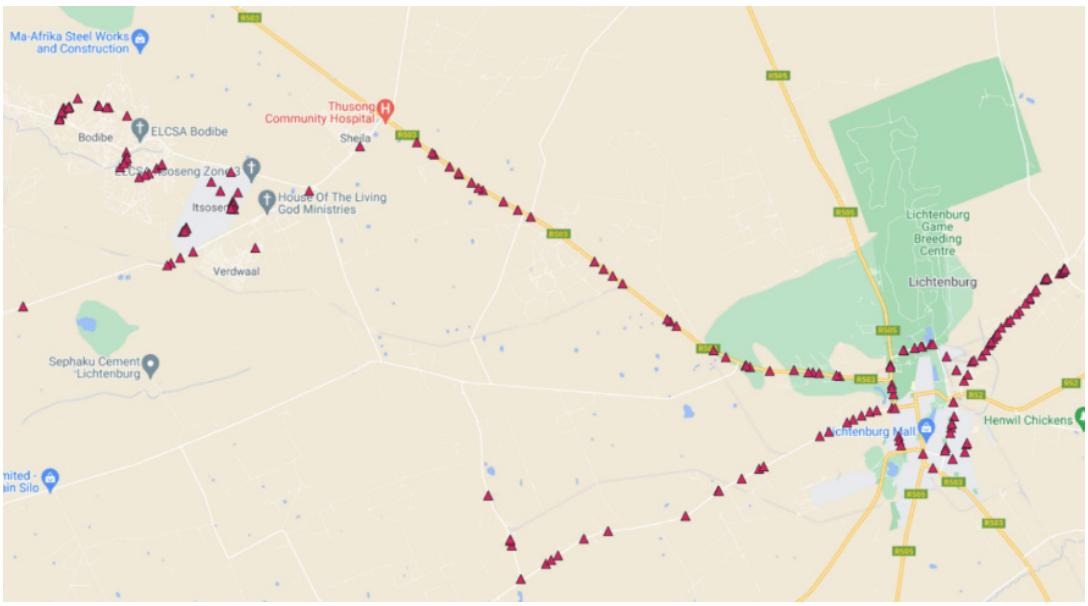

Figure 6: Topic 4 - Lichtenburg

Figure 7 showcases Topic 8, located between small settlements along the border between South Africa and Botswana. Clusters of points occur at Tshidilamolomo and Semashu. These settlements are larger than the others; therefore, they should have more topic terms. More topic terms mean that there could be more movement, or more people moving, in that area.

Returning to Topic 1 in Ga-Rankuwa, Figure 8 zooms into the industrial area of Figure 5. Investigating the location of the clusters utilising Google Maps' Streetview revealed an auto electrician situated at that corner. This may imply either the bus passengers work there, or that the buses are taken there often for repairs.

