TASTE SPACE VERSUS THE WORLD: AN EMBEDDING ANALYSIS OF LISTENING HABITS AND GEOGRAPHY

Joshua L. Moore, Thorsten Joachims
Cornell University, Dept. of Computer Science
{jlmo|tj}@cs.cornell.edu

Douglas Turnbull
Ithaca College, Dept. of Computer Science
dturnbull@ithaca.edu

ABSTRACT

Probabilistic embedding methods provide a principled way of deriving new spatial representations of discrete objects from human interaction data. The resulting assignment of objects to positions in a continuous, low-dimensional space not only provides a compact and accurate predictive model, but also a compact and flexible representation for understanding the data. In this paper, we demonstrate how probabilistic embedding methods reveal the "taste space" in the recently released Million Musical Tweets Dataset (MMTD), and how it transcends geographic space. In particular, by embedding cities around the world along with preferred artists, we are able to distill information about cultural and geographical differences in listening patterns into spatial representations. These representations yield a similarity metric among city pairs, artist pairs, and city-artist pairs, which can then be used to draw conclusions about the similarities and contrasts between taste space and geographic location.

1. INTRODUCTION

Embedding methods are a type of machine learning algorithm for distilling large amounts of data about discrete objects into a continuous and semantically meaningful representation. These methods can be applied even when only contextual information about the objects, such as co-occurrence statistics or usage data, is available. For this reason and due to the easy interpretability of the resulting models, embeddings have become popular for tasks in many fields, including natural language processing, information retrieval, and music information retrieval. Recently, embeddings have been shown to be a useful tool for analyzing trends in music listening histories [6].

In this paper, we learn embeddings that give insight into how music preferences relate to geographic and cultural boundaries. Our input data is the *Million Musical Tweets Dataset* (MMTD), which was recently collected and curated by Hauger et al. [3]. This dataset consists of over a million tweets containing track plays and rich geographical information in the form of globe coordinates, which Hauger et al. have matched to cities and other geographic descriptors as well. Our goal in this work is to use embedding methods to enable a more thorough analysis of geographic and cultural patterns in this data by embedding cities and the artists from track plays in those cities into a joint space. The resulting *taste space* gives us a way to directly measure city/city, city/artist, and artist/artist affinities. After verifying the predictive fidelity of the learned taste space, we explore the surprisingly clear segmentations in taste space across geographic, cultural, and linguistic borders. In particular, we find that the taste space of cities gives us a remarkably clear image of some cultural and linguistic phenomena that transcend geography.

© Joshua L. Moore, Thorsten Joachims, Douglas Turnbull. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Joshua L. Moore, Thorsten Joachims, Douglas Turnbull. "Taste Space Versus the World: an Embedding Analysis of Listening Habits and Geography", 15th International Society for Music Information Retrieval Conference, 2014.

2. RELATED WORK

Embeddings methods have been applied to many different modeling and information retrieval tasks. In the field of music IR, these models have been used for tag prediction and song similarity metrics, as in the work of Weston et al. [7]. However, instead of a prediction task such as this, we intend to focus on data analysis tasks. Therefore, we rely on generative models like those proposed in our previous work [5, 6] and by Aizenberg et al [1]. Our prior work uses models which rely on sequences of songs augmented with social tags [5] or per-user song sequences with temporal dynamics [6]. The aim of this work differs from that of our previous work in that we are interested in aggregate global patterns and not in any particular playlist-related task, so we do not adopt the notion of song sequences. We also are concerned with geographic differences in listening patterns, and so we ignore individual users in favor of embedding entire cities into the space.

Aizenberg et al. utilize generative models like those in our work for purposes of building a recommendation engine for music from Internet radio data on the web. However, their work focuses on building a powerful recommendation system using freely available data, and does not focus on the use of the resulting models for data analysis, nor do they concern themselves with geographic data.

