ԲԱՆԲԵՐ ԵՐԵՎԱՆԻ ՀԱՄԱԼՍԱՐԱՆԻ. ՏՆՏԵՍԱԳԻՏՈՒԹՅՈՒՆ
2022, № 2, 73-83 *Տնտեսամաթեմատիկական մոդելավորում*
https://doi.org/10.46991/BYSU:G/2022.13.2.073

HOUSING MARKET SEGMENTATION BASED ON APARTMENT DESCRIPTIONS

TIGRAN KARAMYAN

Market segmentation is the process of breaking down diverse markets into homogenous groupings that have comparable demands, interests, and/or behaviours. Housing market, like any other market, has its own challenges when it comes to market segmentation, since the segmentation process depends not only on the product (houses or apartments) but also on the market player (owners or real-estate agents).

This paper aims to show how Natural language processing (NLP) can be used to determine the segments of the housing market in Yerevan by using the unstructured data (apartment descriptions) scrapped from a real-estate website. The collected textual data represents not only the descriptions of the apartments but also gives an idea of who wrote that text. The applied NLP model shows how certain behavioral patterns of market players can be expressed through textual data and how those patterns can affect market segmentation. That means the segmentation of the market using unstructured data represents not only product-related, but also psychographic and geographic picture of the customers (here – sellers) and their apartments.

Key words: *market segmentation, unstructured data, natural language processing, topic modelling, latent Dirichlet allocation, clustering, pyLDAvis*

Introduction: Understanding of the needs and wants of the customers is one of the main goals of marketing managers. One of the ways to acquire individual customers' likings is market segmentation. It is the process of identifying segments of the market and the process of division a broad customer group into sub-groups of consumers. The individuals of the same sub-group should share characteristics such as common needs, common interests, similar lifestyles, similar demographic profiles and, of course, they should have similar market preferences. Customer segmentation enables researchers to customize the company's marketing activities more accurately and adopt a more systematic approach when planning ahead for the future [1].

In order to properly divide a group of customers into segments researchers take into consideration five major factors:

- Demographic Segmentation: division of the market into groups that are identifiable in terms of physical and factual data such as age, gender, occupation, family size, race, religion and nationality.
- Geographic Segmentation: division of the market into groups based on

[1] **Camilleri, M. A.** (2018). Market Segmentation, Targeting and Positioning. In *Travel Marketing, Tourism Economics and the Airline Product* (Chapter 4, pp. 69-83). Springer, 2018, Cham, Switzerland.

geographic variables such as locations, climate, terrain, natural resources and population density.

- Psychographic Segmentation: division of the market into groups according to personality traits, values, motives, interests and lifestyles.
- Behavioural Segmentation: division of the market into groups by individual purchase behaviours.
- Product-related Segmentation: division of the market into groups based on the product or service.

This information helps businesses to develop undifferentiated, differentiated or concentrated marketing strategy, that is to identify a segment of the market which can be easily targeted.

Market segmentation can be applied to almost every field of economy or business, and it is aimed to identify the most profitable market segments to focus on while avoiding mass marketing. Here we will focus on house market segmentation in Yerevan city, Armenia.

Literature review: Market segments need to be identifiable, which is dependent on the quality and quantity of the available data. The required data for segmentation is collected taking into consideration the research objective or problem and there is a wide range of algorithms available to segment unlabeled market data. Cohort analysis is a type of behavioral analytics in which users or customers are grouped based on their common characteristics within a defined time-span to better track and understand their actions. One of the main steps in conducting cohort analysis is to select a key indicator or a metric (retention rate, churn rate, product sales number, transactions, etc.) that will become the main tool of the research. Another algorithm for market segmentation is RFM (recency, frequency, monetary) analysis. The main idea under RFM analysis is that 80% of your business comes from 20% of your customers. In other words, RFM is used to determine quantitatively which customers are the best ones by examining how recently a customer has purchased (recency), how often they purchase (frequency), and how much the customer spends (monetary). Another famous group of algorithms in market segmentation is clustering ones. There are two types of algorithmic clustering methods: hierarchical (AGNES, DIANA) and non-hierarchical (K-means, DBSCAN, OPTICS, etc). The main difference between these two types is that non-hierarchical clustering requires the number of clusters as an input variable, while hierarchical does not require this[2]. Later we will discuss mainly clustering algorithms.

