Manuscript Submitted	12 Nov 2019
Accepted	28 Nov 2019
Published	31 Dec 2019

Entity Clustering of User Reviews Using Topic Modelling and Similarity Scores

Wong Vee Kin, Bong Sze Ting, Lee Chun Mun & Vimala Balakrishnan
Faculty of Computer Science and Information Technology
University of Malaya, Malaysia
wqd180064@siswa.um.edu.my; wqd180066@siswa.um.edu.my; wqd180033@siswa.um.edu.my; vimala.balakrishnan@um.edu.my

Lok Pik Yin
The Nielsen Company, Malaysia
pikyinlok@gmail.com

Ng Jiunn Jye
Persistent System, Malaysia
ngjj@persistent.com

Abstract

The popularity of online review sites such as Amazon and TripAdvisor have created a window into rich insights of what matters to reviewers, and potentially customers. For the business sector, this provides opportunities to improve their businesses and to attract more customers. The unstructured nature of user reviews, however, makes it difficult to analyze, and natural language processing techniques are instrumental in extracting useful information from these resources. To digest the free form review texts, it is desirable to abstract the reviews into high level themes. However, research works in this area is still scarce. In this paper, the process taken to perform this high-level abstraction is detailed, with an example derived from the hotel industry (identity concealed for confidentiality purpose). More than 3000 sets of customer reviews were processed using the Extract, Transform and Load process, supported by topic modelling algorithms including Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA) and K-means. The study aims to identify significant business aspects or features based on the customer reviews. Preliminary results from these algorithms lack usability, which can be mainly attributed to unknown number of hidden topics and the brief nature of the review texts. To address this issue, a cosine similarity-based algorithm was introduced to improve the thematic clustering into one that is more intuitive and actionable. Ten themes emerged, and using cosine similarity, the review texts in the corpus were intuitively clustered for business planning insights. With this method, the review texts can be classified in a matter of a few minutes and demonstrates the strength of natural language processing for insights mining.

***Keywords*: Clustering, Latent Dirichlet Allocation, Latent Semantic Analysis, similarity, topic modeling, agglomerative clustering, business intelligence**

1. Introduction

The ubiquity of the world wide web and popularity of review sharing via social media and industry specific websites have created a vast reservoir of knowledge waiting to be tapped. In the last few decades, natural language processing (NLP) algorithms have advanced by leaps and bounds, making text mining for descriptive and prescriptive analytics feasible. Even so, every text mining endeavor starts with a different goal, and hence the algorithms and approaches taken will need to be tailored accordingly. The goal in this paper is to extract important high-level themes from unbounded user reviews for business planning and improvement. This will summarize a large collection of reviews and abstract it into a few high-level categories to draft an actionable business plan. One of the issues tackled in this work is the reality that clustering decisions can be subjective in nature. The same collection of review items can be clustered in different ways, depending on business needs. In this paper, we first experiment with well-known clustering and topic extraction algorithms and later propose a method which can be tailored to custom needs.

2. Previous Studies

Literature search shows that there have been many studies in user reviews and clustering. For example, Akhtar et al. (2017) manually analyzed more than 1000 reviews in TripAdvisor in order to find recurring words which will assist in aspects identification. Manual inspections by human seem to be inevitable in text classification until Weismayer et al. (2018) ran an experiment involving both RapidMiner Studio and human, showing machine-based models are able to yield a similar accuracy as human in identifying aspects within TripAdvisor reviews. Despite the high accuracy of supervised learning in text classification (Chang et al., 2017), researchers have pointed out that it is time consuming given the large amount of manual labelling (Ekinci & Omurca, 2019; Wu et al., 2018). Hence, there was a diversion of focus to unsupervised learning method, which does not require any label as input.

Latent Dirichlet Allocation (LDA) is one of the most popular tools in unsupervised learning and had been used in many researches to perform topic modelling (Xu et al., 2017; Guo et al., 2017) or aspect identification, and sentiment analysis (Akhtar et al., 2017; Putri & Kusumaningrum, 2017). LDA has proven to have improved performance in discovering latent topics with the aid of word embedding such as word2vec (Devkota et al., 2019) or work in parallel with supervised machine learning algorithm such as Support Vector Machine (SVM) (Dickinger et al., 2017). There are also scholars who have leveraged Part of Speech Tagging (POS Tagging) in mining hidden aspects in the TripAdvisor reviews (Afzaal et al., 2019) as aspects are always in the form of nouns.

The applications of unsupervised learning in analyzing TripAdvisor reviews range from hotel reviews (Devkota et al., 2019) to tourism attractions reviews such as museum (Alexander et al., 2018). In their studies, challenges related to variable-gram entity extraction and theme identification were apparent.