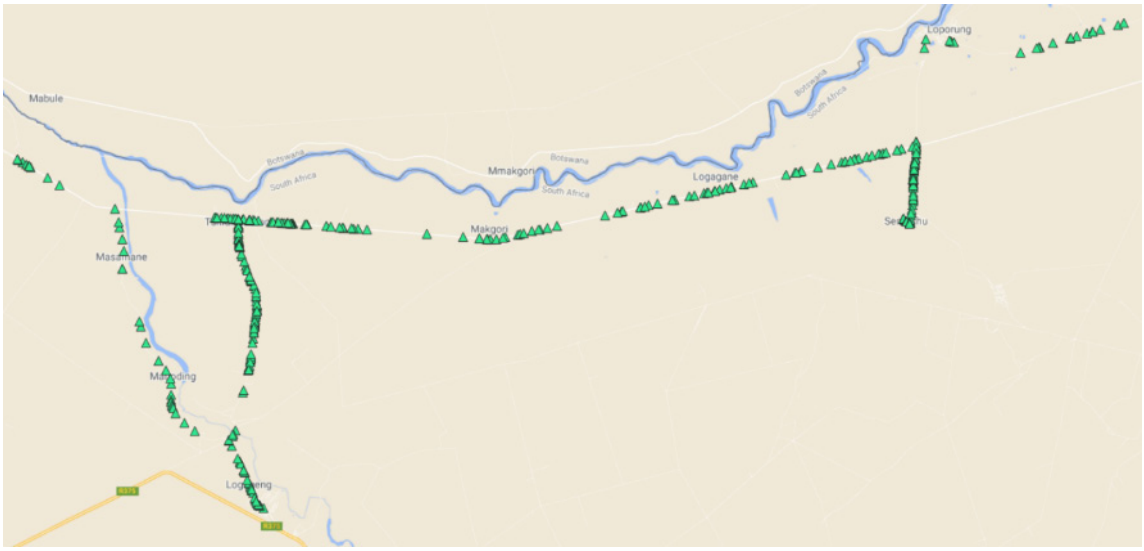

Figure 7: Topic 8 - South Africa/Botswana border

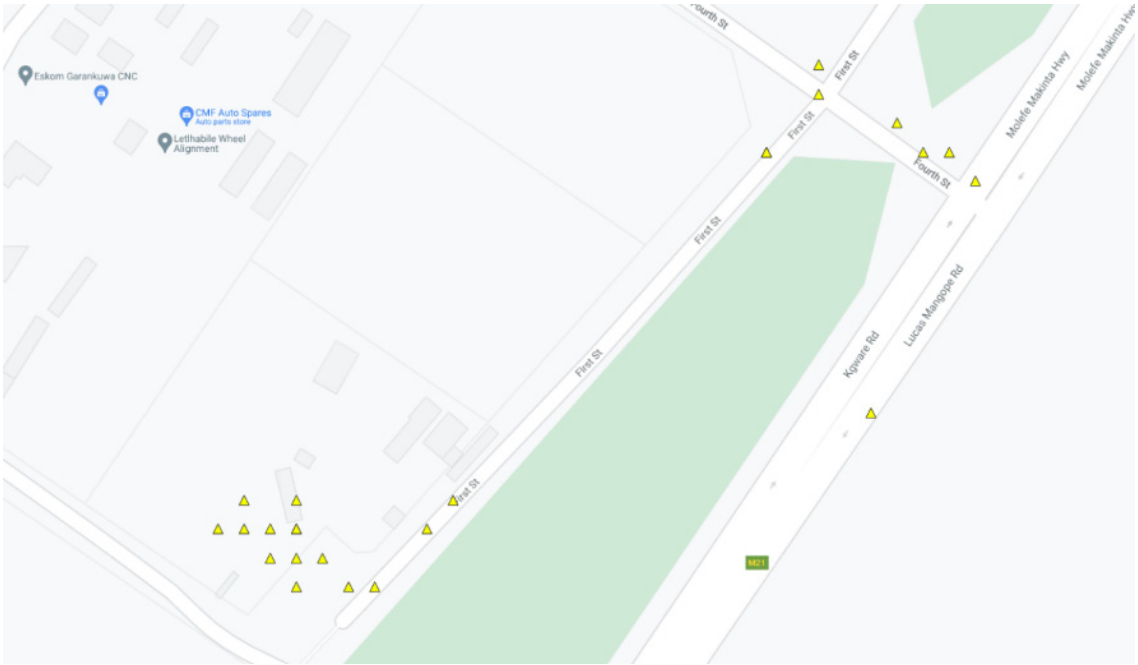

Figure 8: Topic 1 - Ga-Rankuwa (industrial area)