The data set which we will use throughout this work was published by Hauger et al. [3]. The authors of this work crawled Twitter for 17 months, looking for tweets which carried certain key words, phrases, or hashtags in order to find posts which signal that a user is listening to a track and for which the text of the tweet could be matched to a particular artist and track. In addition, the data was selected for only tweets with geographical tags (in the form

of GPS coordinates), and temporal data was retained. The final product is a large data set of geographically and temporally tagged music plays. In their work, the authors emphasize the collection of this impressive data set and a thorough description of the properties of the data set. The authors do add some analyses of the data, but the geographic analysis is limited to only a few examples of coarse patterns found in the data. The primary contribution of our work over the work presented in that paper is to greatly extend the scope of the geographic analysis, presenting a much clearer and more exhaustive view of the differences in musical taste across regions, countries, and languages.

Finally, we describe how geographic information can be useful for various music IR tasks. Knopke [4] also discusses how geospatial data can be exploited for music marketing and musicological research. We use embedding as a tool to further explore these topics. Others, such as Lamere's *Roadtrip Mixtape* [1] app, have developed systems that use a listeners location to generate a playlist of relevant music by local artists.

3. PROBABILISTIC EMBEDDING MODEL

The embedding model used in this paper is similar to the one used in our previous work [6]. However, the following analysis focuses on geographical patterns instead of temporal dynamics and trends. In particular, we focus on the relationships among cities and artists, and so we elect to condense the geographical information in a tweet down to the city from which it came. Similarly, we discard the track name from each tweet and use only the artist for the song. This leads to a joint embedding of cities and artists.

At the core of the embedding model lies a probabilistic link function that connects the observed data to the underlying semantic space. Intuitively, the link function we use states that the probability $\Pr(a|c)$ of a given city c playing a given artist a is proportional to the distance $||X(c) - Y(a)||_2^2$ between that city and that artist in a Euclidean embedding space of a chosen dimension d. $X(c)$ and $Y(a)$ are the embedding locations of city c and artist a respectively. Similar to previous works, we also incorporate a popularity bias term p_a for each artist to model global popularity. More formally, the probability for a city c to play an artist a is:

$$\Pr(a|c) = \frac{\exp(-||X(c) - Y(a)||_2^2 + p_a)}{\sum_{a' \in A} \exp(-||X(c) - Y(a')||_2^2 + p_{a'})}.$$

The sum in the denominator is over the set A of artists. This sum is known as the *partition function*, denoted $Z(\cdot)$, and serves to normalize the distribution over artists.

Determining the embedding locations $X(c)$ and $Y(a)$ for all cities and artists (and the popularity terms p_a) is the learning problem the embedding method must solve. To fit a model to the data, we maximize the log-likelihood formed by the sum of log-probabilities $\log(\Pr(a_i|c_i)$:

$$\begin{aligned}(X,Y,p) &= \max_{X,Y,p} \sum_{(c_i,a_i)\in D} \log(\Pr(a_i|c_i)) \\ &= \max_{X,Y,p} \sum_{(c_i,a_i)\in D} -||X(c_i) - Y(a_i)||_2^2 + p_{a_i} - \log(Z(a_i)).\end{aligned}$$

We solve this optimization problem using a Stochastic Gradient Descent approach. First, each embedding vector $X(\cdot)$ and $Y(\cdot)$ is randomly initialized to a point in the unit ball in $\mathbb{R}^d$ (for the chosen dimension d). Then, the model parameters are updated in sequential stochastic gradient steps until convergence. The partition function $Z(\cdot)$ presents an optimization challenge, in that a naïve optimization strategy requires $O(|A|^2)$ time for each pass over the data. For this work, we used our C++ implementation of the efficient training method employed in [6], an approximate method that estimates the partition function for efficient training. This implementation is available by request, and will later be available on the project website, `http://lme.joachims.org`.