3. Methodology

One of the key issues with unsupervised clustering of text is that the number of hidden topics or clusters is not always known. Furthermore, it can be difficult to identify the cluster themes. In this paper, the process of evaluating various theme discovery and clustering algorithms are described. To test the clustering result of each algorithm, we use a rudimentary test to verify if a subset of entities generally known to belong to a desired theme is clustered correctly. In this work, the clustering results are rejected if food related items (identified via sampling the reviews, i.e. {'dinner', 'rice', 'lunch', 'breakfast', 'buffet', 'pasta', 'ribeye', 'restaurant'} does not belong to the same cluster. While lacking the rigor of formal coherence scoring metrics such as the co-occurrence and pointwise mutual information measures (Mazarura & De Waal, 2016), this logical test offers alignment to intuition and can be expanded as needed.

Entities of interest were first extracted from user reviews before clustering them into groups using k-means algorithm. Topic modeling techniques were applied to cluster entities with similar themes.

Additionally, agglomerative clustering based on cosine similarity distance was administered to search for coherent clusters. The following sub-sections elaborate the steps and processes involved.

3.1 Dataset

The dataset used in this paper consists of a total of 3055 reviews gathered from a hotel review website (i.e. www.tripadvisor.com)., with user name, user location, rating, review title and review text as features. Although the focus of this paper is the review text, the extracted features were used for association analysis. Figures 1 and 2 show the distribution of the sentences and words in each user review. Although there are outliers, the figures show that most of the review texts are brief, mostly about 4-5 sentences and the sentences can be very short (~5 words) or long (~20 words). These by itself present significant challenges to this task, as models that work on long texts do not work as well for short texts as mentioned in (Li et al., 2016; Blei & Lafferty, 2009).

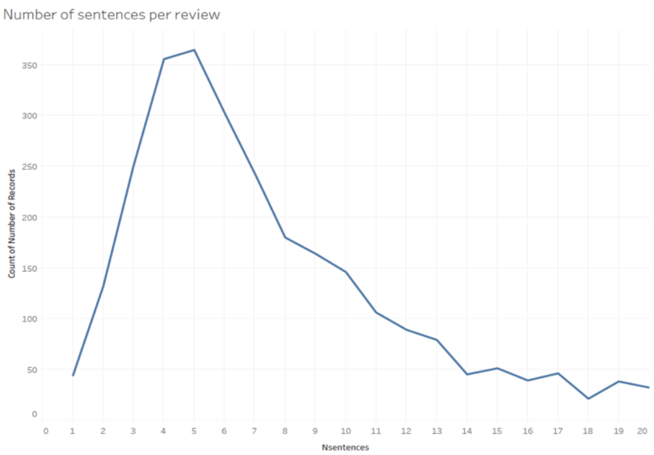

Figure 1: Distribution of sentences for each review

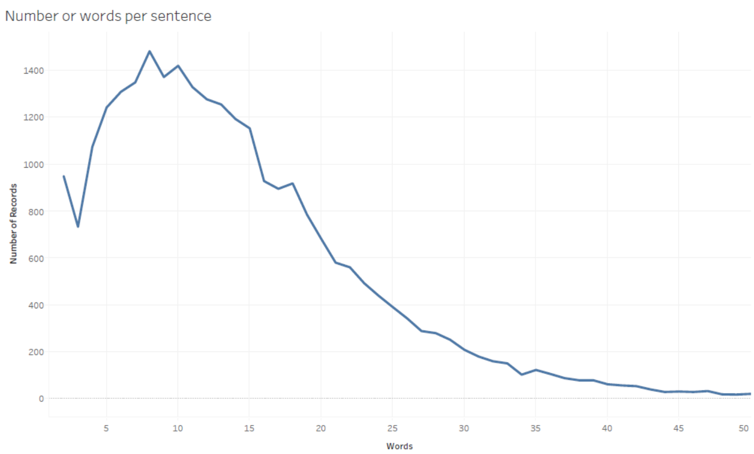

Figure 2: Distribution of words in each sentence in the reviews

In order to extract meaningful insights from these reviews, the entities were extracted first. Since the entities of interest are typically nouns, parts-of-speech (POS) tagging was used to identify these nouns. In this present study, we are interested in nouns that are in the vicinity of an adjective such as “huge” (followed by the noun “room” which is an entity of interest), since such entities are those that the reviewers feel strongly about. Also, since an adjective can precede or follow a noun, we included nouns on both sides of an adjective. For example, “the service was good” and “we enjoyed the good service” have adjectives preceding and following the noun. Entities with phrases of more than one word were merged to form entity phrases. Figure 3 below shows the entity extraction flow that feeds into the clustering algorithms. First, basic cleaning was done to the extracted reviews. These include removing punctuations and non-ASCII characters. POS tagging was then done using the Natural Language ToolKit (NLTK), followed by nouns that are within 5 words in a sentence were appended to a list of entities. If the noun is adjoining another noun, a noun phrase entity will be appended instead. Then, the reviews were transformed into a bag-of-words count vector, where entities (both nouns and noun phrases) were treated as individual tokens. These vectors are used as inputs to the entity clustering algorithms mentioned below.