One of the earliest researches in this field was conducted by Goodman and Thibodeau. In one of their work they have shown that the metropolitan Dallas housing market is segmented by the quality of public education using hierarchical models[3]. Later they have come up with another idea which examines

[2] **Jurowski, C. and Reich, A. Z.** (2000). An explanation and illustration of cluster analysis for identifying hospitality market segments. Journal of Hospitality & Tourism Research, 24(1):67–91.

[3] **Goodman, A.C. and Thibodeau, T.G.** (1998). Housing market segmentation. *Journal of housing economics*, *7*(2), pp.121-143.

whether delineating submarkets with hierarchical models improves hedonic estimates of property value.

With the increasing popularity of machine learning, recently many techniques have been developed and applied for market segmentation. Particularly recent studies have shown how hospitality can make use of hierarchical clustering (an agglomerative approach) in combination with the elbow method to segment guests and develop a marketing strategy[4]. An application of K-means algorithm[5] for marketing segmentation is described in (Hung et al, 2019) where authors tried to determine potential customer zones to make reasonable marketing strategies in a specific case study of Black Friday.

Things are a little different for online market segmentation, where clickstream data sets can be extremely useful. (Lin et al, 2021)[6] have come up with the idea of adoption of the Markov Model to represent customer visiting sessions which gives a detailed mathematical framework and analytical method to model the customer journey in a reliable and repeatable fashion. This formal expression of the customer journey enables to describe the data in a clear and comprehensive manner and use clusters to represent the market segments.

Housing market in Yerevan[7]: Real estate plays an essential role in the economy of RA. During 2020, the RA real estate market was subjected to several shocks. The first was the epidemic of a new type of coronavirus and the following waves of the epidemic that appeared in the world during 2020 and the restrictions caused by it in Armenia. The second was the war in Artsakh and the political crisis at the end of the year. Transactions in the real estate market were also affected by the anticipation of changes in the legislation on the repayment of interest on mortgage loans at the expense of income tax, in particular, the elimination of this mechanism. During 2021, a certain increase (almost 10,000) in transactions was observed in real estate market, that's the highest since 2010.

With this amount of real estate transactions, it is hard to monitor the market. This research aims to segment the housing market based on the similarity of the apartments. Here similarity is the textual descriptions of the apartments left by the owners (sellers or real-estate agents). With the use of scrapping techniques (such as Beautiful Soup) the needed dataset[8] was collected from estate.am website. The sample of the dataset is shown in Table 1.

---

[4] **van Leeuwen, R. and Koole, G.** (2021). Data-Driven Market Segmentation in Hospitality Using Unsupervised Machine Learning. *arXiv preprint arXiv:2111.02848*.

[5] **Hung, P.D., Ngoc, N.D. and Hanh, T.D.** (2019). February. K-means clustering using RA case study of market segmentation. In *Proceedings of the 2019 5th International Conference on E-Business and Applications* (pp. 100-104).

[6] **Lin, J., Holland, C.P., Argyris, N., Prinz, A. and Hengesbach, C.** (2021). A Machine Learning Approach to Online Market Segmentation. *Available at SSRN 3941093*.

[7] All the codes are available here: https://github.com/Tigran-Karamyan/house_market_segmentation

[8] Although there were no limitations related to the location of the apartments, the collected data represents the current situation in Yerevan city. One of the reasons is that online platforms are not so popular outside Yerevan.

