LAB 2_STAT 215A

October 4, 2024

1 Introduction

Dialectology, the study of regional linguistic variation, has long fascinated researchers as it explores the geographical, social, and cultural factors that shape language differences. As dialectology evolved, computational tools enabled researchers to investigate these differences quantitatively through dialectometry. Dialectometry allows linguists to analyze large datasets, examining how language varies across different regions and uncovering patterns that traditional approaches might miss. The works of Nerbonne and Kretzschmar (2003, 2006) emphasize the importance of computational methods for efficiently managing and analyzing complex linguistic data, allowing for deeper exploration of how geography influences language use.

In this report, I aim to investigate the relationship between specific linguistic survey questions and their geographical distribution. By analyzing this data, I hope to uncover distinct regional patterns that correlate with different language features. Specifically, I will explore whether responses to one question can help predict responses to another, and whether distinct geographical groups emerge based on these linguistic preferences.

The data used for this analysis contains categorical responses to various language-related questions from different regions. My primary goal is to analyze two or more survey questions, exploring the geographical patterns in the responses and the relationships between the questions. Using interactive visualizations like maps and contingency tables, I will examine whether specific responses cluster geographically and whether they are interdependent.

The rest of the report is structured as follows:

Data Selection and Cleaning: I will begin by selecting relevant survey questions and cleaning the dataset to ensure valid responses for the analysis. Geographical Visualization: I will visualize the geographical distribution of responses using interactive maps to identify any regional linguistic trends. Statistical Relationship Analysis: I will examine the relationship between the selected questions using contingency tables and statistical tests to determine whether one response can predict the other. Exploration of Additional Questions: To deepen the analysis, I will expand the scope to include more than two survey questions, examining their geographical relationships and interdependencies. Conclusion: I will summarize the key findings, including any distinct

geographical groups that emerge and the strength of relationships between the survey responses. This analysis will provide insights into how geography influences language variation and whether certain linguistic preferences are linked, contributing to the broader understanding of dialectometry.

2 The Data

In this analysis, I will be using data from the Dialect Survey conducted by Bert Vaux. The survey collected responses from over 47,000 respondents across the United States, focusing on lexical differences (i.e., word choices) rather than phonetic differences (i.e., pronunciation). The survey data captures how different terms are used across geographical regions for various everyday concepts, providing a rich dataset for analyzing regional linguistic variation.

The dataset comprises three files:

lingData.txt:

This file contains the responses from 47,471 respondents to survey questions numbered from Q50 to Q121, which focus on lexical differences. The data includes the following variables: ID: A unique identifier for each respondent. CITY, STATE, ZIP: Self-reported location information of the respondents. lat, long: Latitude and longitude coordinates for the center of the respondent's ZIP code, added to the dataset by geocoding the ZIP codes. Q50 - Q121: Categorical responses to the survey questions. A value of 0 indicates no response, while other numbers correspond to specific choices from the survey (e.g., a value of 1 represents response (a)). The responses capture how lexical choices vary by region, such as what word people use for "carbonated beverage" (soda, pop, coke, etc.). lingLocation.txt:

This file aggregates the binary responses from lingData.txt into one-degree latitude by one-degree longitude squares. Within each grid, the binary responses for each question are summed across individuals, representing the total number of people who selected a particular response within that geographical bin. For example, if two respondents in the same latitude/longitude bin selected different answers for a question, the binary response vectors would be summed. question data.Rdata:

This file contains the actual questions and their corresponding answer options from the survey. The questions cover various lexical distinctions, such as terms for "remote control" or "sandwich," which can vary regionally. Overview of the Data The lingData.txt dataset provides individual-level data on lexical preferences, while the lingLocation.txt dataset aggregates these preferences into geographical bins. The data covers a wide geographical area across the U.S., making it ideal for studying how lexical choices are influenced by regional factors. The survey questions cover a range of everyday topics where lexical variation is likely to be present (e.g., terms for soft drinks, sandwiches, or casual footwear).

