MASTERS THESIS

Analysis of Toponym Co-occurrences on Social Media

Author:
Kevin O'DRISCOLL

Supervisor:
Dr Evert J. MEIJERS
Second Supervisor:
Tongjing WANG

A thesis submitted in fulfillment of the requirements for the degree of MSc Applied Data Science

in the

CITYNET Europe Group
Department of Human Geography and Spatial Planning

July 17, 2022

UTRECHT UNIVERSITY

Abstract

Department of Human Geography and Spatial Planning

MSc Applied Data Science

Analysis of Toponym Co-occurrences on Social Media

by Kevin O'DRISCOLL

Cities exist as nodes in a network and today, more than ever, network embeddedness is of critical importance in understanding how they develop. There was a time when a network of cities was highly clustered and relationships were limited to geographical neighbors. As recreational travel grew and communication technology emerged, the network topology of cities changed. People could build their networks and migrate with greater ease. Today, a large part of human interaction and relationship building occurs online, often in an informal and colloquial forum. This paper aims to apply the toponym co-occurrence method in a novel way to online conversations in order to determine how well we can to explain the co-occurrence of city names in social media. To this end, we will focus on European cities and the social media website Reddit. There are some difficulties with disambiguation as the name of a city can often take on multiple meanings. We will approach this task in three stages. First, we will explore the data to get a better understanding of toponym co-occurrences. Then we will regression techniques to determine how well the information collected can predict variability in co-occurrence of city names. Finally, we will use unsupervised machine learning techniques to assess how well we can disambiguate between different word senses when searching for toponyms.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

The development of a city is the result of a complex aggregation of factors. Progress is determined largely by network embeddedness and ties with other regions which dictate the flow of goods, people, and information in and out of the city. The nature of this development has changed drastically in recent history as information has become more central to global society. The influence of spatial factors on the cultivation of relationships between cities has diminished drastically. Improvements in infrastructure and methods of travel coupled with communication technology have altered the network topology of cities throughout the world, resulting in a quicker diffusion of information and strengthening of ties which was not previously possible.

Toponym co-occurrence is a relatively novel approach to measuring the strength of relationships between geographic entities leveraging the vast amount of accessible data on the internet. "This approach builds the urban system on the basis of co-occurrences of place names in a text corpus" (Meijers and Peris, 2019). The application of toponym co-occurrence reflects a paradigm shift in a world increasingly connected by the flow of information. Social media interactions represent a new frontier and novel application of this method. A pattern of co-occurrences of a pairs of city names in social media conversations could provide additional insight into the strength and nature of relationships between cities and contribute to a more complete representation of the way cities are connected.

The social media platform Reddit will be the focus of this study. Reddit hosts online communities called subreddits. Users submit content such as images, videos, and links within a subreddit and the submissions are discussed in a corresponding comment section. Reddit calls itself "the front page of the internet" with its home page ostensibly functioning as a news aggregation website tailored to the user. They can subscribe to different subreddits and the website tracks browsing history and location to personalize recommendations on the user's front page. The corpus used in this study is a comprehensive data dump of all Reddit comments and metadata posted at the time of retrieval.

In this study we will determine how well variability in toponym co-occurrences on Reddit can be explained by population, geographical distance, comment metadata, and other engineered features.

Toponym disambiguation will be addressed with unsupervised learning techniques to explore the possibility of alternative meanings for a given search term. In the case that there is one other dominant meaning for a given search term, results can be misleading. For instance, 'cologne' can refer to a city in Germany or a fragrance marketed toward men. This study aims to develop a novel technique to identify these cases and reduce the noise in our model due to ambiguity of expressions.

1.2 Literature Overview

Online interactions are not constrained by distance to the extent that other measures of relatedness between cities are. Tobler's first law of geography states "everything is related to everything else, but near things are more related than distant things." Does this principle apply to online, informal measures of relatedness? In order to explore this further, we will use a gravity model to see if toponym co-occurrence frequency decays with distance and size. "In its generic form, the use of gravity model is to predict the interaction intensity given the city sizes and the distances" (Hao Guo and Liu, 2022). The gravity model is derived from Newton's law of universal gravitation, which defines the gravitational force between two bodies as inversely proportional to the square distance between their centers and proportional to the product of the bodies' masses (see (Lenormand, 2016)).

Word sense disambiguation (WSD) is the process of determining which meaning of a word is being used in a given context. "In order to determine the correct resolution of an ambiguous concept it is necessary to consider its context, whilst this context is most readily provided by the other information contained in the content" (Ireson and Ciravegna, 2010). Social media presents a unique set of challenges compared to a more formal corpus like a CommonCrawl archive or Wikipedia. In addition to WSD, issues such as alternate spellings, abbreviations, colloquialisms, punctuation, and typographical errors are more prevalent. The majority of typographical errors identified in patient demographic records by (Sun YC, 2002) were single character omission and single character replacement. Even minor errors such as these are computationally expensive to disambiguate. The effect of this bias would also likely vary with time as the internet evolved along with auto correct algorithms.

Our goal here is to apply machine learning techniques to the former issue to classify a potential toponym by it's true meaning. For this we will focus on the narrow case of a place name having multiple meanings. Other sources of bias, including different places having the same place name or places lending their name to the territories of which they are part of are not in the scope of this study. A more complete list of toponym disambiguation issues can be found in (Meijers and Peris, 2019).

Approaches to disambiguation include supervised, knowledge based approach, and unsupervised approaches. For this study a supervised approach would require annotated comments, or manual labeling for the Reddit data. Knowledge based approaches use a lexicon to identify word sense, which is also not available (Sankar, Raj, and Jayan, 2016). The Reddit data is unlabeled so we will attempt toponym disambiguation with unsupervised learning techniques. When we identify a potential toponym there is no way to know the word sense without examining context. This task is perhaps better suited to a human being than to a machine, although with social media it is often the case that "people are unable to retrieve their own content due to their inconsistent descriptions" (Ireson and Ciravegna, 2010).

We will need to collect and codify context contained within the body of a comment. In (Sankar, Raj, and Jayan, 2016) the authors developed a WSD technique in which they identified a set of 'seed' context words which could be used to identify the word sense of an ambiguous word. These seeds were generated from a dictionary dataset, and the appearance of seed words in the neighborhood of an ambiguous word were used for classification. We will follow a similar approach, but instead of relying on a dictionary to generate seeds, we will focus on a single city and generate an n-gram, a set of words appearing before and after the city name, to capture context. We will then use machine learning techniques to assign the word to a cluster in an attempt to disambiguate between different meanings.

1.3 Research Question

How much of the variability in toponym co-occurrences on social media can be explained with the available data? To answer this question the focus will be on the social media platform Reddit and will include European cities with a population size of at least 1 million people. Toponym disambiguation will be examined with unsupervised learning techniques.

Chapter 2

Data

2.1 Pushift Data

This study was performed on user submission data from the website Reddit. The corpus was collected from Pushshift (*Pushshift* n.d.), a data collection and archiving platform. It consists of all user comments from December 2005 until January 2021. Reddit is a social media website which hosts online communities called subreddits. Users submit content such as images, videos, and links within a subreddit and the submissions are discussed in a corresponding comment section. Reddit calls itself "the front page of the internet" with its front page ostensibly functioning as a news aggregation website tailored to the person using it. The user can subscribe to different subreddits and the website tracks browsing history and location to personalize recommendations on the front page. The corpus used in this study is a comprehensive data dump of all reddit comments and metadata posted at the time of retrieval. The hierarchical structure of the Reddit data can be seen in Figure 2.1.

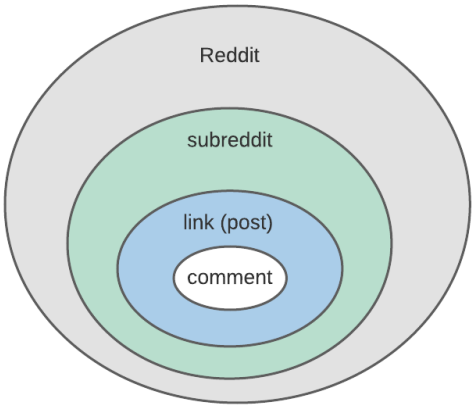


FIGURE 2.1: Reddit Organization

The subset of European cities included in this study was chosen based on population density. Morphological Urban Area (MUA) populations were used to determine the population of a given city and apply a threshold of inclusion (*ESPON 1.4.3 – Final Report* 2007). MUAs can be considered the cultural core of a city and are represented by municipalities within the true urban landscape. They are a subset of larger Functional Urban Areas (FUA) which typically include the sprawling labor basin for the city. For this study, cities with a population greater than 1 million were included.