Table 1

Sample dataset

	addr	rooms	ruler	floor	price	descr	lat	lon
0	Mashtots avenue	2 room	50 m2	floor 13/13, new	Sale - 150,000 $	This flat is in the last...	40.18	44.51
1	A. Isa-hakyan	4 room	180 m2	floor 4/4	Sale - 380,000 $	Three - bedroom flat...	40.18	44.51

Particularly the dataset contains information about the apartment description, pricing and location. Although the dataset is dynamically updated and new apartments are continuously added, for this research we used housing information for 1724 apartments and this data were used to build different models. While locations of the apartments are used for geographical segmentation, the descriptions are used for actual (product-related) housing market segmentation ignoring the prices of the apartments and focusing only on how the apartments can be separated into different clusters according to their textual description.

Methodology: Natural Language Processing (NLP) is a branch of Data Science which deals with unstructured data that is a text data in our case. But before using apartments descriptions to divide the housing market in Yerevan into different clusters, text preprocessing is needed. To prepare the text data for the model building these preprocessing steps have been applied in this particular order:

1) Removing punctuations: clearing the text data from symbols (. , ! $( ) * % @, etc.).

2) Removing stop words: stop words are a set of commonly used words in a language that carry very little useful information, so eliminating these words will clear the text data. Examples of stop words in English are “a”, “the”, “is”, “are” and etc.

3) Lower casing: during the text preprocessing apartments addresses were concatenated with the text data since the names of the streets also are somehow informative. But since case sensitivity might be a problem, it is a good practice to lower the words.

4) Tokenization: after removing all unnecessary words and lowercasing the text data, it needs to be split into words. This technique is called word tokenization.

5) Stemming and Lemmatization[9]: this step can be treated as text standardization step where the words are stemmed or diminished to their root/base form. Unlike stemming, lemmatization stems the word but makes sure that it does not lose its meaning. So here we tokenized the texts and then with stemming or lemmatization we brought back the base form of the words (in case they were transformed).

6) Word filtering with POS-tagging[10]: after stemming or lemmatization we removed all the words which corresponding particular part of speech is not noun, adjective, adverb or verb. Here we assume, that all other parts of speech carry relatively little information and eliminating them will help in model training process.

[9] Here we used both techniques but the final model was built with the lemmatized words.

[10] Part-of-speech tagging is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context.

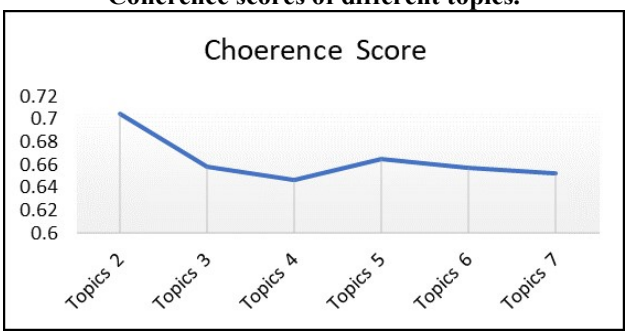
Once the text preprocessing is done it is time for Topic modeling[11]. Topic modeling uses an algorithm to discover the abstract topic or set of topics that best describes a given text document. You can think of each topic as a word or a set of words that can be considered as a class of a document. Here as a topic model is used latent Dirichlet allocation (LDA) that is a generative statistical model and allows sets of observations to be explained by unobserved groups which explain why some parts of the data are similar. It is a generalization of probabilistic latent semantic analysis and it yields better disambiguation of words and a more precise assignment of documents to topics. One should note that these assignments have probabilities as within a topic, certain terms will be used much more frequently than others. In other words, the terms within a topic will also have their own probability distribution.

There is a classical problem with non-hierarchical clustering models: how to find the optimal number of topics. This same problem goes also for LDA. To answer that question, one can compare the goodness-of-fit of LDA models fit with varying numbers of topics. As an evaluation of the goodness-of-fit of an LDA model perplexity or Coherence score can be considered. The perplexity indicates how well the model describes a set of documents. The overall Coherence score of a topic is the average of the distances between words.