3.1 Interpretation of Embedding Space

As defined above, the model gives us a joint space in which both cities and artists are represented through their respective embedding vectors $X(\cdot)$ and $Y(\cdot)$. Related works have found such embedding spaces to be rich with semantic significance, compactly condensing the patterns present in the training data. Distances in embedding space reveal relationships between objects, and visual or spatial inspection of the resulting models quickly reveals a great deal of segmentation in the space. In particular, joint embeddings yield similarity metrics among the various types of embedded objects, even though individual dimensions in the embedding space have no explicit meaning (e.g. the embeddings are rotation invariant). In our case, this specifically entails the following three measures of similarity:

City to Artist: this is the only similarity metric explicitly formulated in the model, and it reflects the distribution $\Pr(a|c)$ that we directly observe data for. In particular, we directly optimize the positions of cities and artists so that cities have a high probability of listening to artists which they were observed playing in the dataset. This requires placing the city and artist nearby in the embedding space, so proximity in the embedding space can be interpreted as an affinity between a city and an artist.

Artist to Artist: due to the learned conditional probability distributions' being constrained by the metric space, two artists which are placed near each other in the space will have a similar probability mass in each city's distribution. This implies a kind of exchangeability or similarity, since any city which is likely to listen to one artist is likely to listen to the other in the model distribution.

City to City: finally, the form of similarity on which we will most rely in this work is that among cities. Again due to the metric space, two nearby cities will assign similar masses to each artist, and so will have very similar distributions over artists in the model. This implies a similarity in musical taste or preferred artists between two cities.

The third type of similarity will form the basis for most of the analyses in this paper. In particular, we are interested

[1] http://labs.echonest.com/CityServer/roadtrip.html

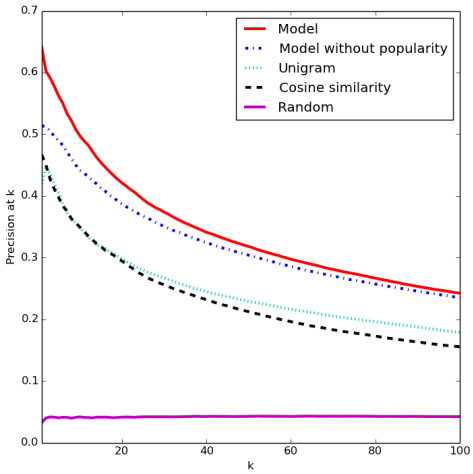

Figure 1: Precision at k of our model, a cosine similarity baseline, a tweet count ranking baseline, and a random baseline on a city/artist tweet prediction task.

in the connection between the metric space of cities in the embedding space and another metric space: the one formed by the geographic distribution of cities on the Earth's surface. As we will see, these two spaces differ greatly, and the taste space of cities gives us a clear image of some cultural and linguistic phenomena that transcend geography.

4. EXPERIMENTS

We use the MMTD data set presented by Hauger et al. [3]. This data set contains nearly 1.1 million tweets with geographical data. We pre-process the data by condensing each tweet to a city/artist pair, which results in a city/artist affinity matrix used to train the model. Next, we discard all cities and artists which have not appeared at least 100 times in the data, as well as all cities for which fewer than 30 distinct users tweeted from that city. The post-processed data contains 1,017 distinct cities and 1,499 distinct artists.

For choosing model parameters, we randomly selected 80% of the tweets for the training set, and the remaining 20% for the validation set. This resulted in a training set of 390,077 tweets and a validation set of 97,592 tweets. We used the validation set both to determine stopping criteria for the optimization as well as to choose the initial stochastic gradient step size η_0 from the set $\{0.25, 0.1, 0.05, 0.01\}$ and to evaluate the quality of models of dimension $\{2, 50, 100\}$. The optimal step size varied from model to model, but the 100-dimensional model consistently out-performed the others (although the difference between it and the 50-dimensional model was small).

We will analyze the data through the trained embedding models, both through spatial analyses (i.e. nearest neighbor queries and clusterings) and through visual inspection. In general, the high-dimensional model better captures the data, and so we will use it when direct visual inspection is not required. But first, we evaluate the quality of the model through quantitative means.