TABLE 2.1: Description of Pushshift Variables.

Variable	Description	Applications
created-utc	time the comment was posted	temporal analysis
subreddit	community which contains the post	classification, heterogeneity
body	text of the comment	natural language processing
link-id	post identifier	hierarchical cluster analysis
score	sum of upvotes and downvotes	sentiment analysis

All data wrangling and analysis were done in Python with Jupyter Notebook. A data dump of comments from 2006 to 2020 organized by month were analyzed. Compressed files for each month were available for download in .zst, .bz2, or .xz file formats totaling approximately 828 gigabytes before decompression. The files were decompressed into JSON objects and the relevant metadata was stored in .csv files. Variables available in the dump varied with time, so only those which were relevant to our goals and available for every month were considered for inclusion. Some of the variables include 'created-utc', 'gilded', 'subreddit-id', 'distinguished', 'score-hidden', 'name', 'subreddit', 'id', 'author-flair-css-class', 'author', 'retrieve-on', 'body', 'edited', 'downs', 'link-id', 'author-flair-text', 'ups', 'archived', 'controversiality', 'score', and 'parent-id'. Table 2.1 shows the relevant factors that were compiled for analysis along with a description and application use. Some comments were deleted prior to the time of retrieval by the user. In those cases, the body of the comment is 'deleted.'

2.2 Feature Engineering

In addition to the Pushshift data, additional explanatory variables were constructed from existing data for regression and disambiguation purposes. Year and month variables were created from the file names in the data dump, which were named according to the following convention: "RC-YYYY-MM.txt". Dummy variables were created to encode categorical variables in both models, to encode a toponym occurrence in a comment and the categorical variable 'subreddit' to account for heterogeneity in co-occurrences due to subreddit.

Context is important for disambiguation, and for this we measured sentiment and subjectivity of a set of words neighboring a city name. All text analysis included preprocessing steps of punctuation removal, stop words removal, tokenization, and lemmatization. We considered TextBlob and Vader Python libraries for sentiment analysis. "TextBlob is a powerful Natural Language Processing (NLP) library for Python, which is built upon NLTK and provides an easy to use interface to the NLTK library." (Nemes and Kiss, 2021) It performs a number of NLP tasks, including sentiment analysis. Given a document, TextBlob outputs a lexicon-based polarity score that lies between -1 and 1, with -1 indicating a negative sentiment and 1 indicating a positive sentiment. We also considered TextBlob's subjectivity score for context, which lies between 0 and 1 with 1 indicating strong personal opinion. Vader is another Python library which was developed to measure sentiment in tweets, assigning a positive, negative, and neutral score as output to a given text. Each of these outputs will be produced for a given comment as a measure of context and used in an unsupervised learning model for WSD.

It was important to get a distance measure between city-pairs to act as an independent variable in the gravity model. To do this, OpenStreetMap data combined with QGIS processing tools were used. First, geographic coordinates were obtained using Nominatim (*Nominatim* n.d.), an OpenStreetMap search engine which takes search terms and returns coordinates. These labeled coordinates were subsequently converted into geometries in QGIS using WGS84 coordinate reference system. The geometries were then validated using QGIS quick map services to add a map layer to verify locations of the city nodes. Finally, the Distance Matrix tool was used to calculate distances between all pairs of cities.

We decided to transform the time variable to account for cyclical patterns that may occur with season or time of year. City names may appear more frequently during popular vacation periods or some other confounding variable. For this step, a sine function with a period of one year was fitted to the time the comment was created.

Chapter 3

Methods

3.1 Co-occurrence

Co-occurrence is determined on the comment level, meaning that any individual comment that contains more than one city name will be included in the analysis. In order to identify comments with co-occurrences, a set of city names was created from the list of MUA's in Europe with populations greater than 1 million (39 cities in total). This set consisted of city names in the local language, so the English variation was included as well to form a dictionary of search terms. For this, the guidelines for naming convention on Wikipedia were deemed sufficient and city names were scraped from Wikipedia pages of individual cities (*Naming Conventions* n.d.). All text was converted to lower case characters before searching for occurrences throughout this study due to the informal nature of the corpus and inconsistencies with capitalization.

The next step was to construct a co-occurrence matrix with a row for each comment. Every toponym and every token in each comment was searche, and a dummy vector the length of the set of toponyms was created with 1 indicating inclusion and zero indicating no inclusion. In addition, the metadata outlined in Table 2.1 was appended to the vector, and each of the vectors were added as rows in a Pandas dataframe (Pandas, 2022). Recall that the set of toponyms include English and non-English translation of the same entity, meaning that a row would have redundant information if, say, "Athens" and "Athinia" appeared in the same comment. For this reason, the dummy columns referring to the same city were merged, taking the maximum value of the two columns. In total, the number of reddit comments in the 15 year periods containing toponym co-occurrence of the 39 largest European cities amounts to 1,121,437.

The entire corpus of comments with at least one toponym were identified, indexed, and saved locally totalling approximately 9 gigabytes of data. The text body of each comment was then split into a list of tokens. Tokenization consisted of converting text to all lowercase and spitting by empty space. The comments were then searched for additional city names, and the indices of comments with more than two toponyms were recorded. The decision to do this in two stages was one of many pragmatic decisions to due to computational cost, including batch processing of monthly comment dumps during decompression. This process yielded a text file containing a line for each month and the corresponding indices of comments with toponym co-occurrences.

3.2 Exploratory Data Analysis

How are subreddits represented in the co-occurrence data? The frequency distribution in Figure 3.1 was created from the co-occurrence matrix. The distribution of subreddit counts in the corpus is heavily skewed, forming an approximate power law distribution with a small number of subreddits containing the majority of co-occurrences. In total there are 22,001 different subreddits, most of which contribute very little to overall patterns.

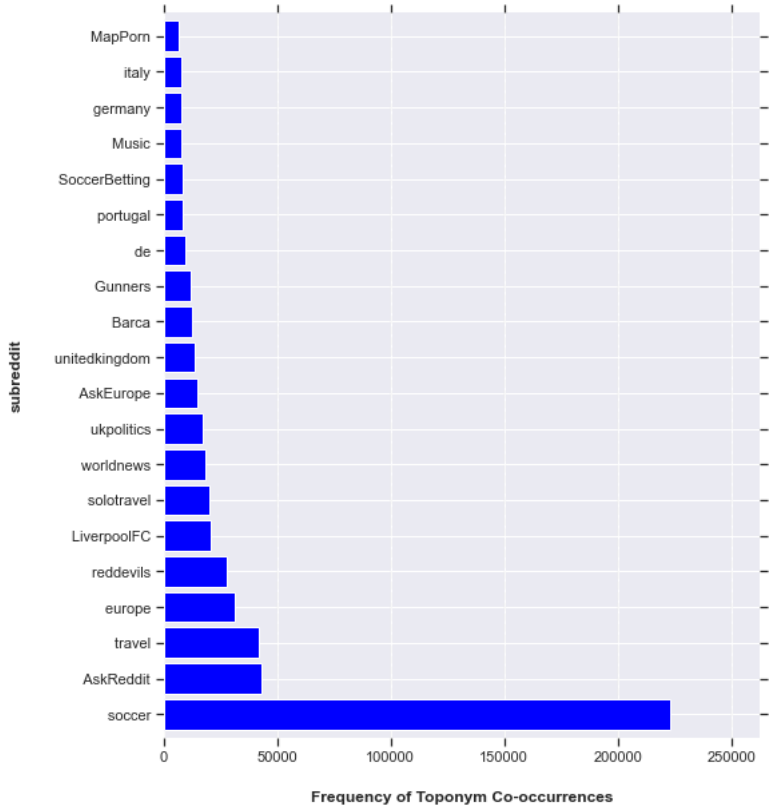


FIGURE 3.1: Top 20 Subreddits with Co-occurrences

3.2.1 Test for Independence

Reddit is host to diverse communities of people with common, and often very niche, interests. Communities have volunteers called moderators who make editorial decisions about which content to post. "The culture of each community is shaped explicitly, by the community rules enforced by moderators, and implicitly, by the upvotes, downvotes, and discussions of its community members" (*Content Policy* n.d.). The hierarchical structure of Reddit with subreddit communities opens up the possibility of between group variability. Figure 3.1 shows that 'r/soccer' is heavily represented in the co-occurrence data, and that 6 out of the top 20 subreddits are related to soccer. To investigate between group variability further we performed a chi-square test for independence. This test can be used to determine if there is significant evidence that subreddit community affects toponym co-occurrence frequency.