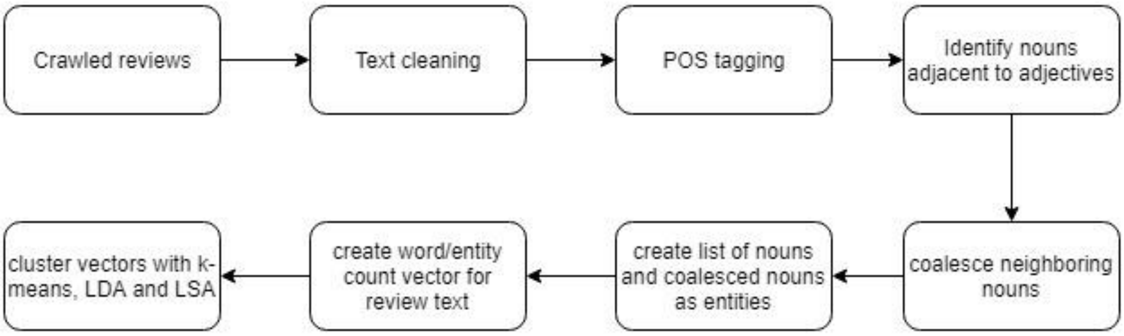

Figure 3: Parts-of-speech tagging and initial entity clustering approach using k-means

3.2 Clustering and topic modelling

This section briefly describes the k-means, LDA and LSA algorithms used in the initial phase of this work. The proposed method is discussed in the later section.

K-means

K-means is one of the best known and versatile clustering algorithms. The algorithm begins by a random selection of k centroids {C1, C2,…Ck} and distances from each entity in the dataset to these centroids are computed. The entities are clustered to the closest centroid and the cluster centroid is updated. The entire process repeats until the cost function meets the limit criteria.

LDA

LDA attempts to distribute the entities into a fixed set of topics based on a generative model approach. Each topic z is associated with a multinomial distribution over the vocabulary Φz, and a given document Di is assumed to have a topic distribution Θi. The LDA algorithm searches for Φz and Θi that best fits the corpus.

LSA

LSA learns topics by first forming a term by document matrix and then smoothing the counts to enhance the weight of informative words. The matrix is then decomposed, in this work by singular value decomposition into the input number of categories.

4. Clustering Results

4.1 K-means clustering

In order to get a general overview of the different reviews, k-means clustering was used to analyze reviews that were grouped together. The cluster sum squared error (SSE) shows that no optimum clustering is possible up to 100 clusters. If the number of clusters is too large, the insights would not be very valuable since it does not describe macro-level, interpretable behavior. Hence, we picked a cluster size of 10 to explore if any meaningful trends emerge. While the cluster sizes were disproportionate, the number of unique entities within each cluster was quite significant, even for the smallest clusters, as indicated in Figure 4.

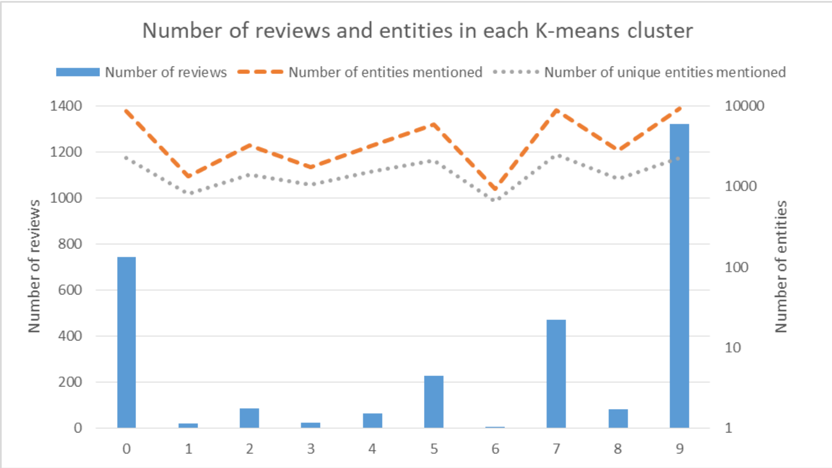

Figure 4: K-means clustering of the reviews

In order to determine if a theme exists within the clusters, unique entities within each cluster were analyzed. As shown in Table 1 no clear distinction of themes was found within each cluster. For example, food entities appear within many clusters, as highlighted in bold.