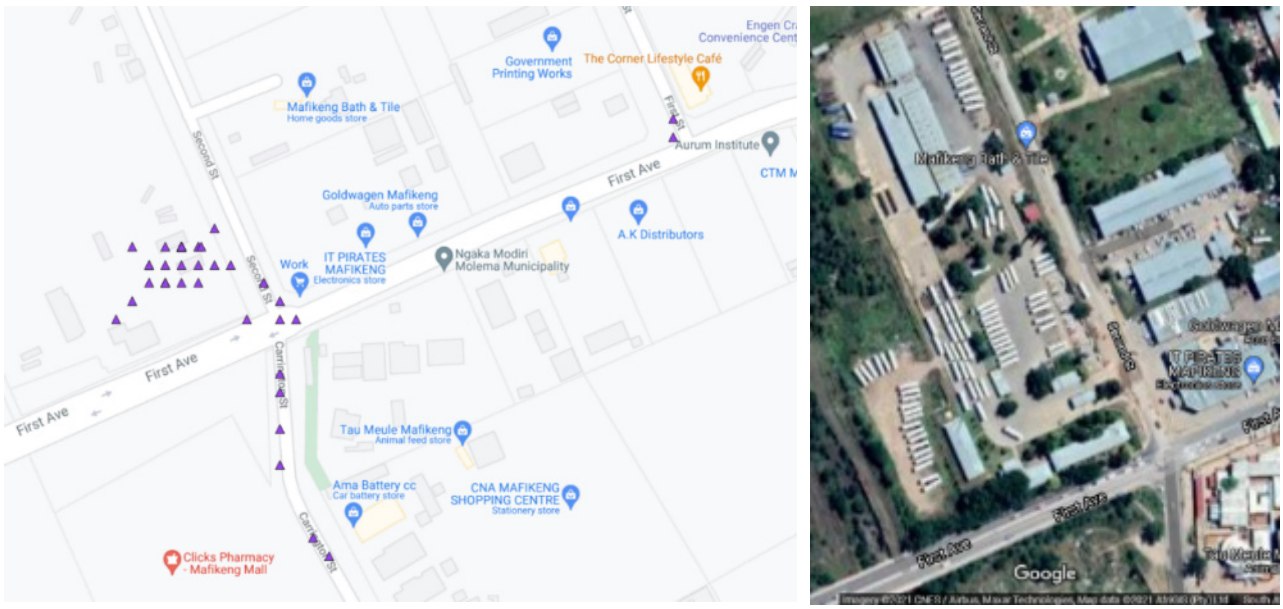

Figure 9: Mahikeng (industrial area)

Figure 9 shows Mahikeng's industrial area, where the corner of First Avenue and Second Street has a dense cluster of points. Google Maps Streetview indicates a bus depot, and Figure 9 also shows the satellite view with the buses visible.

Figure 10 shows the topic terms (topics 3, 10, 15, 19) present in the centre of Rustenburg. The clustering of terms indicates three main clusters. The blue cluster is situated between a police station, hospital, hardware store, and motor dealership (Figure 11). The green cluster is at the bus depot and taxi rank, providing transport for people staying in the surrounding rural areas (Figure 12). The red cluster is along Beneden street, which goes through Rustenburg, and leads to the airfield and numerous settlements.