The binary and categorical responses in the datasets allow for various types of analysis:

Geographical analysis: We can map the distribution of responses across different U.S. regions. Lexical relationships: By comparing responses to different questions, we can explore whether certain lexical choices correlate with others. Relevance to the Problem of Interest The data is highly relevant to the problem of understanding how geography influences language variation. By analyzing the geographical distribution of lexical choices, we can investigate whether certain regions have distinct linguistic identities or whether lexical preferences form geographic clusters. Furthermore, the data allows us to explore relationships between different lexical choices, revealing potential linguistic patterns across regions.

This dataset will enable us to investigate questions such as:

Do certain regions of the U.S. have a clear preference for specific lexical choices? Can the response to one question (e.g., the term for "carbonated beverage") help predict responses to another question (e.g., the term for "casual footwear")? Do different regions exhibit distinct lexical profiles based on multiple survey questions? By linking this data to geographical information (latitude and longitude), we can create visual maps and perform statistical analyses that will deepen our understanding of how geography shapes linguistic variation.

2.1 Data Cleaning

- This dataset isn't as bad as the TBI data, but there are still some issues. You should discuss them here and describe your strategies for dealing with them.
- Remember to record your preprocessing steps and to be transparent!

```
Requirement already satisfied: pandas in
/opt/miniconda3/envs/215a/lib/python3.12/site-packages (2.2.2)
Requirement already satisfied: pyreadr in
/opt/miniconda3/envs/215a/lib/python3.12/site-packages (0.5.2)
Requirement already satisfied: numpy>=1.26.0 in
/opt/miniconda3/envs/215a/lib/python3.12/site-packages (from pandas) (1.26.4)
Requirement already satisfied: python-dateutil>=2.8.2 in
/opt/miniconda3/envs/215a/lib/python3.12/site-packages (from pandas)
(2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in
/opt/miniconda3/envs/215a/lib/python3.12/site-packages (from pandas) (2024.1)
Requirement already satisfied: tzdata>=2022.7 in
/opt/miniconda3/envs/215a/lib/python3.12/site-packages (from pandas) (2023.3)
Requirement already satisfied: six>=1.5 in
/opt/miniconda3/envs/215a/lib/python3.12/site-packages (from python-
dateutil>=2.8.2->pandas) (1.16.0)
```

[4]:

	ID	CITY	STATE	ZIP	Q050	Q051	Q052	Q053	Q054	Q055	…	Q110	\
0	1	Boise	ID	83704	4	1	3	2	3	1	…	8	
1	2	Pittsfield	MA	1201	4	2	3	2	2	2	…	8	
2	3	Burlington	VT	5401	4	1	2	2	2	2	…	4	
3	4	Easton	PA	18042	7	1	1	2	2	2	…	8	
4	5	Bedford	MA	1730	8	2	3	1	2	2	…	8	

	Q111	Q115	Q117	Q118	Q119	Q120	Q121	lat	long
0	2	1	6	7	3	2	3	43.631230	-116.287161
1	2	1	1	7	1	1	3	42.453840	-73.254003
2	2	1	4	7	1	2	1	44.484038	-73.221265
3	1	1	3	7	1	2	1	40.681798	-75.220820
4	3	1	5	8	1	1	1	42.496679	-71.275046

[5 rows x 73 columns]

[6]:
```
   Number of people in cell  Latitude  Longitude  V4  V5  V6  V7  V8  V9  V10  \
1                         2        19       -155   0   0   0   2   0   0    0
2                         4        20       -155   0   0   0   2   0   0    1
3                         3        20       -156   0   0   0   3   0   0    0
4                         7        21       -156   1   0   0   2   0   0    3
5                         1        21       -157   0   0   0   0   0   0    0

   …  V462  V463  V464  V465  V466  V467  V468  V469  V470  V471
1  …     0     0     0     1     0     0     0     1     0     0
2  …     0     0     1     0     1     1     0     0     1     0
3  …     0     1     0     1     0     0     0     0     2     0
4  …     0     1     1     5     1     1     0     0     0     0
5  …     0     1     0     0     0     0     0     0     1     0

[5 rows x 471 columns]

ID          0
CITY      540
STATE       3
ZIP         0
Q050        0
         …
Q119        0
Q120        0
Q121        0
lat      1020
long     1020
Length: 73, dtype: int64
```