In total there are 741 unique toponym pairs and 22001 unique subreddits. Including all of these categories in one test would lead to over 16 million degrees of

freedom and a very sparse two-way table. Because many of the toponym pairs and cities are underrepresented, including all of them would would make it impossible to do a chi-square test for independence because the the expected counts assumption of 5 observations per cell would not be met. For this reason, we decided to do the test on the most common occurring subreddits and city-pairs, as this would give insight into the dominant trends in the data. The 10 most common city pairs and the 15 most common subreddits in the co-occurrence matrix were included in the test. We will use a significance level of 0.05.

	Chi-square test	results
0	Pearson Chi-square (126.0) =	19184.6984
1	p-value =	0.0000
2	Cramer's V =	0.2848

FIGURE 3.2: Results

The results of the test can be seen in Figure 3.2. A two-way table of the expected counts two-way table can be seen in Appendix B.1. The test returned a p-value of 0, meaning there is significant evidence of association between subreddit community and of toponym co-occurrences frequencies. Cramer's V measure's how strongly two categorical variables are associated, and the results indicate a moderate association (*Cramer's V* n.d.).

3.2.2 Time Series

The results of the chi-square test for independence suggest that toponym co-occurrence is strongly dependent on subreddit. For this reason, measuring relatedness with this method could be best suited at the subreddit level. Communities are open forums created for people with common sets of interests. Confirmation bias, or "the tendency to seek, select, and interpret information coherently with one's system of beliefs," (Nickerson, 1998) plays a big role in the communities people choose to join and contribute to. Further, the 'action of this cognitive bias may lead to the emergence of homogeneous and polarized communities - i.e., echo-chambers" (Brugnoli, 2001). Reddit users are particularly susceptible to this recursive pattern because it is reinforced by the recommender algorithm used to suggest content based on past activity. The clusters resulting from these biases and mechanism likely translate to between subreddit variability in toponym co-occurrences.

Figure 3.3 shows the cumulative co-occurrence count for all of Reddit from 2006-2020, with the top ten city-pairs labeled and highlighted.

The monotonic trend in co-occurrences coincides the growth of Reddit since 2005. Figure 3.3 corrects for this and shows the percent of all toponym co-occurrences for each year per subreddit. There was much more variability in the early years, and the distribution of toponym co-occurrences has since converged significantly.

The time series graphs show trends of toponym co-occurrences for all of Reddit, but gives no imformation on the subreddit level. Are these patterns consistent across subreddits? Figure 3.1 shows that community representation in the corpus of comments with co-occurrences is far from uniform. To explore this further distributions of a set of city pairs were compared for different subreddits.

Figure 3.4 shows the relative frequency of 5 city pairs from the top ten subreddits. To create this graph, the most common co-occurrence pair for each of the top ten

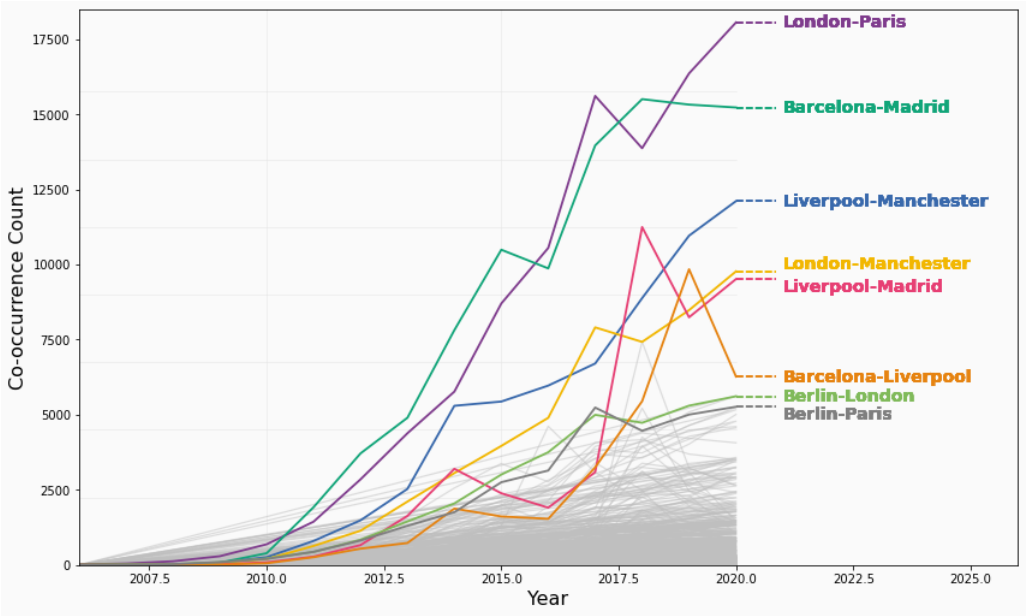


FIGURE 3.3: Frequency of toponym co-occurrences on Reddit for European cities with over 1 million inhabitants from 2006 to 2020

subreddits was identified and included. Five city pairs are represented in this group. The relative frequency of those five city-pairs was then calculated by dividing the number of co-occurrences by the total number of co-occurrences of all five city pairs. Figure 3.5 shows the between subreddit variability demonstrated in the chi-square test for independence.

A closer look at co-occurrences for the ten most represented subreddits in the co-occurrence data set can be seen in Appendix A.1. 'r/soccer' and 'r/AskReddit' do not share one unique pair of cities in their respective top ten. Further, the overall trends of co-occurrence line plots vary significantly across subreddits. It appears that the key to understanding relatedness between cities is to analyze them on the subreddit level. This could lead to a better understanding of domain specific relations, as subreddits corresponds to differenct communities and interests.

3.3 Regression

Regression will consist of the following steps:

- Fit a gravity model to the co-occurrence data
- Identify outliers based on the gravity model
- Fit a linear regression with additional explanatory variables

The final two steps will serve a similar purpose: to control for variability between subreddits in a prediction model. Based on exploratory data analysis, failure to take subreddit into account will lead to unreliable inference due to the ecological fallacy.

3.3.1 Gravity Model

A gravity model will be constructed to determine how well population and distance between cities predict toponym co-occurrence on reddit. While the role of physical

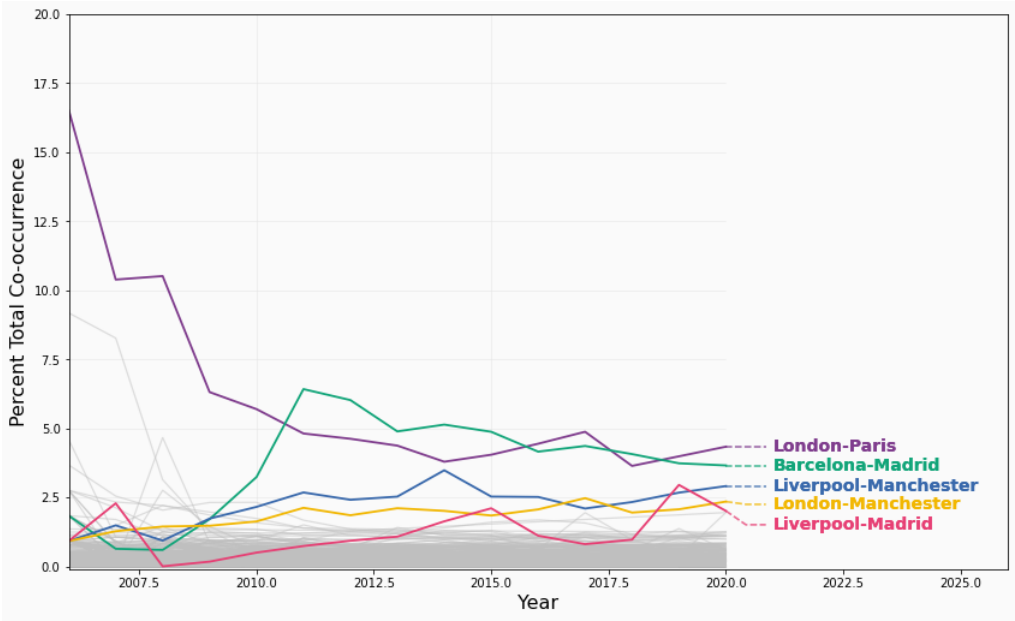


FIGURE 3.4: Relative frequency of toponym co-occurrences on Reddit for European cities with over 1 million inhabitants from 2006 to 2020

distance may be less important when modeling online interactions, it could still play an important role in explaining toponym co-occurrences. Further, the advantage of a linear regression model over non-linear alternatives is interpretability. The gravity model and subsequent linear regression model can provide a measure of the strength of the relationship between the independent variables and toponym co-occurrences, along with a p-value indicating the significance for each independent variable. For more explanation on the gravity model see (Hao Guo and Liu, 2022). Figure 3.6 shows the linear workflow for the gravity model.