Fig 1. shows Coherence scores of different topics. According to the chart we can say that 5-topics model can be considered the best among the tested models. The higher the Coherence score the better the model, though if it is higher than 0.8 the model is probably overfitted. To visualize how 1724 apartments were divided into 5 clusters, pyLDAvis is deployed (LDA topic modeling visualization package) and shown in Fig. 2a and Fig 2b.

Fig. 1

Coherence scores of different topics.



[11] **Aggarwal, C.C.** (2018). *Machine learning for text* (Vol. 848). Cham: Springer, (pp. 46-56).

Fig. 2a

Visualization of the 5-topic model

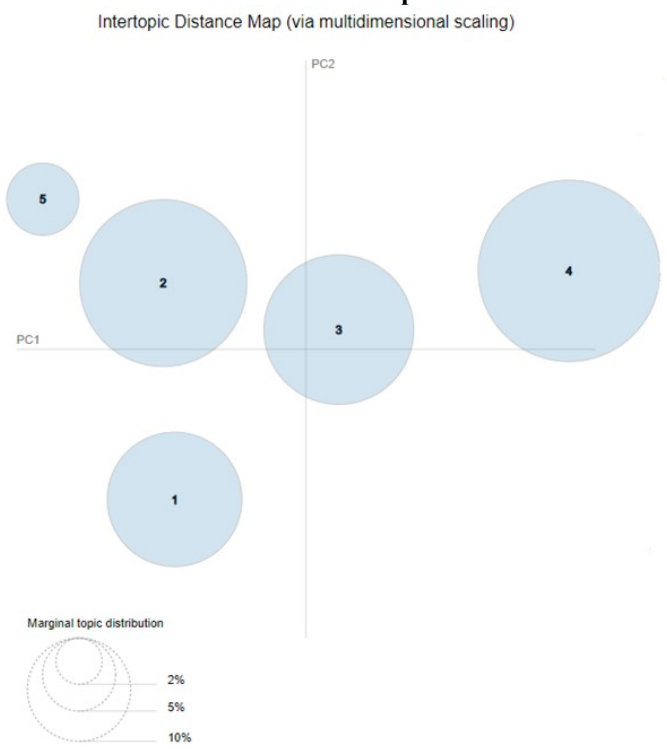
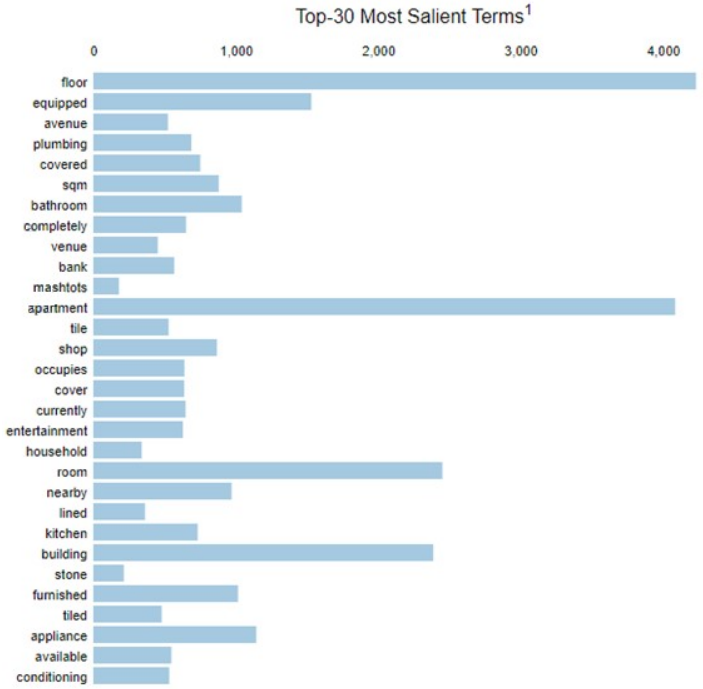


Fig. 2b

The most frequent 30 words of all the descriptions



In Fig. 2 each bubble represents a topic. The larger the bubble, the more prevalent is that topic. A topic model can be considered good if it has fairly big, non-overlapping bubbles scattered throughout the chart instead of being clustered in one quadrant.