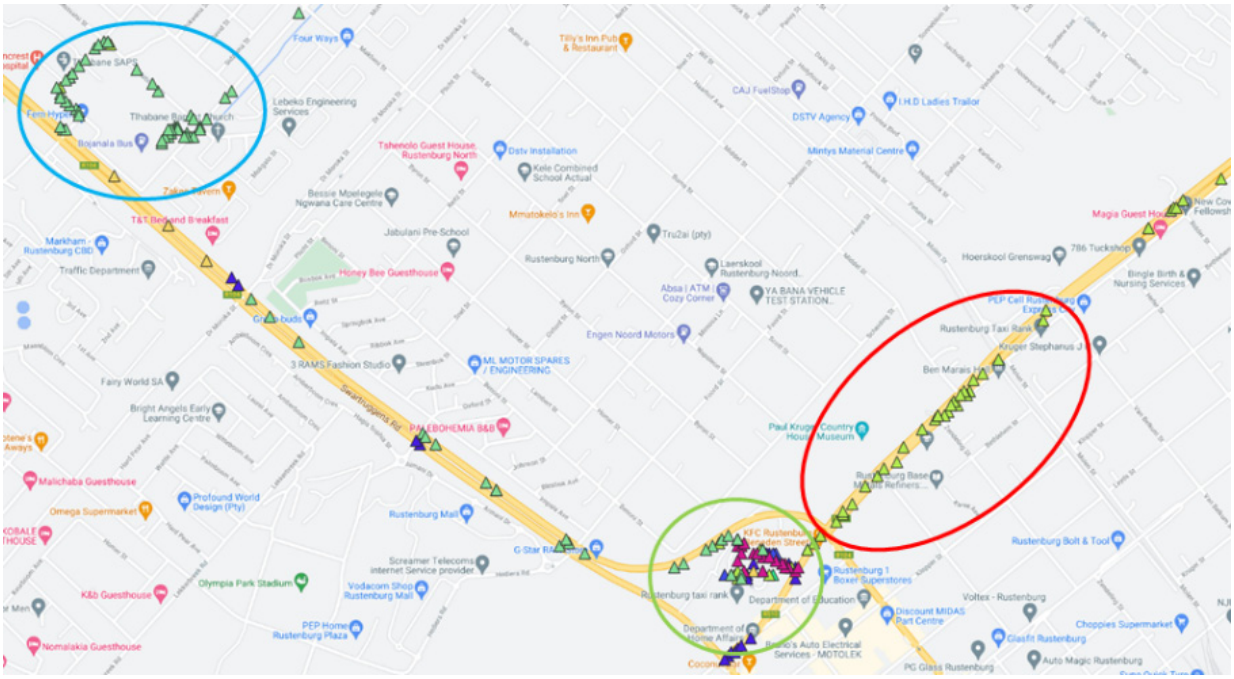

Figure 10: Rustenburg

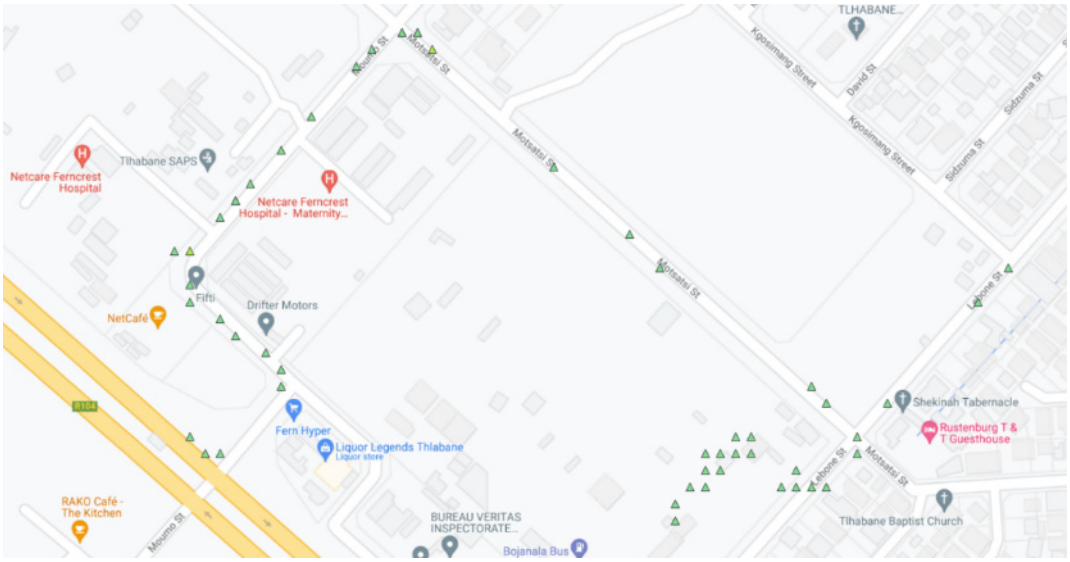

Figure 11: Rustenburg

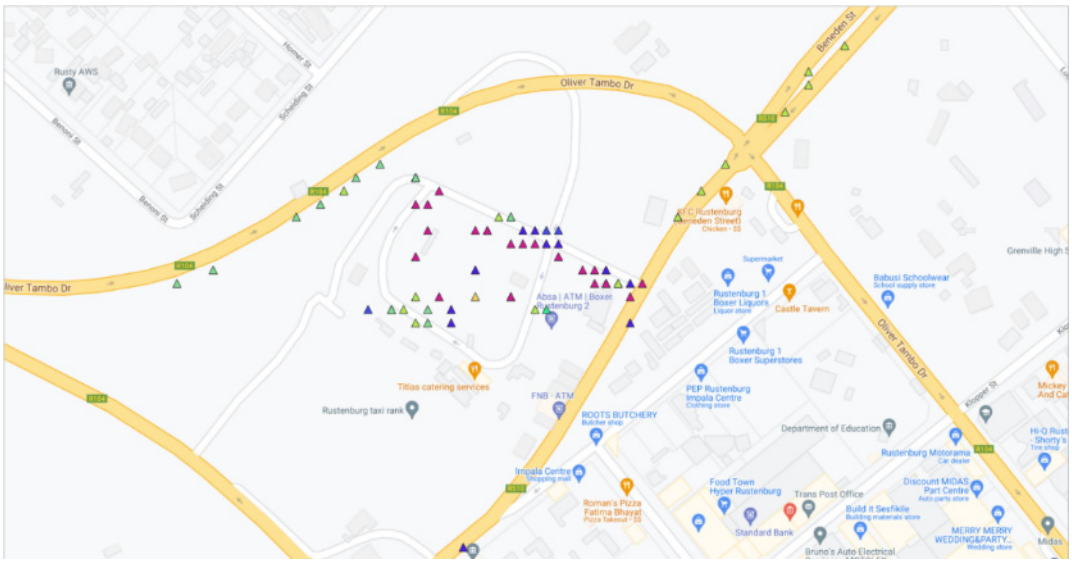

Figure 12: Rustenburg

5. CONCLUSION

Topic modelling and, more specifically, LDA is used to find interesting spatial patterns in bus route data collected over 17 months in the Gauteng and Northwest provinces of South Africa. Topic modelling algorithms are traditionally applied to text documents, whereas in this case, the data consists of GPS coordinates. A dictionary is used to convert GPS coordinates to words for the LDA algorithm to process. The LDA yields topics and topic terms, which are converted back to GPS coordinates to plot on a map.

It is difficult to derive meaning from points on a map that represent topics. The terms cannot be connected to meaningful words in a document to be interpreted as part of NLP, as with a traditional topic modelling application. It is not easy deciding why a topic term is deemed important by the LDA algorithm. However, using street views and satellite views in Google Maps assist in interpreting the clusters in the topic terms (GPS coordinates). The proximity of busy intersections, towns, settlements, and points of interest is used as anchor points to establish the importance and meaning of topic terms. The advantage of

topic modelling outputs over direct coordinate plotting, or a heatmap, is that a heatmap displays a more general area of interest. In contrast, topic modelling shows specific points of interest along the bus route. This could be extremely beneficial in city planning or