The independent variable in the gravity model is the natural logarithm of co-occurrence counts for a unique pair of cities. To prepare the data, dummy variables were created for each unique pair of cities (741 unique combinations for 39 cities). Comments with more than 2 toponym occurrences were then expanded so that there was a single instance of the comment for each unique pair of cities. Counts per unique pair of toponyms were calculated by aggregating by sum. MUA population data for each city was added from the (*ESPON 1.4.3 – Final Report* 2007) data set, and distance calculated in QGIS. A sample of resulting model data can be seen in Figure 3.7. (Note that explanatory variables have also been log transformed).

Residuals represent the difference between the predicted and observed number of co-occurrences and the observed, and can be used to identify outliers. Figure 3.8 shows the residual plot for the gravity model.

In isolation the residual plot does not appear to show any discernible pattern, which indicates that the chosen model is appropriate. The plot of the studentized residuals in figure 3.9 can be used to identify outlier city pairs. If the absolute value of the residual is greater than a threshold of 3 it will be considered a an outlier.

The graph appears to show several negative residuals at the bottom of the figure. The outlier data can be seen in figure 3.10.

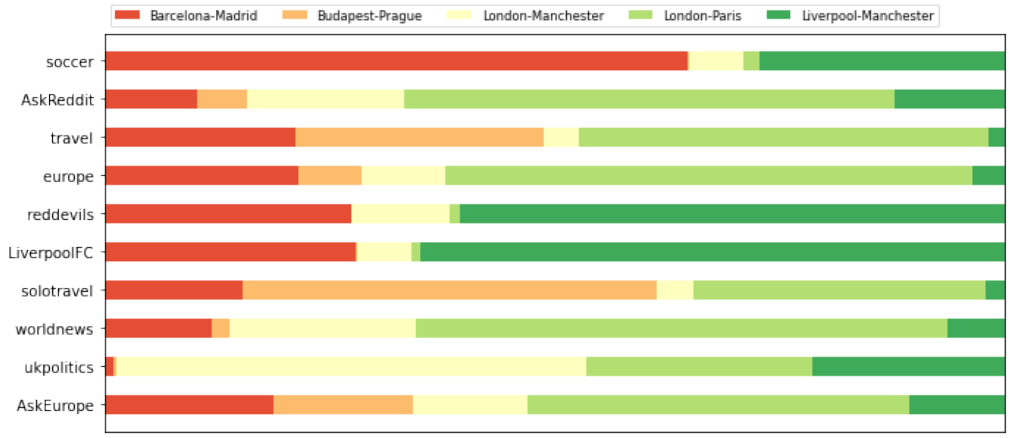


FIGURE 3.5: Proportion of total co-occurrence subset of city pairs for the 10 subreddits with the most toponym co-occurrences

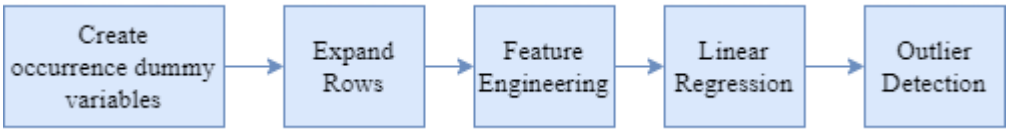


FIGURE 3.6: Gravity Model Workflow

3.3.2 Linear Regression

For linear regression a random sample 50,000 comments from the co-occurrence matrix was taken. The goal is to use as much information as possible, including contextual factors from the text body derived from TextBlob and Vader NLP capabilities. We would like to explore the possibility that certain sentiments may be more common in online interactions depending on the places being discussed. Processing the text body came at a heavy computational cost so a sufficiently large sample of the corpus had to be used instead.

Figure 3.11 shows a correlation matrix for all of the candidate features for the model. One assumption of linear regression is that the predictors are independent, so strong correlations indicate that variables should be removed or combined. It was clear that the five sentiment scores from the two different packages would likely correlate in some way, so that subset of predictors would need to be reduced. Further, 'num-tops' appears to be moderately correlated with other predictors while not being a particularly interesting or informative predictor.

An initial model was created with all comment metadata, sentiment features, and dummy variables for subreddit. The model was iteratively tuned using visualizations and descriptive statistics. The results failed to explain variability in toponym co-occurrence better than the gravity model. Further, scatterplots of the candidate features were created to see how the individual variables are associated with co-occurrence counts, and none of them appeared to fit a linear model. The scatterplots can be seen in Apendix A1.

It was clear from exploratory data analysis that toponym co-occurrence was influenced by subreddit. Without building this into the model the results and any subsequent inference would be biased. Soccer teams referred to by city names in

tops	ln_pop1	ln_pop2	ln_distance	ln_counts
['Amsterdam', 'Athens']	6.958448	8.111028	14.588456	7.302496
['Amsterdam', 'Barcelona']	6.958448	8.204945	14.029741	8.803124
['Amsterdam', 'Berlin']	6.958448	8.236421	13.266508	9.678530
['Amsterdam', 'Birmingham']	6.958448	7.767687	13.043953	6.336826
['Amsterdam', 'Brussels']	6.958448	7.311886	12.066258	9.178230

FIGURE 3.7: Gravity Model Data

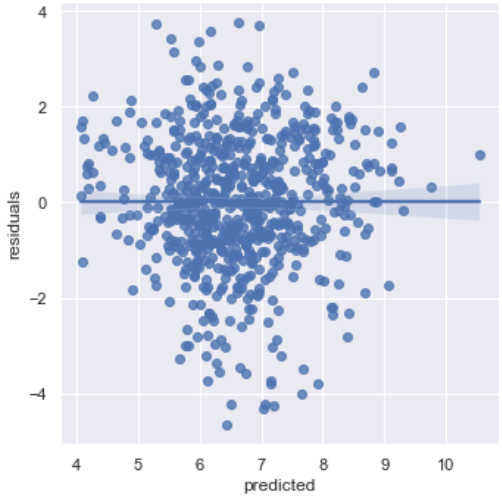


FIGURE 3.8: Gravity Model Residaul Plot

particular appear to be influential on the global data, as in addition to the soccer subreddit, there are four other subreddits in the top 20 in terms of co-occurrences that are dedicated to soccer teams. We created a dummy variable for the soccer subdreddit. We also compiled a list of related subreddits from (*Related Subreddits* n.d.) and created a dummy variable indicating if a comment was posted in any of those subreddits. In both cased the dummy variable was included as a proportion of all occurrence counts for a given pair of cities. It was important to normalize this way because the soccer counts would likely positively correlate with total counts due to the growth of Reddit and the soccer subreddit. The model which included a dummy variable for only the soccer subreddit did the best at explaining variability in counts, improving on the gravity model with an adjusted R^2 value of 0.395. Results can be seen in Appendix B.3.

3.4 Disambiguation

Toponym disambiguation will be addressed with unsupervised learning techniques as mentioned above. We will attempt word sense disambiguation with the city of Paris. The goal is to determine if a place name with at least one other dominant meaning can be separated into distinct word sense clusters. As mentioned in the literature review, we will use context, in the form of n-grams around the token 'paris'

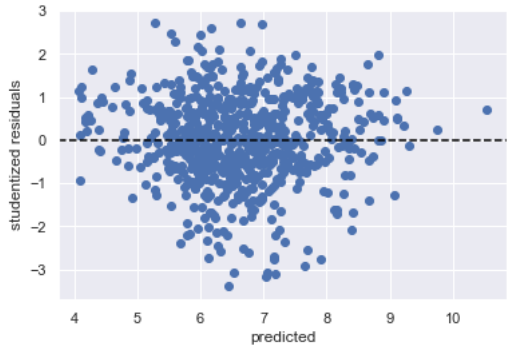


FIGURE 3.9: Gravity Model Studentized Residuals

	const	ln_pop1	ln_pop2	ln_distance	student_resid	ln_count
AthensKatowice	1.0	8.111028	7.731492	14.158947	-3.099772	2.944439
DüsseldorfKatowice	1.0	6.923629	7.731492	13.676153	-3.390482	1.791759
KatowiceNaples	1.0	7.731492	7.744137	13.916221	-3.150082	2.708050
KatowiceStuttgart	1.0	7.731492	7.458763	13.499297	-3.074502	2.833213
KatowiceTurin	1.0	7.731492	7.177019	13.838849	-3.073001	2.302585

FIGURE 3.10: Gravity Model Outliers

which will be used to sort the candidate into one of two clusters, along with the other available reddit data. Once again, all comments will be converted to lowercase so that capitalization does not affect the search. Figure 3.12 shows the disambiguation workflow.