4.1 Quantitative Evaluation of the Model

Before we inspect our model in order to make qualitative claims about the patterns in the data, we first wish to evaluate it on a quantitative basis. This is essential in order to confirm that the model accurately captures the relations among cities and artists, which will offer validation for the conclusions we draw later in the work.

4.1.1 Evaluating Model Fidelity

First, we considered the performance of the model in terms of perplexity, which is a reformulation of the log-likelihood objective outside of a log scale. This is a commonly used measure of performance in other areas of research where models similar to ours are used, such as natural language processing [2]. The perplexity p is related to the average log-likelihood L by the transformation $p = \exp(-L)$.

Our baseline is the unigram distribution, which assumes that $\Pr(a|c)$ is directly proportional to the number of tweets artist a received in the entire data set independent of the city. Estimating the unigram distribution from the training set and using it to calculate the perplexity on the validation set yielded a perplexity of 589 (very similar to the perplexity attained when estimating this distribution from the train set and calculating the perplexity on the train set itself). Our model offered a great improvement over this – the 100-dimensional model yielded a perplexity on the validation set of 290, while the 2-dimensional model reached a perplexity of 357. This improvement suggests that our model has captured a significant amount of useful information from the data.

4.1.2 Evaluating Predictive Accuracy

Second, we created a task to evaluate the predictive power of our model. To this end, we split the data chronologically into two halves, and further divided the first half into a training set and a validation set. Using the first half of the data, we trained a 100-dimensional model. Our goal is to use this model to predict which new artists various cities will begin listening to in the second half of the data.

We accomplish this by considering, for each city, the set of artists which had no observed tweets in that city in the first half of the data. We then sorted these artists by their score in the model – namely, for city c and artist a, the function $-||X(c) - Y(a)||_2^2 + p_a$. Using this ordering as a ranking function, we calculated the precision at k of our ranking for various values of k, where an artist is considered to be relevant if that artist receives at least one tweet from that city in the second half of the data. We average the results of each city's ranking.

We compare the performance of our model on this task to three baselines. First, we consider a *random* ranking of all the artists which a city has not yet tweeted. Second, we sort the yet untweeted artists by their raw global tweet count in the first half of the data – which we label the *unigram* baseline. Third, we use the raw artist tweet counts for a city's nearest neighbor city in the first half of data to rank untweeted artists for that city. In this case, the nearest

neighbor is not determined using our embedding but rather based on the maximum *cosine similarity* between the vector of artist tweet counts for the city and the vectors of tweet count for all other cities.

The results can be seen in Figure 1. At $k = 1$, our model correctly guesses an artist that a city will later tweet with 64% accuracy, compared to 46% for the cosine similarity, 42% for unigram and around 5% for the random baseline. This advantage is consistent as k increases, with our method attaining about 24% precision at 100, compared to 18% for unigram and 14% for cosine similarity. We also show the performance of the same model at this task when popularity terms are excluded from the scoring function at ranking time. Interestingly, the performance in this case is still quite good. We see precision at 1 of about 51% in this case, with the gap between this method and the method with popularity terms growing smaller as k increases. This suggests that proximity in the space is very meaningful, which is an important validation of the analyses to follow. Finally, the good performance on this task invites an application of the space to making marketing predictions – which cities are prone to pick up on which artists in the near future? – but we leave this for future work.

4.2 Visual Inspection of the Embedding Space

In Figure 2 we present plots of the two-dimensional embedding space, with labels for some key cities (left) and artists (right). Note that the two plots are separated by city and artists only for readability, and that all points lie in the same space. In this figure, we can already see a striking segmentation in city space, with extreme distinction between, e.g., Brazilian cities, Southeast Asian cities, and American cities. We can also already see distinct regional and cultural groupings in some ways – the U.S. cities largely form a gradient, with Chicago, Atlanta, Washington, D.C., and Philadelphia in the middle, Cleveland and Detroit on one edge of the cluster, and New York and Los Angeles on the opposite edge. Interestingly, Toronto is also on the edge of the U.S. cluster, and on the same edge where New York and Los Angeles – arguably the most "international" of the U.S. cities shown here – end up.