Fig. 2b represents the top 30 frequent words of all the documents. Each cluster contains a certain set of words with some probabilities that represent a cluster. For example, cluster number 1 can be represented by the following equation:

$$Clus\,1 = 0.042 * "floor" + 0.033 * "apartment" + 0.025 * "building" + 0.024 * "room" + 0.018 * "new"$$

where the words are top 5 keywords of the topic and the numbers are weights representing how important a keyword is to that topic. Each word has its own weight. Similarly, each document (here apartment description) has a topic percentage contribution score. In other words, that score shows what's the probability (with the threshold of 0.5) of a certain apartment to be included in a certain cluster.

Results: The final 5-topic model "divided" the apartments data into 5 clusters based on the description. It is crucial to understand that the research is directly connected to consumer behavior analysis since each owner describes their apartment in a way that is as attractive as possible for the potential buyer. In other words, the apartments are clustered based on the subjective opinions of their owners while these opinions are formed by personality traits, values, motives, interests and lifestyles and can be expressed through text. This kind of market segmentation is known as psychographic segmentation as it focuses on the behavior of the people. Therefore, the 5-topic model not only defines the housing market segments by examining the similarity of the apartments, it also gives the psychological picture of the participants of the market.

In order to find some patterns within each cluster, we decided to use locations of the apartments and project them on a map. This segmentation technique is known also as geographic segmentation when one can segment the market based on variables like locations. With the help of Streamlit, the appropriate app was created that interactively shows the apartments locations, their pricing, flooring and other useful information. The app allows to dive into each cluster and find out how the apartments are connected or considered similar.

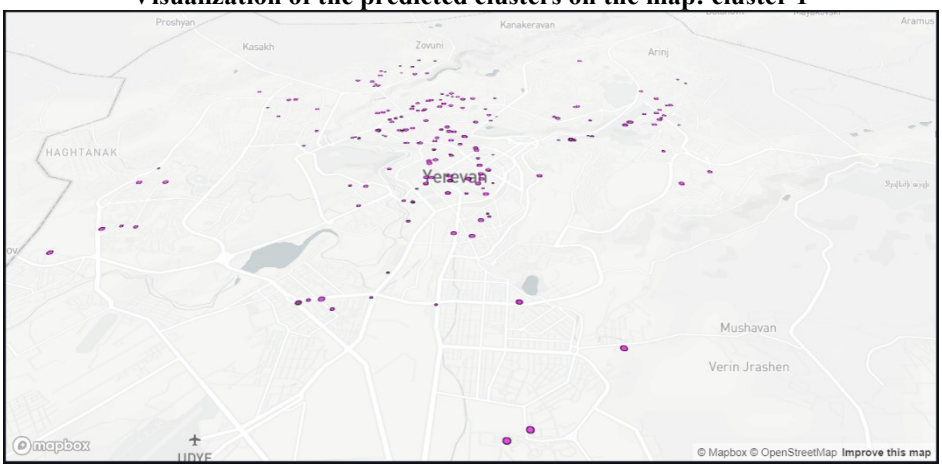
The following series of figures represent how apartments within a certain cluster are distributed in Yerevan. The bubbles represent the apartments and the size of the bubbles represents the topic contribution percentage: the bigger the bubble, the higher the probability of that apartments to be in that cluster.

Cluster 1 has the largest distribution (Fig 3a). The prices of the apartments within cluster vary mainly from 50,000$ to 120,000$ and these apartments are mainly located in old building. One can notice that there are some apartments which prices might be greater than 120,000$ (like the apartments in the Center of the city or in Arabkir community) and these apartments also have been included in this cluster. This is due to the fact that the descriptions of those

apartments are “similar”. Another interesting fact: real-estate brokers, who are also members of this market, describes the apartments similarly. For example, all the apartments located in “Multi City House” complex (in Nork-Marash community) have been included in Cluster 1.