During exploratory data analysis we checked the head and tail of the data and found that in the early history of Reddit the string 'paris hilton' appeared more often than the substring 'paris' on its own. The former refers to a socialite who appeared on television around that time. Place names which doubled as football teams names were considered for this task, but 'paris' offers a unique model evaluation opportunity that would not otherwise be possible. As this is an unsupervised learning technique, even if there are two distinct clusters of comments it will be impossible to determine if the toponyms were correctly grouped without labels. Therefore, comments containing the string 'paris hilton' will be given a pseudo label so that the performance of the model can be evaluated with a confusion matrix and an accuracy score.

Sentiment scores and other metadata used in the gravity model will be applied here as well. In addition to the preprocessing steps outlined in the Chapter 2, French stop words will be removed. N-grams of varying lengths with 'paris' at the center will be created. The n-grams will be converted into vectors using the TF-IDF (term frequency-inverse document frequency) method, which determines the relevance of a given word to a document and "is one of the most commonly used term weighting schemes in today's information retrieval systems" (Aizawa, 2003). This will emphasize discriminatory words in the comment relative to the whole corpus. The idea is that toponyms will tend to be surrounded by uniquely identifying words which will help the algorithm sort them into the correct cluster.

K-means clustering will be used to group comments in one of two clusters. This method was chosen because it is a simple and effective method for unsupervised learning which works well with large datasets. However, clusters are created based

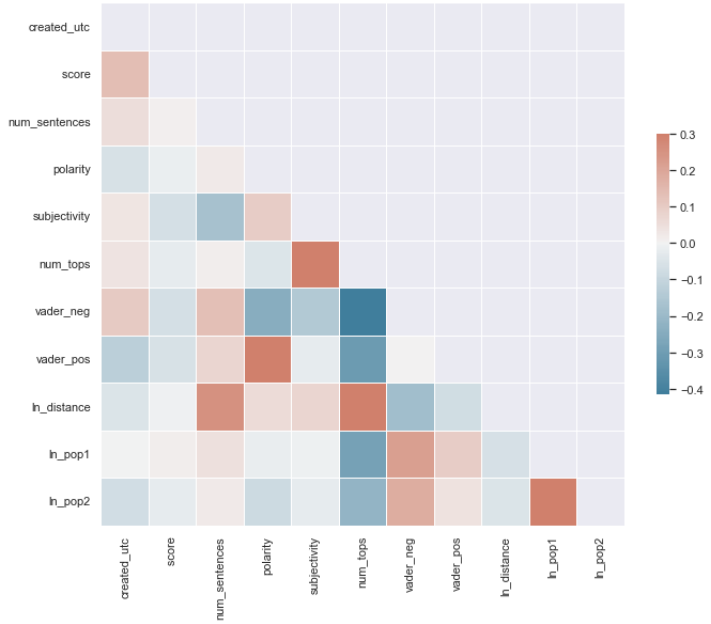


FIGURE 3.11: Correlation Matrix

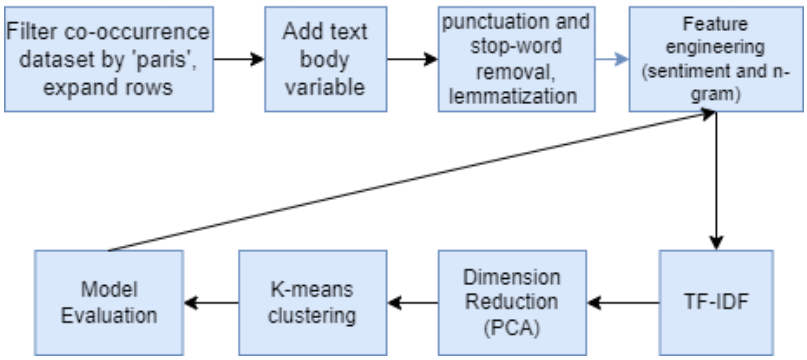


FIGURE 3.12: Disambiguation Workflow

on Euclidean distance and nearest neighbors, and K-means clustering becomes less effective as the number of dimensions in the dataset increases as Euclidean distance is not a reliable metric in high dimensional space (see (Zimek, Schubert, and Kriegel, 2012)). This is problematic for TF-IDF because it produces a large and very sparse matrix with a column for every word in the corpus and a row for every comment. For this reason, we will reduce the dimensions in the TF-IDF vector using Pricipal Component Analysis (PCA) which will transform the data into a lower dimensional space. This will be done iteratively, as the more information contained in the principal components, the more dimensions in the output of the algorithm. This setting will be adjusted to tune the model.

There is an issue with imbalanced data which will need to be addressed. If the k-means algorithm determines that every comment containing 'paris' is referring to the city then the model will have an accuracy greater than 99%. There are 570 instances of 'paris hilton' out of 215,616 comments containing 'paris.' To treat this imbalance, the majority class will be randomly undersampled so that there is a random sample of 570 comments with 'paris' and without the 'paris hilton.'

Chapter 4

Results

4.1 Regression

The gravity model shows that toponym co-occurrences are indeed affected by the distance-decay effect. Roughly 34% of the variabiliy in toponym co-occurrences can be explained by population and distance explanatory variables. Additionally, the p-values for the log transformed explanatory variables were all approximately zero, indicating that they were all statistically significant and helped to explain co-occurrence variation. The coefficient for distance was negative, meaning that as distance between cities increased the number of co-occurrences decrease as expected. Both population variable have a positive coefficient, meaning that larger cities tend to appear more often in comments with co-occurrences.

Outliers were identified using studentized residuals. All of them were negative outliers, indicating that predicted number of co-occurrences was less than the observed for those city pairs. Additionally, they all included the city of Katowice.

There were many features that were considered for inclusion in the linear regression model, including temporal and sentiment variables. The only one that improved the model and while also agreeing with the between subreddit variability finding in exploratory data analysis was the dummy variable for the soccer subreddit. This helped to explain roughly 40% of variability in co-occurrences when added to the gravity model. All explanatory variable coefficients in this model had p-values approximately zero, and the coefficients were all positive except for the distance variable. The results are shown in full in appendix B.

4.2 Disambiguation

The results of the disambiguation model can be seen in Table 4.1. True negatives and true positives are the number of comments that were correctly labeled. The accuracy of the model is approximately 51%. The model was tuned iteratively by changing the size of the n-gram and the PCA threshold. Increasing the number of words around 'paris' helped to increase the amount of context captured in the model. This step did not change the results significantly, as the accuracy was resistant to n-gram tuning and the accuracy remained approximately 50% throughout the iterations.

TABLE 4.1: Confusion Matrix for 'paris' Disambiguation

True Negative	False Positive	False Negative	True Positive
296	277	284	282

Chapter 5

Conclusion

The results of the gravity model showed that distance and population do indeed influence the frequency of co-occurrences for European cities. Exploratory data analysis showed that patterns are heavily influenced by subreddit. The inclusion of the soccer dummy variable helped to explain more of the variability. This was no surprise, as the soccer subreddit was the most represented subreddit in the corpus by a substantial margin. In the end the correlation matrix in Figure 3.10 was not used for the multicollinearity problem because the highly correlated variables were removed due to their non-linearity.

Future work may consider continuing to reduce the noise in the data by adding more subreddit dummy variables. Subreddits can be clustered together to collapse the number of inputs based on some criteria like similarity scores. While soccer improved the model, it's reasonable to assume that it was not the only influential subreddit within the corpus. The co-occurrence time series graphs in Appendix A.2 show that between subreddit variability is common throughout the corpus.