It is also interesting to note that the space has a very clear segmentation in terms of genre – just as clear as embeddings produced in previous work from songs alone [5] or songs and individual users [6]. Of course, this does not translate into an effective user model – surely there are many users in Recife, Brazil that would quickly tire of a radio station inspired by Linkin Park – but we believe it is still a meaningful phenomenon. Specifically, this suggests that the taste of the average listener can vary dramatically from one city to the next, even within the same country. More surprisingly, this variation in the average user is so dramatic that cities themselves can form nearly as coherent a taste space as individual users, as the genre segmentation is barely any less clear than in other authors' work with user modeling.

4.3 Higher-dimensional Models

Directly visualizing two-dimensional models can give us striking images from which rough patterns can be easily gleaned. However, higher dimensional models are able to achieve perplexities on the validation set which far exceed those of lower dimensional models. For example, as mentioned before, our best performing 2-dimensional model attains a validation perplexity of 357, while our best performing 100-dimensional model attains a perplexity of 290 on the validation set. This suggests that higher dimensional models capture more of the nuanced patterns present in the data. On the other hand, simple plotting is no longer sufficient to inspect high-dimensional data – we must resort to alternative methods, for example, clustering and nearest neighbor queries. First, in Figure 3, we present the results of using k-means clustering in the city space of the 100-dimensional model. The common algorithm for solving the k-means clustering problem is known to be prone to getting stuck in local optima, and in fact can be difficult to validate properly. We attempted to overcome these problems by using cross validation and repeated random restarts. Specifically, we used 10-fold cross-validation on the set of all cities in order to find a validation objective for each candidate value of k from 2 to 20. Then, we selected the parameter k by choosing the largest value for which no larger value offers more than a 5% improvement over the immediately previous value.

Once the value of k was chosen, we tried to overcome the problem of local optima by running the clustering algorithm 10 times on the entire set of cities with that value of k and different random initializations, finally choosing the trial with the best objective value. This process resulted in optimal k values ranging from 6 to 13. Smaller values resulted in some clusterings with granularity too coarse to see interesting patterns, while larger values were noisy and produced unstable clusterings. Ultimately, we found that $k = 9$ was a good trade-off.

Additionally, in Table 1, we obtain a complementary view of the 100-dimensional embedding by listing the results of nearest-neighbor queries for some well-known, hand-selected cities. These queries give us an alternative perspective of the city space, pointing out similarities that may not be apparent from the clustering alone. By combining these views, we can start to see many interesting patterns arise:

The French-speaking supercluster: French-speaking cities form an extremely tight cluster, as can also be seen in the 2-dimensional embedding in Figure 2. Virtually every French city is part of this cluster, as well as French-speaking cities in nearby European countries, such as Brussels and Geneva. Indeed even beyond the top 10 listed in Table 1, almost all of the top 100 nearest neighbors for Paris are French-speaking. Language is almost certainly the biggest factor in this effect, but if we consider the countries near France, we see that despite linguistic divides, in the clustering, many cities in the U.K. still group closely with Dutch cities and even Spanish cities. Furthermore, this grouping can be seen in every view of the data – in the two-dimensional space, the clustering, and the nearest neighbor queries. It should be noted that in our own trials clustering the data, the French cluster is one of the first