Fig. 3a

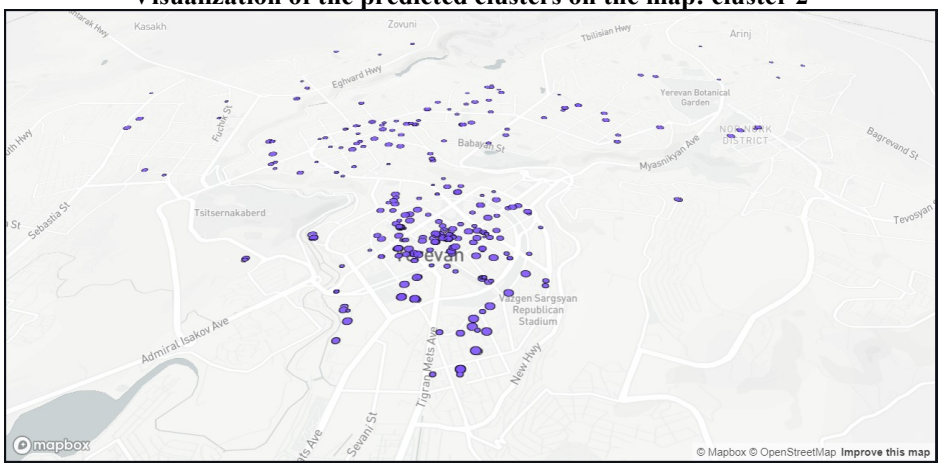
Visualization of the predicted clusters on the map: cluster 1



The apartments in Cluster 2 (Fig 3b) are mainly located in Arabkir community and in the Center of the city, therefor the prices of the apartments vary mainly form 100,000$ to 600,000$ (and even higher). It can be noticed that apartments are located mainly in new buildings and the blocks of apartments like at Northern avenue or at Aram street are included in this cluster. Though this cluster also represents the apartments in Arabkir community, one can notice that they have pretty low topic contribution rate (small bubbles), which means that these apartments are not the representative ones for this cluster.

Fig. 3b

Visualization of the predicted clusters on the map: cluster 2



Cluster 3 (Fig 3c) and Cluster 5 (Fig 3d) are relatively small in comparison with the other clusters. The prices in Cluster 3 vary mainly from 100,000$ to

300,000$. The “handwriting” of certain real-estate agents can be seen here too: the whole chunk of apartments located at Paronyan street (near Dvin Music Hall) has been included in Cluster 3. For Cluster 5 the representative apartments are the ones located along the Mesrop Mashtots avenue in old buildings with the price range from around 100,000$ to 200,000$.