Table 1: Entities within each cluster from k-means

Cluster	Entities
0	'sun_beds', 'extension', **'dining_experiences'**, 'bit_cranky', **'prawns_dish'**, 'tasty_options', 'toilet_facilities', 'exhausting', 'widespread', 'safes'
1	‘braces', 'applause', 'frame', 'packets', 'angle_bracket', 'morale', 'injury', 'mid', 'hips', 'hotel_perfect'
2	'room_door', 'Need', 'aircond_walkway', **'cuisine_level'**, **'food_evening'**, 'pillow_list', 'hotel_quests', 'hotel_set-up', 'air_stream', 'nightFood'
3	'cooling', 'strainer', 'concept_layout', 'dirt', **'coffee_pods'**, 'copycat_brand', 'distinctive', 'interiors/furnishings', 'packaging', 'Feedback'
4	**'buffet_spread_variety'**, 'bath_switch', 'telekung', 'vanity_corridor', 'Hence', 'peeves', 'glass_separator', 'fares', 'basics', 'power_point'
5	**'food_restaurant'**, 'term_stay', **'beautiful_restaurant'**, 'fitness_trainer', 'gym_etc', 'visit/vacation', 'layout/surroundings', 'species', 'security_staff', 'accrediton'
6	'registration', 'view_infront', 'dropping', 'lobby_afterall', 'hand_prints', 'amuse-bouche', 'sirloin', 'sarcasm', 'blind', 'rag'
7	'office_area', 'club_offer', 'garden_views', 'bathroom_space', 'hotel_hardware', 'tad_pricey', 'night_clubs', 'room_cleaning_sign','hotels_amenities', 'front_desk_service'
8	'paidment', **'superb_lunch_buffet'**, **'buffet_lunch'**, 'trademark', 'front-of-house_team', 'kind_consideration', 'in-room_team', **'gourmet_experience'**, **'kuay_teow'**, 'classics'
9	'checkin_experience', 'client', **'bfast_spread'**, **'breackfast'**, 'condominium_view', 'pavilion_shopping_mall', 'shop', 'shove', 'continental', 'twins_towers'

To address this issue, topic modeling approaches were explored, as they are well known to extract hidden topic within text (Stevens et al., 2016). The topics can be interpreted as themes of the entities.

4.2 Latent Dirichlet Allocation (LDA)

One of the requirements of LDA is to input the number of topics (Blei et al., 2003). We tested the algorithm with 10 topics in the present study. The results are shown in Figure 5 below.

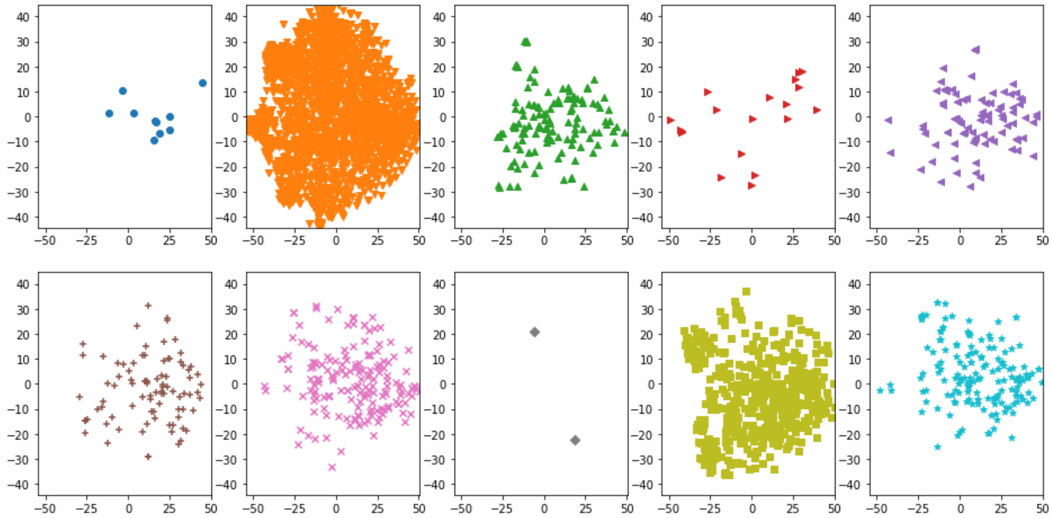
Figure 5: Grouping of entities from LDA (clusters are numbered 0-9 from top left to bottom right)

A look at the entities list show that the clustering is not all intuitive. For example, we looked at known food-themed entities 'dinner', 'rice', 'lunch', 'breakfast', 'buffet', 'pasta', 'ribeye' and 'restaurant' to test the coherency of the clustering. The result shows that these were split among three clusters, as follows:

Cluster1: {dinner, rice, lunch, breakfast, buffet, restaurant}
Cluster6: {ribeye}
Cluster8: {pasta, vegetables}

4.3 Latent Semantic Analysis (LSA)

For LSA, Singular Value Decomposition (SVD) was used to decompose the probability of each entity falling into individual topics. The results are depicted in Figure 6, again the entities list did not show a clear distinction between the clusters. The same basic coherency test on food entities was done, with the following results.