Another option to account for heterogeneity between subreddits is a Linear Mixed Model (LMM). The hierarchical organization of reddit shown in Figure 2.1 makes this method suitable. LMMs contain both fixed and random effects, whereas the regression models in this study consist of only fixed effects. They can be applied when the assumption of independence among predictors is not met. In the case of this study, it appears that knowing the subreddit a comment was posted in makes it easier to predict the class of co-occurrence, so independence is not satisfied. The solution would be to assign a random effect to the intercept or the slope of the regression line based on subreddit clustering. This would have a similar effect to introducing subreddit dummy variables. It was not implemented in this study because the LMM python package did not provide adjusted R^2 values for comparison with our other models. Including a soccer dummy improved the adjusted R^2 value considerably, and a linear mixed model would allow for this flexibility with all subreddits.

Outliers identified with the studentized residual all contained the city of Katowice. Similarly, all pairs that were excluded from the model entirely contained the city Frankfurt Am Main. City pairs with no occurrences had to be excluded because the gravity model required a log transformation of all variables, which cannot be done with a value of 0. The coincidental nature of these two outliers suggests that there may be some underlying reason this occurred. In the case of 'Frankfurt Am Main', both the (*ESPON 1.4.3 – Final Report* 2007) dataset and Wikipedia agreed on this spelling, however 'Frankfurt' is likely the more common way of referring to the city. In this case, it would be worth re-indexing the Reddit corpus for 'Frankfurt' instead and include it in the study.

The scatterplots in Appendix A.1 showed that additional created features were not appropriate for a linear model. As mentioned earlier, the linear regression model

was chosen for interpretability, however it would be interesting to build on the research question of this paper to determine if co-occurrence frequency can be predicted with non-linear machine learning techniques. Natural language processing is a notoriously complex task, and the sentiment analysis did not help to produce a robust model in this study for linear regression or disambiguation.

The disambiguation model did not prove to be effective in discriminating word sense in the case of 'paris'. There was always a danger of overfitting in this task, as the classifier was created with a specific application of WSD in mind. Even if the accuracy of the model was high, there is no guarantee that it could be generalized to other cities with ambiguous names.

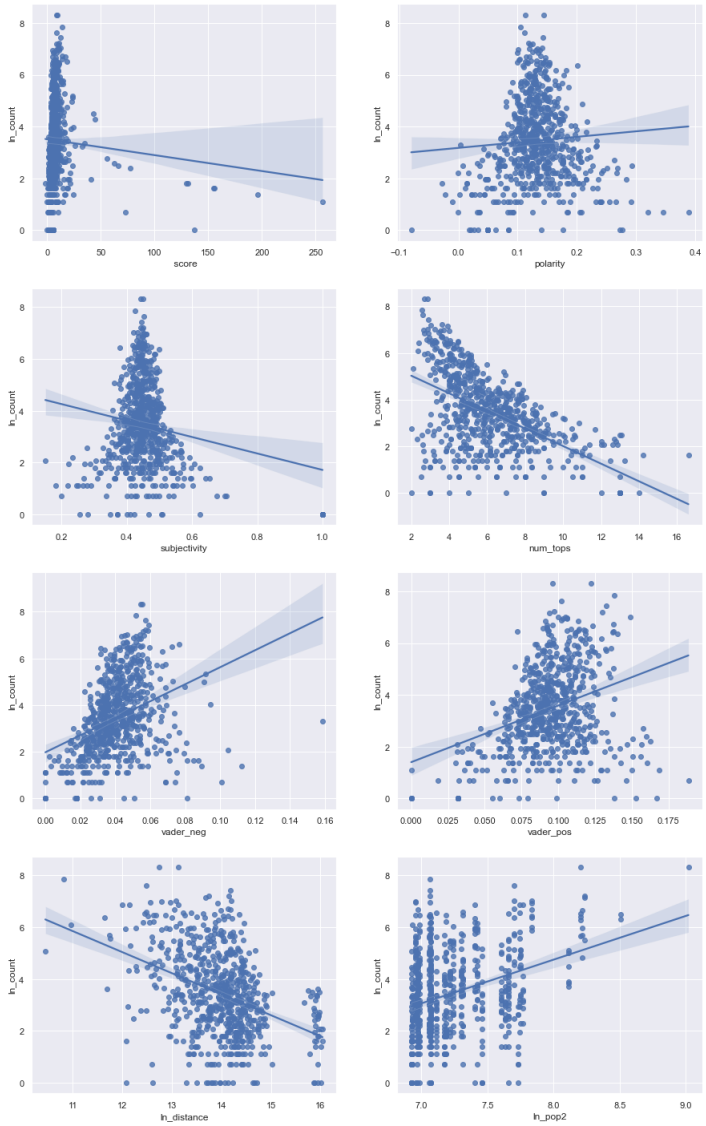
The model was likely ineffective because although we took dimensionality reduction measures with PCA, there was still far too many of them for the K-means algorithm to be effective. Euclidean distance can be useful, however "the distinction in distance decreases fastest in the first 20 dimensions, quickly reaching a point where the difference in distance between a query point and the nearest and farthest data points drops below a factor of four" (Beyer et al., 1999). PCA only managed to reduce the number of dimensions to several hundred from approximately 20 thousand. The trade-off between the effectiveness of PCA to capture the meaning of the text and the reduction of dimensions was not possible.

For future studies, alternative methods for codifying context can be considered. K-means clustering, TF-IDF, and PCA were incompatible for reasons already stated. Part of Speach (POS) tagging is another possiblility for capturing the context around a word, and would result in fewer dimensions. Additionally, alternate methods to vectorize the n-grams that result in smaller vectors should be explored.

Appendix A

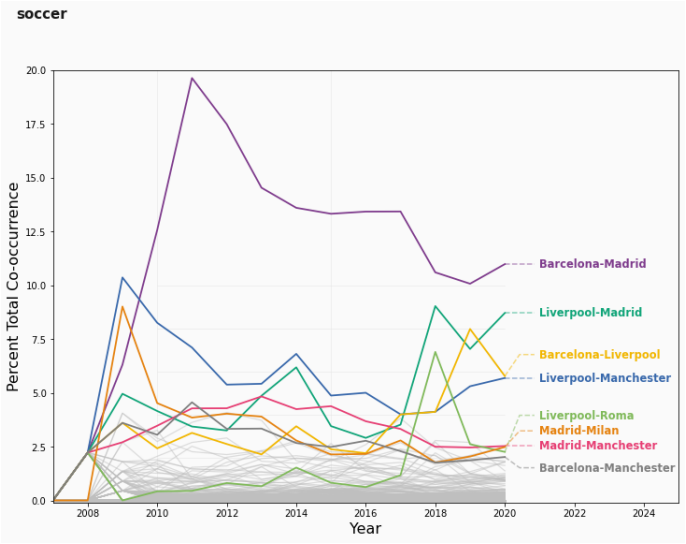
Graphs and Figures

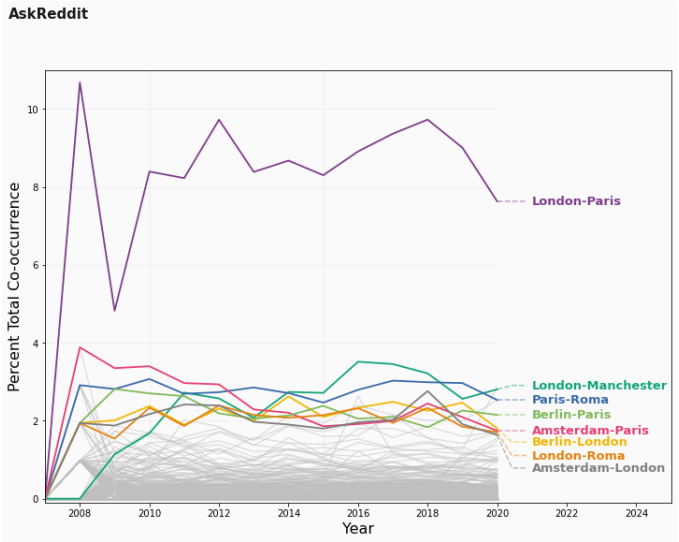
A.1 Scatterplots of Independent Variables vs Log Transformed Counts