Fig. 3c

Visualization of the predicted clusters on the map: cluster 3

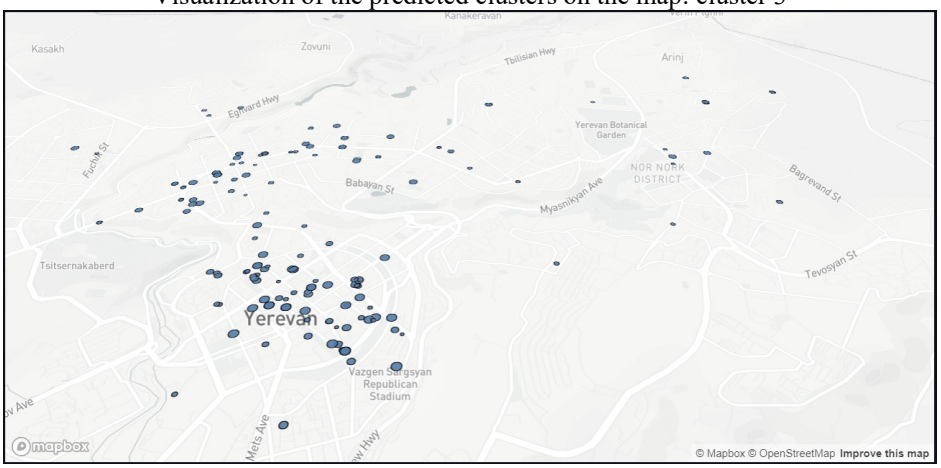
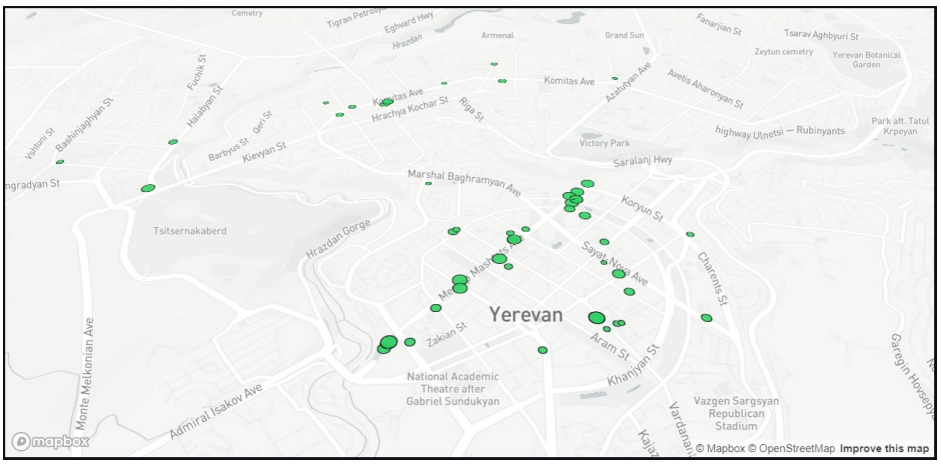


Fig. 3d

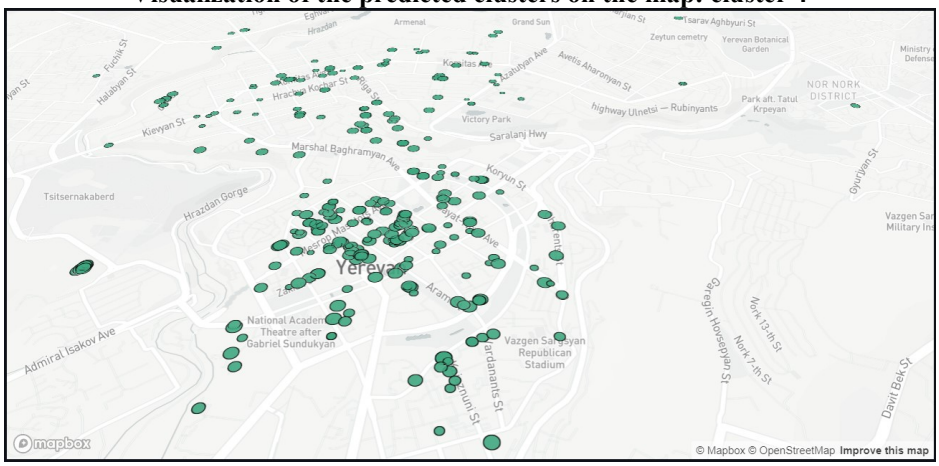
Visualization of the predicted clusters on the map: cluster 5



And, finally, Cluster 4 (Fig 3e) – one of the biggest clusters, represents the apartments located in the Center of the City and Arabkir community. Here can be found distinguishable blocks of apartments located at Verin Antarayin street, Tsitsernakaberd highway, Northern avenue, Aram and Byuzand streets and Vardanants street. All the buildings are new so prices in this cluster are relatively high and all these blocks of apartments can be considered similar in description.

Fig. 3e

Visualization of the predicted clusters on the map: cluster 4



Conclusion: The aim of this research is to determine the segments of the housing market by using the text descriptions of apartments. It can be said that the usage of unstructured data allows to use a wider range of analytical techniques for market segmentation. The result shows that topic modelling can be a good option in terms of market segmentation. The proposed model tries to divide the housing market into clusters so that the apartments with similar description are in the same cluster or segment. In order to define those clusters, the housing market in Yerevan was segmented geographically, psychologically and by description that shows the distribution of the apartments among certain clusters based on apartment descriptions. It was already mentioned, that these descriptions provide information not only about apartments but also about owners whose psychological picture can be the expressed through text. Since each "owner" describes the apartment in their own subjective way, one can observe the patterns of similarity between apartments within each segment especially after a graphical visualization on a map.