Cluster0: {lunch, buffet}
Cluster6: {breakfast}
Cluster7: {ribeye}
Cluster8: {rice}
Cluster9: {dinner, pasta, vegetables}

As indicated in the results above, LSA appears to be less discriminative than LDA in clustering coherency in our experiment since these similarly themed entities were spread out to more clusters. From the results of LDA and LSA, it can be concluded that increasing the number of topics will not improve clustering accuracy and coherency since similarly themed entities will be split further. Therefore, to further address this issue, we rely on cosine similarity-based clustering to improve coherency, as elaborated in the following sub-section.

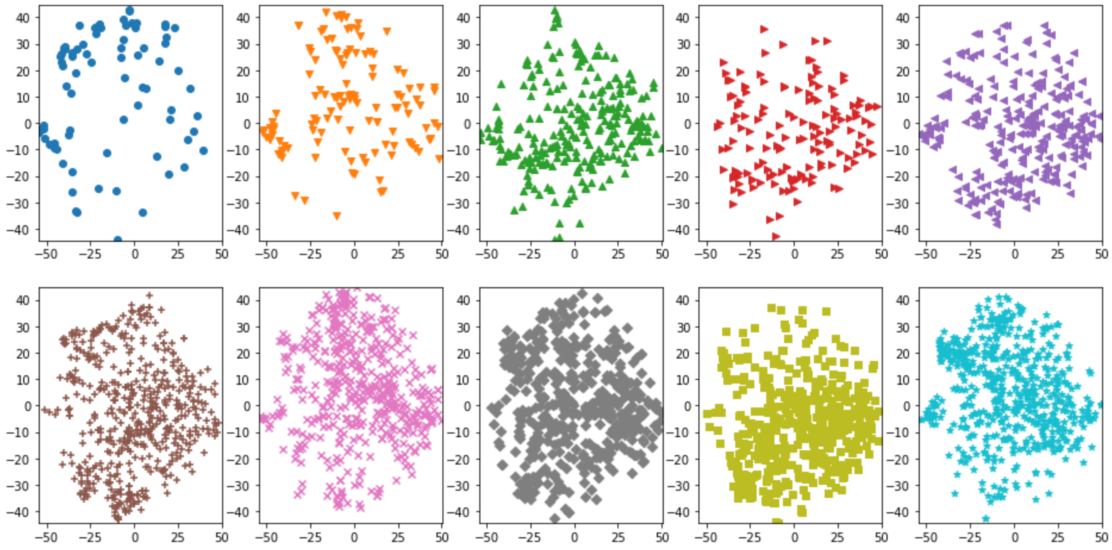

Figure 6: Grouping of entities from LSA (clusters are numbered 0-9 from top left to bottom right)

5. Proposed Method

5.1 Agglomerative Clustering based on Cosine Similarity

Both LDA and LSA represent a top-down approach to cluster the entities, resulting in two key issues in the study:

i. the clusters do not represent an expected intuitive clustering, and

ii. the semantic coherency within each cluster is low, even if only based on a simple test. In other words, the noise level is high

An alternative approach is a bottom-up approach. In this case, an agglomerative clustering approach was used based on seed identification. In order to improve the accuracy of similarity scores, a word vector trained on the review texts was generated. The clustering algorithm used is as follows:

Algorithm (1):

$Set\ Sim_{th} = 0.9,\ Sim_{th}\ is\ similarity\ threshold$

$Pairs = \{(x_i, x_j) \mid \mathrm{Sim}(x_i, x_j) > Sim_{th}\}, x_i \in X, x_j \in X;\ i = 1\ to\ N\ where\ X = all\ entities$

$While\ Count(Pairs) > 0:$

$$Seeds = \{Seed_m \cup P_\alpha \mid P_\alpha \in Pairs\ ;\ Seed_m \cap P_\alpha \neq \emptyset\}$$

First, we extract only the word vectors for entities of interest. Then, we generate a matrix of cosine similarity distances for all the entities and identify those that meet a similarity threshold of 0.9. The entities that exist in these pairs are ranked according to the count of occurrence. In other words, we group entities that are pivoted on the same entity and use it as a seed.