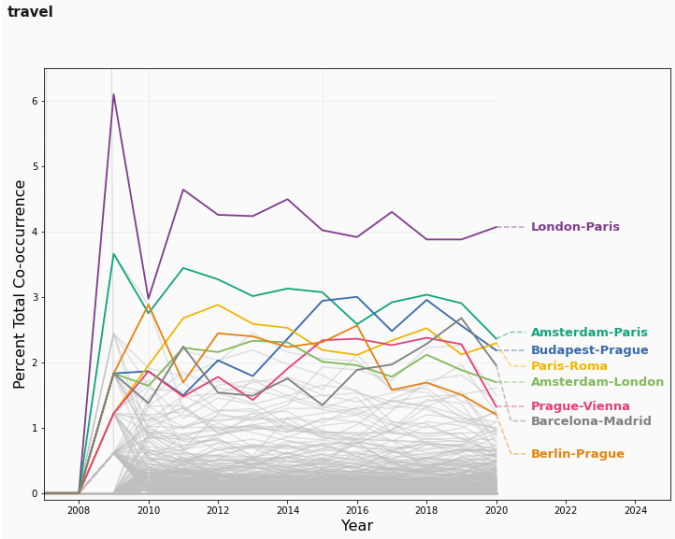


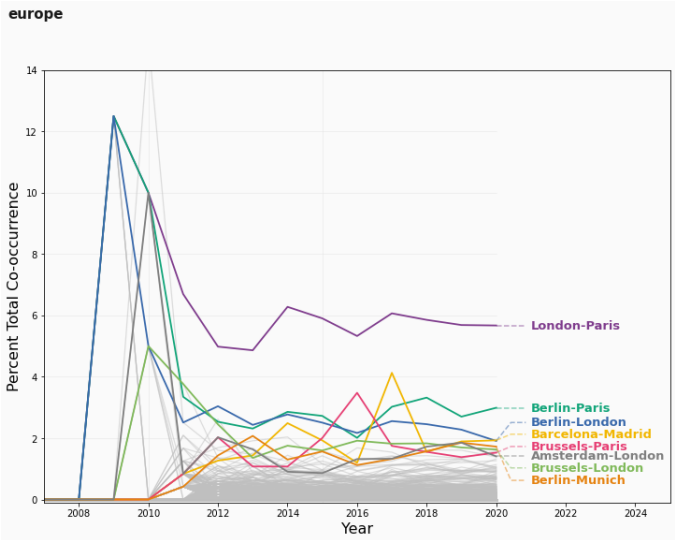
A.2 Co-occurrence Time Series by Subreddit

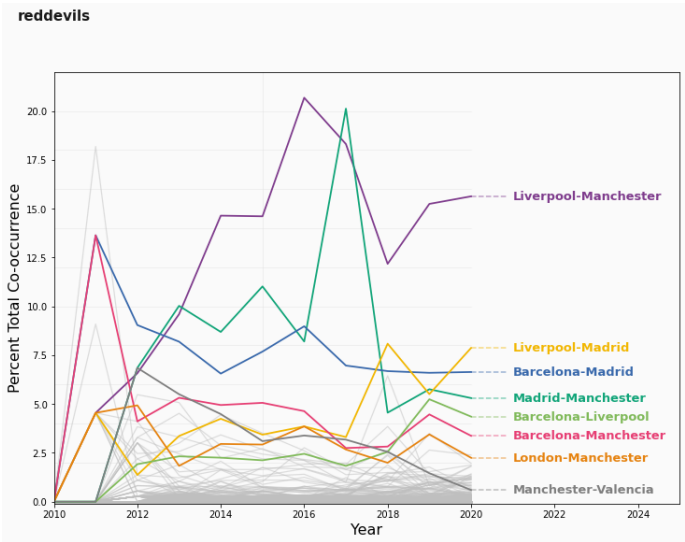
Proportion of toponym co-occurrences on Reddit for European cities with over 1 million inhabitants from 2006 to 2020 by subreddit

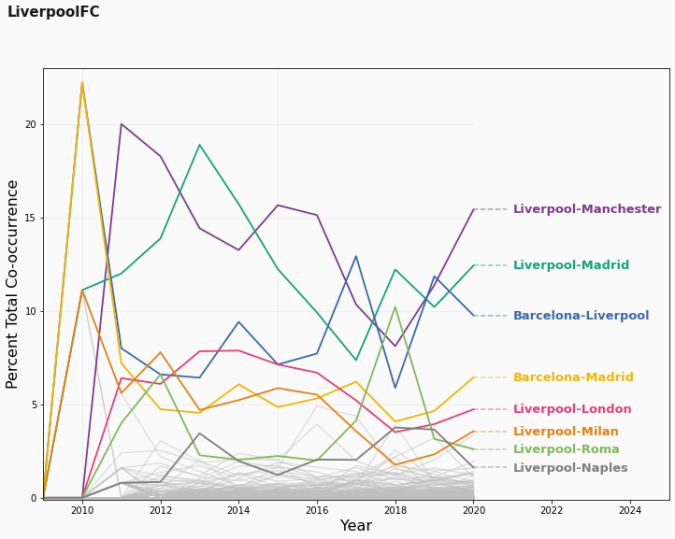


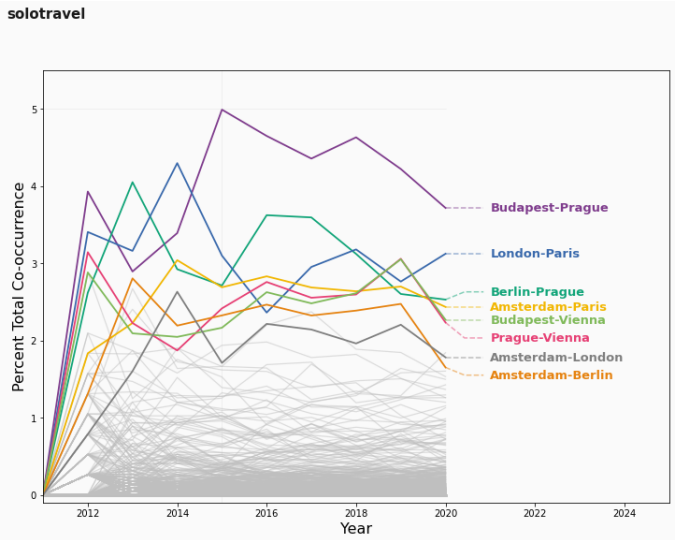


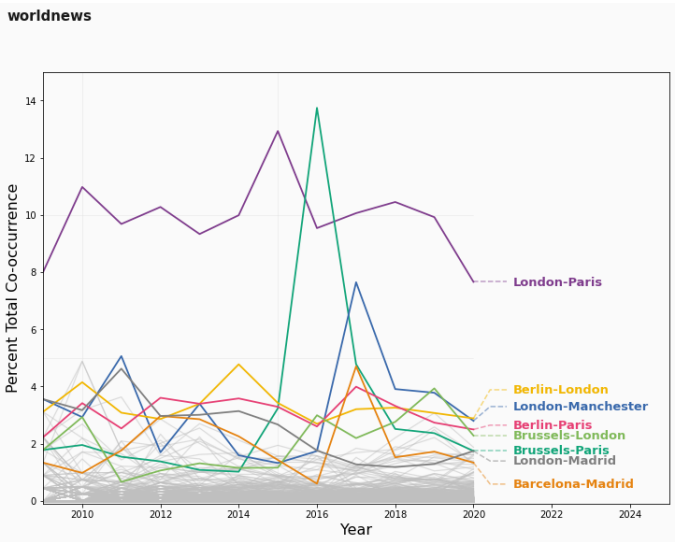

travel
Percent Total Co-occurrence
Year
London-Paris
Amsterdam-Paris
Budapest-Prague
Paris-Roma
Amsterdam-London
Prague-Vienna
Barcelona-Madrid
Berlin-Prague


europe
Percent Total Co-occurrence
Year
London-Paris
Berlin-Paris
Berlin-London
Barcelona-Madrid
Brussels-Paris
Amsterdam-London
Brussels-London
Berlin-Munich


reddevils
Percent Total Co-occurrence
Year
Liverpool-Manchester
Liverpool-Madrid
Barcelona-Madrid
Madrid-Manchester
Barcelona-Liverpool
Barcelona-Manchester
London-Manchester
Manchester-Valencia


LiverpoolFC
Percent Total Co-occurrence
Year
Liverpool-Manchester
Liverpool-Madrid
Barcelona-Liverpool
Barcelona-Madrid
Liverpool-London
Liverpool-Milan
Liverpool-Roma
Liverpool-Naples


solotravel
Percent Total Co-occurrence
Year
Budapest-Prague
London-Paris
Berlin-Prague
Amsterdam-Paris
Budapest-Vienna
Prague-Vienna
Amsterdam-London
Amsterdam-Berlin


worldnews
Percent Total Co-occurrence
Year
London-Paris
Berlin-London
London-Manchester
Berlin-Paris
Brussels-London
Brussels-Paris
London-Madrid
Barcelona-Madrid