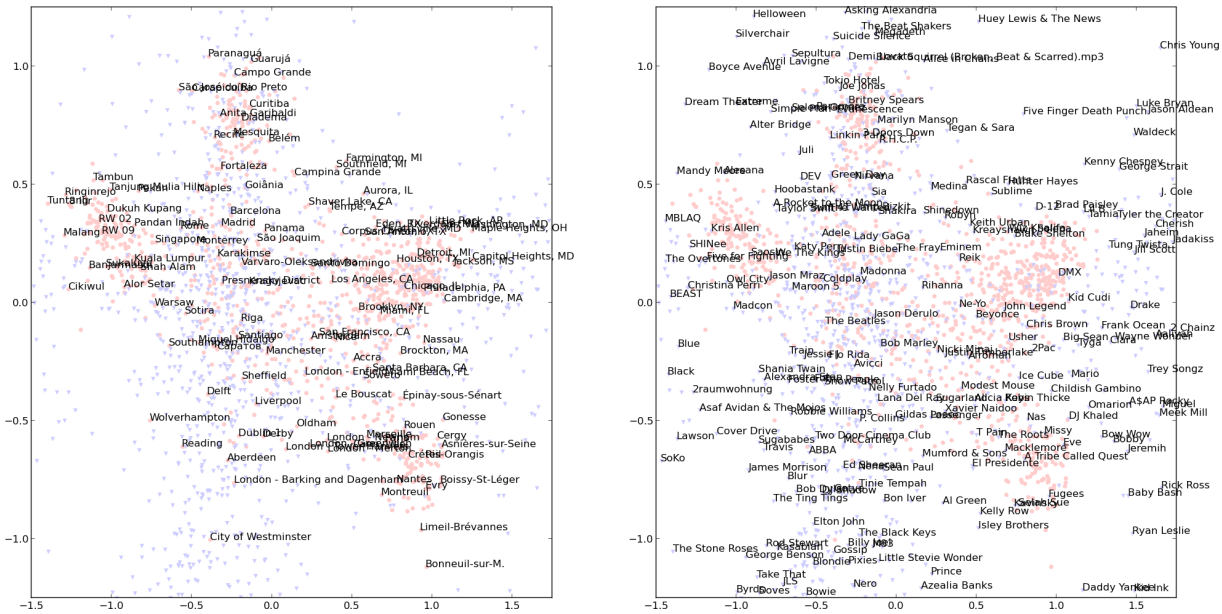

Figure 2: The joint city/artist space with some key cities and artists labeled.

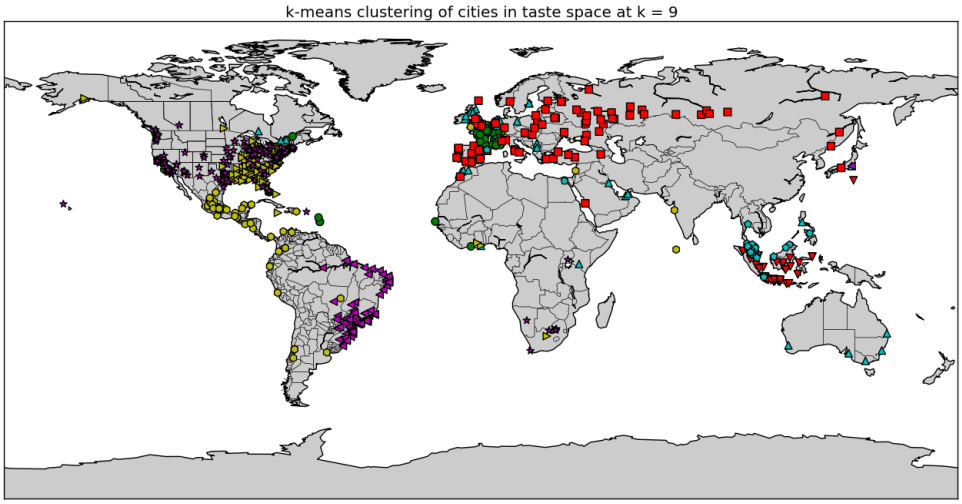

Figure 3: A k-means clustering of cities around the world with $k = 9$.