The results of the seed formation are shown in Fig. 7. Table 2 on the other hand, shows the entities that fall within each seed.

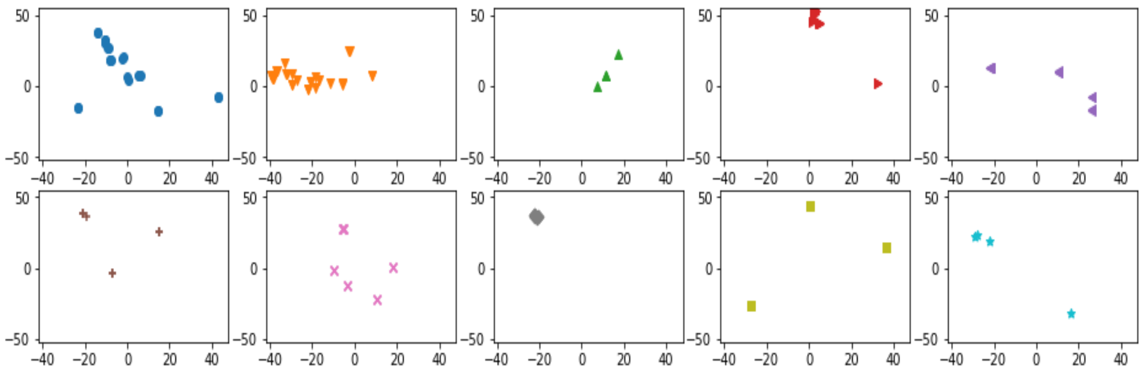

Figure 7: Coherent clusters after first round agglomeration with 0.9 similarity score and grouping.

Table 2: Entities in each seed from cosine similarity-based clustering

Seed	Entities
cluster 0	'converter', 'dock', 'cable', 'safes', 'closest', 'lifting', 'loans', 'remover', 'deodorant', 'makeup', 'make-up', 'lifts', 'elevators', 'escalators', 'wi-fi', 'mattress', 'firm', 'shoe', 'polish', 'rugs', 'wifi_connection'
cluster 1	'carte', 'ala', 'chicken', 'rice', 'sauce', 'dahl', 'noodle', 'tarik', 'chilli', 'pastries', 'vegetables', 'turkey', 'congee', 'eats', 'classics', 'kuay_teow', 'cuisine', 'western', 'digestives', 'liqueurs', 'food-', 'nasi_lemak', 'nasi', 'lemak', 'porridge', 'pineapple', 'tart', 'ribeye', 'sirloin'
cluster 2	'hips', 'injury', 'discharge', 'symptoms'
cluster 3	'walkways', 'passage', 'bukit', 'underground', 'link', 'mall', 'pavillion', 'walkway', 'pavilion', 'suria', 'tunnel', 'bintang', 'bridge', 'connect', 'roads', 'tunnels', 'bridges', 'star', 'stars'
cluster 4	'evening_canape', 'calmer', 'healing', 'show_kitchens', 'rejuvenate', 'ramuan', 'revenues', 'oil', 'sultanate', 'quietness', 'room_quality'
cluster 5	'wall', 'glass', 'privacy', 'humus', 'suspects', 'report', 'police'
cluster 6	'consierge', 'gound_floor', 'hotel_mgmt', 'downturn', 'stars_service', 'logistic', 'packaging', 'toothbrushes', 'toothpaste', 'separation', 'room_cleaning_sign', '*****']
cluster 7	'bath', 'tub', 'vanity', 'closet', 'wardrobe', 'toilet', 'sink'
cluster 8	'bus_stop', 'hop-', 'traffic_jam', 'peroid'
cluster 9	'assortment', 'free-weights', 'strength', 'cardio', 'sauna', 'locker', 'jacuzzi', 'steam'

It is easy to see the themes that emerge from these small collections of entities. Cluster0 relates to facilities or amenities, cluster1 is food, cluster2 is health related, cluster3 for landmark, cluster4 for relaxation theme, cluster5 for security, cluster6 for service, cluster7 for bath and changing, cluster8 for transport and cluster9 for workout.

From these seeds, new entities that meet a new similarity threshold were added based on algorithm (2). While the number of entities increase for each cluster, as the similarity threshold is relaxed there are more irrelevant entities included in each cluster. Trimming of the list is necessary to reduce the number of irrelevant entities in a cluster which if included will cause more irrelevant words (noise) to be included. The result for this step is shown in Fig. 8.