Appendix B

Results

B.1 Chi-Square Expected Counts Table

sub pair	askeurope	askreddit	barca	de	europe	gunners	liverpoolfc	portugal	reddevils	soccer	solotravel	travel	ukpolitics	unitedkingdom	worldnews
barcelona_liverpool	37.039831	130.112075	65.313855	5.328236	70.556152	57.923077	128.737046	3.351632	171.019174	1292.097124	34.289774	100.892719	60.071559	52.938598	49.329149
barcelona_madrid	105.626645	371.041163	186.255802	15.194552	201.205280	165.179487	367.119988	9.557864	487.696112	3684.678917	97.784296	287.716199	171.306323	150.965229	140.672145
liverpool_madrid	52.468995	184.311040	92.520733	7.547744	99.946740	82.051282	182.363235	4.747774	242.258236	1830.327931	48.573385	142.920186	85.094727	74.990489	69.877501
liverpool_manchester	60.109792	211.151335	105.994065	8.646884	114.501484	94.000000	208.919881	5.439169	277.537092	2096.869436	55.646884	163.732938	97.486647	85.910979	80.053412
liverpool_roma	24.381686	85.647036	42.993228	3.507342	46.444001	38.128205	84.741916	2.206231	112.574374	850.530511	22.571445	66.413224	39.542456	34.847143	32.471201
london_manchester	30.464810	107.015598	53.719851	4.382409	58.031576	47.641026	105.884653	2.756677	140.661188	1062.734155	28.202922	82.983033	49.408126	43.541353	40.572624
london_paris	46.123526	162.020924	81.331507	6.634939	87.859431	72.128205	160.308681	4.173591	212.960131	1608.972647	42.699041	125.635776	74.803584	65.921327	61.426691
madrid_manchester	27.906947	98.030434	49.209465	4.014456	53.159172	43.641026	96.994446	2.525223	128.851099	973.505668	25.834969	76.015674	45.259758	39.885566	37.166096
milan_roma	23.840600	83.746329	42.039108	3.429506	45.413300	37.282051	82.861295	2.157270	110.076086	831.655254	22.070532	64.939359	38.664917	34.073804	31.750590
naples_roma	23.037168	80.924066	40.622385	3.313931	43.882865	36.025641	80.068858	2.084570	106.366507	803.628357	21.326752	62.750894	37.361904	32.925512	30.680590

B.2 Gravity Model Results Summary

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                ln_count   R-squared:                       0.340
Model:                            OLS   Adj. R-squared:                  0.337
Method:                 Least Squares   F-statistic:                     119.9
Date:                Sun, 10 Jul 2022   Prob (F-statistic):           1.16e-62
Time:                        18:27:09   Log-Likelihood:                -1222.7
No. Observations:                 703   AIC:                             2453.
Df Residuals:                     699   BIC:                             2472.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.2772      1.424      1.599      0.110      -0.518       5.073
ln_distance   -0.7387      0.066    -11.130      0.000      -0.869      -0.608
ln_pop1        0.9419      0.094     10.017      0.000       0.757       1.127
ln_pop2        1.0005      0.095     10.573      0.000       0.815       1.186
==============================================================================
Omnibus:                       17.543   Durbin-Watson:                   1.236
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               19.325
Skew:                          -0.331   Prob(JB):                     6.36e-05
Kurtosis:                       3.470   Cond. No.                         482.
==============================================================================
```

B.3 Gravity Model w/ Soccer Dummy Results Summary

OLS Regression Results

Dep. Variable:	ln_count	R-squared:	0.398
Model:	OLS	Adj. R-squared:	0.395
Method:	Least Squares	F-statistic:	115.4
Date:	Sun, 10 Jul 2022	Prob (F-statistic):	1.60e-75
Time:	21:36:46	Log-Likelihood:	-1190.2
No. Observations:	703	AIC:	2390.
Df Residuals:	698	BIC:	2413.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.7640	1.362	1.295	0.196	-0.910	4.438
ln_pop1	0.9065	0.090	10.078	0.000	0.730	1.083
ln_pop2	1.0546	0.091	11.632	0.000	0.877	1.233
ln_distance	-0.7355	0.063	-11.598	0.000	-0.860	-0.611
r_soccer_norm	2.9263	0.356	8.219	0.000	2.227	3.625

Omnibus:	16.351	Durbin-Watson:	1.329
Prob(Omnibus):	0.000	Jarque-Bera (JB):	16.936
Skew:	-0.378	Prob(JB):	0.000210
Kurtosis:	3.086	Cond. No.	483.

Bibliography

Aizawa, Akiko (2003). "An information-theoretic perspective of tf–idf measures". In: *Information Processing Management* 39.1, pp. 45–65. ISSN: 0306-4573. DOI: https://doi.org/10.1016/S0306-4573(02)00021-3. URL: https://www.sciencedirect.com/science/article/pii/S0306457302000213.

Beyer, Kevin et al. (1999). "When Is "Nearest Neighbor" Meaningful?" In: *Database Theory — ICDT'99*. Ed. by Catriel Beeri and Peter Buneman. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 217–235. ISBN: 978-3-540-49257-3.

Brugnoli, Emanuele (2001). "Recursive patterns in online echo chambers". In: *Scientific Reports* 9. URL: https://doi.org/10.1038/s41598-019-56191-7.

Content Policy (n.d.). https://www.redditinc.com/policies/content-policy. Accessed: 2022-05.

Cramer's V (n.d.). https://www.ibm.com/docs/en/cognos-analytics/11.1.0?topic=terms-cramrs-v. Accessed: 2022-05.

ESPON 1.4.3 – Final Report (2007). Tech. rep. Institut de Gestion de l'Environnement et d'Aménagement du Territoire, Université.

Hao Guo Weiyu Zhang, Haode Du Chaogui Kang and Yu Liu (2022). "Understanding China's urban system evolution from web search index data". In: *EPJ Data Science* 11.

Ireson, Neil and Fabio Ciravegna (2010). "Toponym Resolution in Social Media". In: pp. 370–385.

Lenormand Maxime Bassolas, Aleix Ramasco José J. (2016). "Systematic comparison of trip distribution laws and models". In: *Journal of Transport Geography*.

Meijers, Evert and Antoine Peris (2019). "Using toponym co-occurrences to measure relationships between places: review, application and evaluation". In: *International Journal of Urban Sciences* 23.2, pp. 246–268. DOI: 10.1080/12265934.2018.1497526. eprint: https://doi.org/10.1080/12265934.2018.1497526. URL: https://doi.org/10.1080/12265934.2018.1497526.

Naming Conventions (n.d.). https://en.wikipedia.org/wiki/Wikipedia:Naming_conventions. Accessed: 2022-05.

Nemes, László and Attila Kiss (2021). "Prediction of stock values changes using sentiment analysis of stock news headlines". In: *Journal of Information and Telecommunication* 5.3, pp. 375–394. DOI: 10.1080/24751839.2021.1874252. eprint: https://doi.org/10.1080/24751839.2021.1874252. URL: https://doi.org/10.1080/24751839.2021.1874252.

Nickerson, R. S (1998). "Confirmation bias: A ubiquitous phenomenon in many guises". In: *Review of General Psychology 2*, pp. 175–220.

Nominatim (n.d.). https://nominatim.openstreetmap.org/ui/search.html. Accessed: 2022-05.

Pushshift (n.d.). https://files.pushshift.io/reddit/comments/. Accessed: 2022-05.

Related Subreddits (n.d.). https://www.reddit.com//r/soccer/wiki/relatedsubreddits?st=J30SQFK0&sh=5b62ae41. Accessed: 2022-05.

Sankar, K.P. Sruthi, P.C. Reghu Raj, and V. Jayan (2016). "Unsupervised Approach to Word Sense Disambiguation in Malayalam". In: *Procedia Technology* 24. International Conference on Emerging Trends in Engineering, Science and Technology (ICETEST - 2015), pp. 1507–1513. ISSN: 2212-0173. DOI: https://doi.org/10.1016/j.protcy.2016.05.106. URL: https://www.sciencedirect.com/science/article/pii/S2212017316301955.

Sun YC Tang DD, Zeng Q Greenes R. (2002). "Identification of Special Patterns of Numerical Typographic Errors Increases the Likelihood of Finding a Misplaced Patient File". In: *J Am Med Inform Assoc.*

Zimek, Arthur, Erich Schubert, and Hans-Peter Kriegel (2012). "A survey on unsupervised outlier detection in high-dimensional numerical data". In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* 5.5, pp. 363–387. DOI: https://doi.org/10.1002/sam.11161. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sam.11161. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sam.11161.