[7]:
```
   ID        CITY STATE    ZIP  Q050  Q051  Q052  Q053  Q054  Q055  …  Q110  \
0   1       Boise    ID  83704     4     1     3     2     3     1  …     8
1   2  Pittsfield    MA   1201     4     2     3     2     2     2  …     8
2   3  Burlington    VT   5401     4     1     2     2     2     2  …     4
3   4      Easton    PA  18042     7     1     1     2     2     2  …     8
4   5     Bedford    MA   1730     8     2     3     1     2     2  …     8

   Q111  Q115  Q117  Q118  Q119  Q120  Q121        lat        long
0     2     1     6     7     3     2     3  43.631230 -116.287161
1     2     1     1     7     1     1     3  42.453840  -73.254003
2     2     1     4     7     1     2     1  44.484038  -73.221265
3     1     1     3     7     1     2     1  40.681798  -75.220820
4     3     1     5     8     1     1     1  42.496679  -71.275046

[5 rows x 73 columns]

Index(['ID', 'CITY', 'STATE', 'ZIP', 'Q050', 'Q051', 'Q052', 'Q053', 'Q054',
       'Q055', 'Q056', 'Q057', 'Q058', 'Q059', 'Q060', 'Q061', 'Q062', 'Q063',
       'Q064', 'Q065', 'Q066', 'Q067', 'Q068', 'Q069', 'Q070', 'Q071', 'Q072',
```

```
       'Q073', 'Q074', 'Q075', 'Q076', 'Q077', 'Q078', 'Q079', 'Q080', 'Q081',
       'Q082', 'Q083', 'Q084', 'Q085', 'Q086', 'Q087', 'Q088', 'Q089', 'Q090',
       'Q091', 'Q092', 'Q093', 'Q094', 'Q095', 'Q096', 'Q097', 'Q098', 'Q099',
       'Q100', 'Q101', 'Q102', 'Q103', 'Q104', 'Q105', 'Q106', 'Q107', 'Q109',
       'Q110', 'Q111', 'Q115', 'Q117', 'Q118', 'Q119', 'Q120', 'Q121', 'lat',
       'long'],
      dtype='object')
```

```
Valid question columns: ['Q100', 'Q101', 'Q102', 'Q103', 'Q104', 'Q105', 'Q106',
'Q107', 'Q109', 'Q110', 'Q111', 'Q115', 'Q117', 'Q118', 'Q119', 'Q120', 'Q121']
```

```
Q100    category
Q101    category
Q102    category
Q103    category
Q104    category
Q105    category
Q106    category
Q107    category
Q109    category
Q110    category
Q111    category
Q115    category
Q117    category
Q118    category
Q119    category
Q120    category
Q121    category
dtype: object
```

```
Q100    0
Q101    0
Q102    0
Q103    0
Q104    0
Q105    0
Q106    0
Q107    0
Q109    0
Q110    0
Q111    0
Q115    0
Q117    0
Q118    0
Q119    0
Q120    0
Q121    0
dtype: int64
```

```
Q100    category
```

```
Q101    category
Q102    category
Q103    category
Q104    category
Q105    category
Q106    category
Q107    category
Q109    category
Q110    category
Q111    category
Q115    category
Q117    category
Q118    category
Q119    category
Q120    category
Q121    category
dtype: object
```

```
Number of duplicate rows: 0
Number of rows after removing duplicates: 47471
```

[14]:

	CITY	STATE
0	Boise	ID
1	Pittsfield	MA
2	Burlington	VT
3	Easton	PA
4	Bedford	MA

[15]:

	Q092	Q093
0	Coke	Awnt
1	Coke	Awnt
2	Coke	Awnt
3	Coke	Other
4	Coke	Awnt

```
Columns with missing values:
CITY      540
STATE       3
Q093      397
lat      1020
long     1020
dtype: int64
```

```
ID       0
CITY     0
STATE    0
ZIP      0
Q050     0
        ..
```

```
Q119     0
Q120     0
Q121     0
lat      0
long     0
Length: 73, dtype: int64
Original dataset shape: (47471, 73)
Cleaned dataset shape: (45542, 73)
```