Kuala Lumpur	**Paris**	**Singapore**	**Los Angeles, CA**	**Chicago, IL**	**São Paulo**
Kulim	Boulogne-Billancourt	Hougang	Grand Prairie, TX	Buffalo, NY	Osasco
Sungai Lembing	Brussels	Seng Kang	Ontario, CA	Clarksville, TN	Jundiaí
Ipoh	Rennes	USJ9	Riverside, CA	Cleveland, OH	Carapicuíba
Kuching	Lille	Subang	Sacramento, CA	Durham, NC	Ribeirão Pires
Sunway City	Aix-en-Provence	Kota Bahru	Salinas, CA	Birmingham, AL	Shinjuku
Seremban	Limoges	Bangkok	Paterson, NJ	Flint, MI	Vargem Grande Paulista
Seri Kembangan	Amiens	Alam Damai	San Bernardino, CA	Montgomery, AL	Santa Maria
Taman Cheras Hartamas	Marseille	Kota Padawan	Inglewood, CA	Nashville, TN	Itapevi
Kuantan	Geneva	Glenmarie	Modesto, CA	Jackson, MS	Cascavel
Selayang	Grenoble	Budapest	Pomona, CA	Paterson, NJ	Embu das Artes
Brooklyn, NY	**Atlanta, GA**	**Madrid**	**Amsterdam**	**Sydney**	**Montréal**
Minneapolis, MN	Savannah, GA	Sevilla	Eindhoven	Toronto	Montpellier
Winston-Salem, NC	Tallahassee, FL	Granada	Tilburg	Denver, CO	Geneva
Arlington, VA	Cleveland, OH	Barcelona	Emmen	Windhoek	Raleigh, NC
Waterbury, CT	Washington, DC	Murcia	Nijmegen	Angers	Limoges
Washington, DC	Memphis, TN	Sorocaba	Enschede	Rialto, CA	Angers
Syracuse, NY	Flint, MI	Ponta Grossa	Zwolle	Hamilton	Ontario, CA
Jersey City, NJ	Huntsville, AL	Huntington Beach, CA	Amersfoort	Rotterdam	Anchorage, AK
Louisville, KY	Montgomery, AL	Istanbul	Maastricht	Ottawa	Nice
Tallahassee, FL	Jackson, MS	Vigo	Antwerp	London - Tower Hamlets	Lyon
Ontario, CA	Lafayette, LA	Oxford	Coventry	London - Southwark	Rennes

Table 1: Nearest neighbor query results in 100-dimensional city space. Brooklyn was chosen over New York, NY due to having more tweets in the data set. In addition, only result cities with population at least 100,000 are displayed.

Country	Least typical	Most typical
Brazil	Criciúma, Santa Catarina	Itapevi, São Paulo
Canada	Surrey, BC	Toronto, ON
Netherlands	Leiden	Emmen
Mexico	Campeche, CM	Cuauhtémoc, DF
Indonesia	Panunggangan Barat	RW 02
France	Bordeaux	Mantes-la-Jolie, Île-de-France
United States	Huntington Beach, CA	Jackson, MS
Malaysia	Kota Damansara	Kuala Lumpur
United Kingdom	Wolverhampton, England	London Borough of Camden
Russia	Ufa	Podgory
Spain	Álora, Andalusia	Barcelona

Table 2: Most and least typical cities in taste profile for various countries.

clusters to become apparent, as well as one of the most consistent to appear. We can also see that the French cluster is indeed a linguistic and cultural one which is not just due to geographic proximity: although Montreal has several nearest neighbors in North America, it is present in the French group in the k-means clustering (as is Quebec City) and is also very close to many French-speaking cities in Europe, such as Geneva and Lyon. We can also see that Abidjan, Ivory Coast joins the French k-means cluster, as do Dakar in Senegal, Les Abymes in Guadeloupe and Le Lamentin and Fort-de-France in Martinique – all cities in countries which are members of the Francophonie.