for the location of new businesses where a specific location (for example, next to a motorway or street) is more important than a general area. As part of future work, this may be integrated with other data sources, including footfall data, for a more accurate prediction.

6. ACKNOWLEDGEMENTS

The authors would like to thank the CSIR Smart Mobility group for their assistance in collecting the data, and for their review of this paper.

7. REFERENCES

Agrawal, A, Fu, W & Menzies, T, 2018. What is wrong with topic modeling? And how to fix it using search-based software engineering. Information and Software Technology, 98, 74-88. Available at: https://doi.org/10.1016/j.infsof.2018.02.005.

Chen, H, Wang, X, Pan, S & Xiong, F, 2019. Identify Topic Relations in Scientific Literature Using Topic Modeling. IEEE Transactions on Engineering Management, 0018, 1-13.

Chen, X & Wang, H, 2019. Automated chat transcript analysis using topic modeling for library reference services. *Proceedings of the Association for Information Science and Technology*; 56(1):368-371.

Chu, D, Sheets, D, Zhao, Y, Wu, Y, Yang, J, Zheng, M & Chen, G, 2014. Visualising Hidden Themes of Trajectories with Semantic Transformation. *IEEE Pacific Visualization Symposium*, pp. 137-144.

Cmore, 2021. A Platform for Shared Awareness. Available at: https://www.csir.co.za/platform-shared-awareness. Accessed 10 October 2021.

Eker, S, Rovenskaya, E, Langan, S & Obersteiner, M, 2019. Model validation: A bibliometric analysis of the literature. Environmental Modelling and Software, 117, 43-54. https://doi.org/10.1016/j.envsoft.2019.03.009.