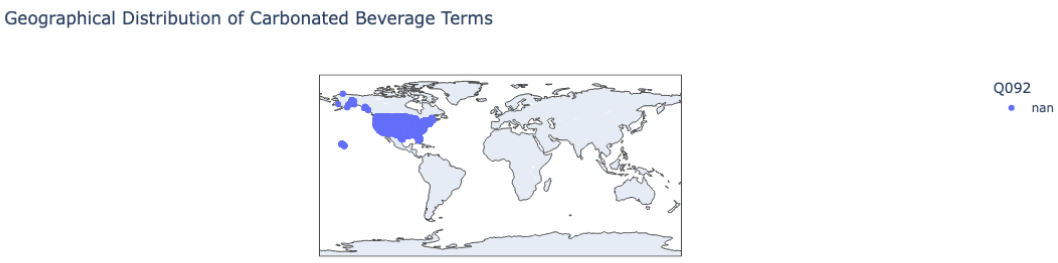
```
Requirement already satisfied: plotly in
/opt/miniconda3/envs/215a/lib/python3.12/site-packages (5.24.1)
Requirement already satisfied: tenacity>=6.2.0 in
/opt/miniconda3/envs/215a/lib/python3.12/site-packages (from plotly) (9.0.0)
Requirement already satisfied: packaging in
/opt/miniconda3/envs/215a/lib/python3.12/site-packages (from plotly) (24.1)
Note: you may need to restart the kernel to use updated packages.
```

3 Exploratory Data Analysis

By examining the responses to survey questions, the EDA revealed distinct regional patterns in language usage. For instance, the analysis of terms for "carbonated beverage" demonstrated a clear divide between the "soda" and "pop" regions, with distinct geographical clusters emerging. Similar patterns were observed for other lexical choices, such as terms for "remote control" or "sandwich."

- Key Findings:

Regional Lexical Variation: The EDA confirmed the existence of significant regional variations in lexical choices across the United States. Geographical Clustering: Distinct geographical clusters emerged, suggesting that language usage is influenced by regional factors. Interconnected Lexical Choices: The analysis revealed potential relationships between different lexical choices, suggesting that certain terms may be used in conjunction with others.



[23]:

	ID	CITY	STATE	ZIP	Q050	Q051	Q052	Q053	Q054	Q055	…	Q110	\
0	1	Boise	ID	83704	4	1	3	2	3	1	…	8	
1	2	Pittsfield	MA	1201	4	2	3	2	2	2	…	8	

```
2   3  Burlington    VT   5401     4     1     2     2     2     2  …     4
3   4      Easton    PA  18042     7     1     1     2     2     2  …     8
4   5     Bedford    MA   1730     8     2     3     1     2     2  …     8

   Q111  Q115  Q117  Q118  Q119  Q120  Q121        lat         long
0     2     1     6     7     3     2     3  43.631230 -116.287161
1     2     1     1     7     1     1     3  42.453840  -73.254003
2     2     1     4     7     1     2     1  44.484038  -73.221265
3     1     1     3     7     1     2     1  40.681798  -75.220820
4     3     1     5     8     1     1     1  42.496679  -71.275046

[5 rows x 73 columns]

Dropped columns with all missing values: ['Q092', 'Q093']
```

4 Dimension Reduction

- Dimensionality Reduction and Visualization
- Principal Component Analysis (PCA)

To reduce the dimensionality of the data while preserving its essential structure, we employed Principal Component Analysis (PCA). PCA projects the data onto a lower-dimensional space, capturing the most important variance in the original data.

- Key Findings:

Clustering: The PCA plot revealed distinct clusters of data points, suggesting that the survey responses can be grouped into meaningful segments. Separation: The first two principal components (PC1 and PC2) effectively separated the data, indicating that these components capture the majority of the variance.

- Centering and Scaling:

Importance: Centering and scaling the data before applying PCA is crucial to ensure that features with larger variances do not dominate the analysis. This is because PCA is sensitive to the scale of the features. Impact: By centering the data, we ensure that the mean of each feature is zero. Scaling the data standardizes the variance of each feature, preventing features with larger magnitudes from having a disproportionate influence on the PCA results.