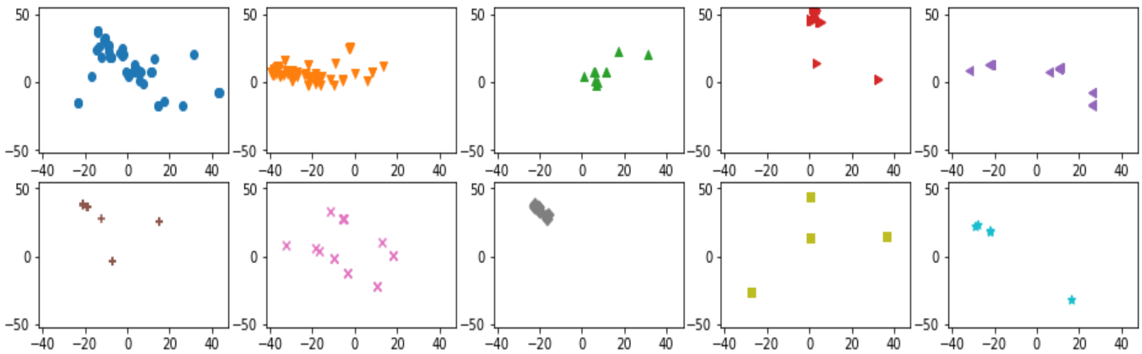

Figure 8: Cluster expansion after one iteration of single link clustering

Algorithm (2):

$Set\ Sim_{th} = 0.85,$
$\mathrm{New}Pairs = \{(x_i, x_j) \mid \mathrm{Sim}(x_i, x_j) > S_{th}\}$
$While\ Count(NewPairs) > 0:$
$\quad Clusters = \{Seed_m \cup P_\beta \mid P_\beta \in NewPairs\ ;\ Seed_m \cap P_\beta \neq \emptyset\}$

Finally, we expanded the clusters using a minimum single link clustering by including other entities that have better than 0.6 cosine similarity score to any entity within the cluster. Each of these new entities is assigned to the existing cluster where any of its entity is nearest to the new candidate. The final clustering results are depicted in Fig. 9.

Algorithm (3):

$Set\ Sim_{th} = 0.6$
$NewPairs = \{(x_i, x_j) \mid \mathrm{Sim}(x_i, x_j) > Sim_{th}\}$
$for\ Pair\ in\ NewPairs:$
$\quad for\ c\ in\ Pair:$
$\qquad k = \{\max Sim(c, x_{\alpha,i})\}\ where\ \alpha = 1\ to\ N_{clusters}\ and\ i = 1\ to\ N_{entities\ in\ cluster}$
$\qquad Cluster_m \leftarrow c;\ \ m = argmax(k)$

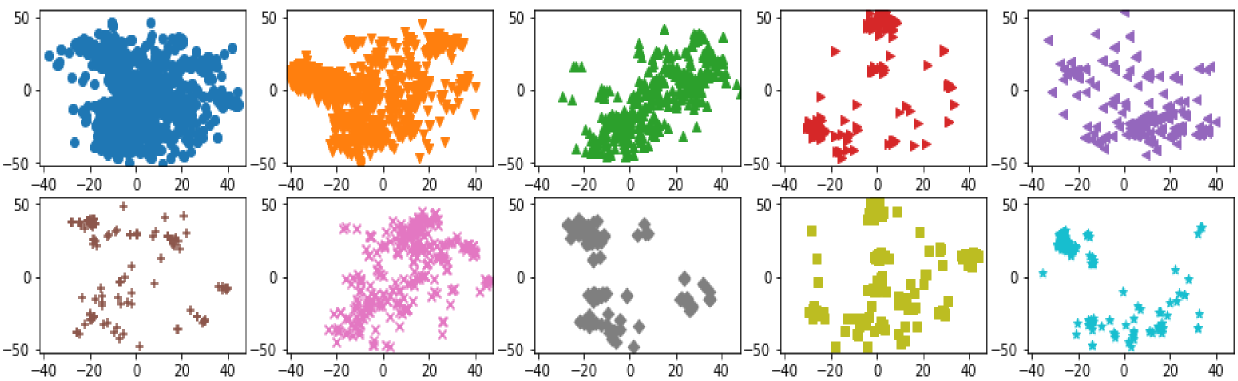

Figure 9: Cluster entities after nearest entity linkage based on similarity score threshold of 0.6

The quick coherency test shows that all food related entities are in the same cluster.

Cluster1= 'dinner', 'rice', 'lunch', 'breakfast', 'buffet', 'pasta', 'restaurant', 'vegetables', 'cuisine', 'ribeye'

With this algorithm, we intuitively select and assign themes to small clusters of entities in the seeds, and systematically include all other entities into the defined seeds, and the coherency within the clusters are governed by the set cosine similarity threshold.