Australia: Here again, despite the relatively tight geographical proximity of Australia and Southeast Asia, and the geographic isolation of Australia from North America, Australian cities tend to group closely with Canadian cities and some cities in the United Kingdom. One way of seeing this is the fact that Sydney's nearest neighbors include Toronto, Hamilton, Ontario, Ottawa, and two of London's boroughs. In addition, other cities in Australia also belong to a cluster that mainly includes cities in the Commonwealth (e.g., U.K., Canada).

Cultural divides in the United States: the cities in the U.S. tend to form at least two distinct subgroups in terms of listening patterns. One group contains many cities in the Southeast and Midwest, as well as a few cities on the southern edge of what some might call the Northeast (Philadelphia, for example). The other group consists primarily of cities in the Northeast, on the West Coast, and in the Southwest of the country, including most of the cities in Texas. Intuitively, there are two results that might be surprising to some here. The first is that the listening patterns of Chicago tend to cluster with listening patterns in the South and the rest of the Midwest, and not those of very large cities on the coasts (after all, Chicago is the third-largest city in the country). The second is that Texas groups with the West Coast and Northeast, and not with the Southeast, which would be considered by many to be more culturally similar in many ways.

4.4 Most and least typical cities

We can also consider the relation of individual cities to their member countries. For this analysis, we considered all the countries which have at least 10 cities represented in the data. Then for each country we calculated the average position in embedding space of cities in that country. With this average city position, we can then measure the distance of individual cities from the mean of cities in their country and find the cities which have the most and least typical taste profiles for that country.

The results are shown in Table 2. We can see a few interesting patterns here. First, in Brazil, the most typical city is an outlying city near São Paulo city, while the least typical is a city in Santa Catarina, the second southernmost state in Brazil, which is also less populous than the southernmost, Rio Grande do Sul, which was also well-represented in the data. In Canada, the least typical city is an edge city on Vancouver's east side, while the most typical is the largest city, Toronto. In France, the most typical city is in Île-de-France, not too far from Paris. We also see in England that the least typical city is Wolverhampton, and edge city of Birmingham towards England's industrial north, while the most typical is a borough of London.

5. CONCLUSIONS

In this work, we learned probabilistic embeddings of the Million Musical Tweets Dataset, a large corpus of tweets containing track plays which has rich geographical information for each play. Through the use of embeddings, we were able to easily process a large amount of data and sift through it visually and with spatial analysis in order to uncover examples of how musical taste conforms to or transcends geography, language, and culture. Our findings reflect that differences in culture and language, as well as historical affinities among countries otherwise separated by vast distances, can be seen very clearly in the differences in taste among average listeners from one region to the next. More generally, this paper shows how nuanced patterns in large collections of preference data can be condensed into a taste space, which provides a powerful tool for discovering complex relationships. *Acknowledgments:* This work was supported by NSF grants IIS-1217485, IIS-1217686, IIS-1247696, and an NSF Graduate Research Fellowship.

6. REFERENCES

[1] N. Aizenberg, Y. Koren, and O. Somekh. Build your own music recommender by modeling internet radio streams. In *WWW*, pages 1–10. ACM, 2012.

[2] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *JMLR*, 3:1137–1155, 2003.

[3] D. Hauger, M. Schedl, A. Košir, and M. Tkalčič. The million musical tweets dataset - what we can learn from microblogs. In *ISMIR*, 2013.

[4] I. Knopke. Geospatial location of music and sound files for music information retrieval. *ISMIR*, 2005.

[5] J. L. Moore, S. Chen, T. Joachims, and D. Turnbull. Learning to embed songs and tags for playlist prediction. In *ISMIR*, 2012.

[6] J. L. Moore, Shuo Chen, T. Joachims, and D. Turnbull. Taste over time: the temporal dynamics of user preferences. In *ISMIR*, 2013.

[7] J. Weston, S. Bengio, and P. Hamel. Multi-tasking with joint semantic spaces for large-scale music annotation and retrieval. *JNMR*, 40(4):337–348, 2011.