Finch, WH, Hernández Finch, ME, McIntosh, CE & Braun, C, 2018. The use of topic modeling with latent Dirichlet analysis with open-ended survey items. *Translational Issues in Psychological Science*, 4(4):403-424. Available at: https://doi.org/10.1037/tps0000173

Gholampour, I, Mirzahossein, H & Chiu, YC, 2020. Traffic Pattern Detection Using Topic Modeling for Speed Cameras Based on Big Data Abstraction. *Transportation Letters,* pp. 1-8.

Hecking, T & Leydesdorff, L. Can topic models be used in research evaluations? Reproducibility, validity, and reliability when compared with semantic maps. Research Evaluation; 2019; 28(3):263-272.

Jia, Y, Wang, W, Liang, J, Liu, L, Chen, Z, Zhang, J, Chen, T & Lei, J, 2018. Trends and characteristics of global medical informatics conferences from 2007 to 2017: A bibliometric comparison of conference publications from Chinese, American, European and the Global Conferences. Computer Methods and Programs in Biomedicine, 166, 19-32. Available at: https://doi.org/10.1016/j.cmpb.2018.08.017

Kim, J & Kang, P, 2018. Analysing international collaboration and identifying core topics for the "internet of things" based on network analysis and topic modeling. International Journal of Industrial Engineering: Theory Applications and Practice, 25(3):349-369.

K'Sharma, A, 2020. *Understanding Latent Dirichlet Allocation (LDA).* Available at: https://www.mygreatlearning.com/blog/understanding-latent-dirichlet-allocation/. Accessed 23 August 2021.

Lossio-Ventura, JA, Gonzales, S, Morzan, J, Alatrista-Salas, H, Hernandez-Boussard, T & Bian, J, 2021. Evaluation of clustering and topic modeling methods over health-related tweets and emails. *Artificial Intelligence in Medicine*; vol. 117, 102096.

Maier, D, Waldherr, A, Miltner, P, Wiedemann, G, Niekler, A, Keinert, A, Pfetsch, B, et al. Applying LDA topic modeling in communication research: Toward a valid and reliable methodology." Communication Methods and Measures; 2018; 12(2-3):93-118.

Manthiramoorthi, M, 2021. *Topic Modelling Techniques in NLP.* Available at: https://iq.opengenus.org/topic-modelling-techniques/. Accessed 23 August 2021.

Nolasco, D & Oliveira, J, 2019. Subevents detection through topic modeling in social media posts. *Future Generation Computer Systems*; 98:290-303.

Price, E, Wordlist10000. Available at: https://www.mit.edu/~ecprice/wordlist.10000 Accessed 20 December 2021.

QGIS, 2021. A Free and Open Source Geographic Information System. Available at: https://qgis.org/en/site/. Accessed 15 October 2021.

Sethasathien, N & Prasertsom, P, 2020. Research Topic Modeling: A Use Case for Data Analytics on Research Project Data. 2020 1st International Conference on Big Data Analytics and Practices, IBDAP 2020. Available at: https://doi.org/10.1109/IBDAP50342.2020.9245451

Suominen, A & Toivanen, H. Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. Journal of the Association for Information Science and Technology; 2016;67(10):2464-2476.

Tang, T, 2021. *Discover Hidden Trip Themes from GPS Data with Topic Modeling.* Available at: https://towardsdatascience.com/discover-hidden-trip-themes-from-gps-data-with-topic-modeling-f70cf04294c4. Accessed 10 July 2021.

Tang, Y, Sheng, F, Zhang, H, Shi, C, Qin, X & Fan, J, 2018. Visual Analysis of Traffic Data Based on Topic Modeling (ChinaVis 2017). *Journal of Visualization,* 21:661-680.

Weng, J, 2019. *Exploratory Data Analysis: A Practical Guide and Template for Structured Data.* Available at: https://towardsdatascience.com/exploratory-data-analysis-eda-a-practical-guide-and-template-for-structured-data-abfbf3ee3bd9. Accessed 9 Sep 2021.

Zhou, X, et al., 2021. When Intelligent Transportation Systems Sensing Meets Edge Computing: Vision and Challenges. Applied Sciences, 11(20):9680. Available at: http://dx.doi.org/10.3390/app11209680.