6. Discussion and Conclusion

Theme discovery in unstructured texts is challenging. In this paper, top down approach in theme discovery using k-means clustering and topic modeling approaches (i.e. LDA and LSA) were experimented to identify prominent themes among hotel reviews. However, initial clustering results show that using these approaches, the clustering themes are not apparent and intuitive. Furthermore, in a given collection of reviews, there is no fixed rule to determine the number of clusters or topics, which is an important parameter for these algorithms. In addition, noise exists in the clusters which can be attributed to polysemous words, and syntactic inconsistencies in the reviews due to diverse grammatical flavor. Spelling errors were also rampant in the texts. Alternatively, a bottom-up approach was proposed i.e. agglomerative clustering based on cosine distance and it has the advantage of a higher coherency within each seed, and theme discovery is more intuitive.

Although the agglomerative clustering approaches used in this paper were semi-supervised, it is also possible to implement it in unsupervised mode. Having said that, there are merits for the semi-supervised approach in that it helps craft desired themes. Many words are related to more than 1 concept or theme. For example, a ‘restaurant’ can be related to a ‘food’ theme or a ‘landmark’ theme. Likewise, the word ‘wardrobe’ can also be assigned to a ‘furnishing’ theme or ‘vanity’ theme. Supervision helps resolve potential thematic decision choices when it arise, and techniques can be employed to make this process efficient, for example employing a heuristic rule-based filtering algorithm.

Both top-down and bottom-up approaches can be improved by improving the data pre-processing phase. A more refined text clean-up is needed to correct spelling errors. Moreover, a more accurate entity recognition and extraction can further reduce the number of false entities extracted. While these may reduce the noise level in the top down approach, theme discovery can be done more effectively using the bottom up agglomerative clustering approach.

Acknowledgment

The authors wish to extend their gratitude to Fundamental Research Grant Scheme [FP109-2018] for supporting this study.

References

Afzaal, M., Usman, M., Fong, A. C., & Fong, S. (2019). Multiaspect- based opinion classification model for tourist reviews. Expert Systems, 36(2), e12371.

Akhtar, N., Zubair, N., Kumar, A., & Ahmad, T. (2017). Aspect based sentiment oriented summarization of hotel reviews. Procedia Computer Science, 115, 563-571.

Alexander, G. Blank, S. Hale (2018). TripAdvisor Reviews of London Museums: A New Approach to Understanding Visitors. Museum International 70(1-2):154-165

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.

Blei, D. M., & Lafferty, J. D. (2009). Topic models. In Text Mining (pp. 101-124). Chapman and Hall/CRC

Chang, Y. C., Ku, C. H., & Chen, C. H. (2017). Social media analytics: Extracting and visualizing Hilton hotel ratings and reviews from TripAdvisor. International Journal of Information Management. 48, 263-279

Devkota, B., Kim, K. S., Zhuang, C., & Miyazaki, H. (2019). Disaggregate Hotel Evaluation by Using Diverse Aspects from User Reviews. In 2019 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 1-8). IEEE.

Dickinger, A., Lalicic, L., & Mazanec, J. (2017). Exploring the generalizability of discriminant word items and latent topics in online tourist reviews. International Journal of Contemporary Hospitality Management, 29(2), 803-816.

Ekinci, E., & Omurca, I. S. (2019). Concept-LDA: Incorporating Babelfy into LDA for aspect extraction. Journal of Information Science, 1 - 13

Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. Tourism Management, 59, 467-483.

Li, C., Wang, H., Zhang, Z., Sun, A., & Ma, Z. (2016) Topic Modeling for Short Texts with Auxiliary Word Embeddings. Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, pp.165–174

Mazarura, J., & De Waal, A. (2016, November). A comparison of the performance of latent Dirichlet allocation and the Dirichlet multinomial mixture model on short text. In 2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech) (pp. 1-6). IEEE.

Putri, I. R., & Kusumaningrum, R. (2017, January). Latent Dirichlet allocation (LDA) for sentiment analysis toward tourism review in Indonesia. In Journal of Physics: Conference Series (Vol. 801, No. 1, p. 012073). IOP Publishing.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring topic coherence over many models and many topics. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 952-961). Association for Computational Linguistics.

Weismayer, C., Pezenka, I., & Gan, C. H. K. (2018). Aspect-Based Sentiment Detection: Comparing Human Versus Automated Classifications of TripAdvisor Reviews. In Information and Communication Technologies in Tourism 2018. 365-380. Springer, Cham.

Wu, C., Wu, F., Wu, S., Yuan, Z., & Huang, Y. (2018). A hybrid unsupervised method for aspect term and opinion target extraction. Knowledge-Based Systems, 148, 66-73.

Xue, W., Li, T., & Rishe, N. (2017). Aspect identification and ratings inference for hotel reviews. World Wide Web, 20(1), 23-37.