ՏԻԳՐԱՆ ՔԱՐԱՄՅԱՆ – ***Բնակարանային շուկայի սեգմենտավորում բնակարանների նկարագրության հիման վրա*** – Շուկայի սեգմենտավորումը տարբեր շուկաների՝ միատարր խմբավորումների բաժանման գործընթաց է, որոնք ունեն ընդհանուր պահանջներ, շահեր և/կամ վարքագիծ: Բնակարանային շուկան, ինչպես ցանկացած այլ շուկա, ունի իր մարտահրավերները, երբ խոսքը վերաբերում է շուկայի սեգմենտավորմանը, քանի որ վերջինիս գործընթացը կախված է ոչ միայն ապրանքից (տներ կամ բնակարաններ), այլև շուկայի խաղացողներից (սեփականատերերից կամ անշարժ գույքի գործակալներից):

Այս հոդվածը նպատակ ունի ցույց տալ, թե ինչպես կարելի է բնական լեզվի մշակումը կիրառել Երևանի բնակարանային շուկայի սեգմենտավորման համար՝ օգտագործելով անշարժ գործակալության կայքից դուրս բերված ոչ կառուցվածքային (բնակարանների նկարագրություններ) տվյալները: Հա-

վաքված տեքստային տվյալները ներկայացնում են ոչ միայն բնակարանների նկարագրությունը, այլև պատկերացում են տալիս, թե ով է գրել այդ տեքստը: Բնական լեզվի մշակման կիրառված մոդելը ցույց է տալիս, թե շուկայի խաղացողների որոշակի վարքագծային օրինաչափություններն ինչպես կարող են արտահայտվել տեքստային տվյալների միջոցով և ազդել շուկայի սեգմենտավորման վրա: Այսինքն՝ շուկայի սեգմենտավորումը ոչ կառուցվածքային տվյալների հիման վրա ներկայացնում է հաճախորդների (այստեղ՝ վաճառողների) հոգեբանական, ինչպես նաև նրանց բնակարանների «նմանության» և աշխարհագրական պատկերը:

Բանալի բառեր – *շուկայի սեգմենտավորում, ոչ կառուցվածքային տվյալներ, բնական լեզվի մշակում, թոփիկ մոդելավորում, լատենտ Դիրիխլեի ալոկացիա, քլաստերիզացիա, pyLDAvis*

ТИГРАН КАРАМЯН – *Сегментация рынка жилья на основе описания квартир.* – Сегментация рынка — это процесс разделения различных рынков на однородные группы, имеющие сопоставимые требования, интересы и/или поведение. На рынке жилья, как и на любом другом рынке, существуют свои сложности, связанные с сегментацией рынка, поскольку процесс сегментации зависит не только от продукта (дома или квартиры), но и от участников рынка (собственников или агентов по недвижимости).

Эта статья призвана показать, как можно использовать обработку естественного языка для определения сегментов рынка жилья в Ереване с помощью неструктурированных данных, взятых с веб-сайта недвижимости. Собранные текстовые данные представляют собой не только описания квартир, но и дают представление о том, кто написал этот текст. Применяемая модель обработки естественного языка показывает, как определенные модели поведения участников рынка могут быть выражены через текстовые данные и как эти модели могут влиять на сегментацию рынка. То есть сегментация рынка на основе неструктурированных данных представляет продуктовую, психографическую, а также географическую картину участников рынка (здесь – продавцов) и их квартир.

Ключевые слова: *сегментация рынка, неструктурированные данные, обработка естественного языка, тематическое моделирование, скрытое распределение Дирихле, кластеризация, pyLDAvis*