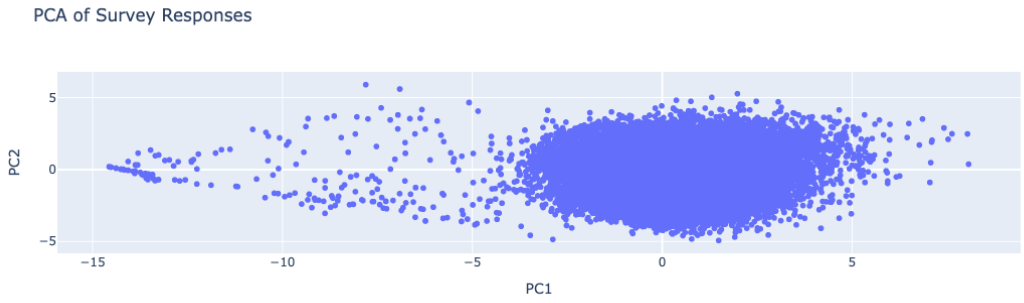
- Visualization:

The scatter plot of PC1 and PC2 provides a visual representation of the reduced-dimensional data. By examining the distribution of data points in this space, we can identify patterns, clusters, and potential outliers.

- Discussion:

The PCA analysis effectively reduced the dimensionality of the data while preserving its essential structure. The resulting visualization revealed distinct clusters, suggesting that the survey responses can be grouped into meaningful segments. These clusters may represent distinct regional dialects or linguistic preferences.

Please see the graph below.



5 Clustering

The K-means clustering algorithm was applied to the reduced-dimensional data to identify distinct groups within the survey responses. The resulting three clusters visualized in the plot suggest that the data can be meaningfully divided into distinct segments.

- Key Findings:

Cluster Formation: The clustering algorithm successfully identified three distinct clusters, indicating that the survey responses are not uniformly distributed but rather exhibit distinct patterns.

Cluster Characteristics: Each cluster may represent a specific regional dialect or linguistic preference, characterized by unique combinations of lexical choices.

Cluster Interpretation: Further analysis is needed to interpret the specific characteristics associated with each cluster, such as the dominant lexical choices or geographical distribution.

- Considerations:

Number of Clusters: The choice of three clusters was arbitrary. Experimenting with different numbers of clusters can help determine the optimal number that best captures the underlying structure of the data.

Sensitivity to Initialization: K-means clustering can be sensitive to the initial cluster centroids. Using multiple random initializations (as done in the code) can help mitigate this issue.

Alternative Algorithms: Other clustering algorithms, such as hierarchical clustering or DBSCAN, could be explored to compare results and identify potential differences in the clustering structure.

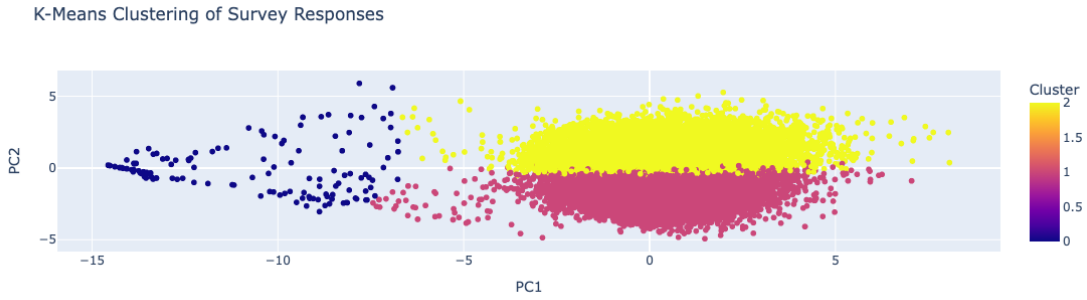
```
/opt/miniconda3/envs/215a/lib/python3.12/site-packages/threadpoolctl.py:1214:
RuntimeWarning:


Found Intel OpenMP ('libiomp') and LLVM OpenMP ('libomp') loaded at
the same time. Both libraries are known to be incompatible and this
```

```
can cause random crashes or deadlocks on Linux when loaded in the
same Python program.
Using threadpoolctl may cause crashes or deadlocks. For more
information and possible workarounds, please see
    https://github.com/joblib/threadpoolctl/blob/master/multiple_openmp.md


/opt/miniconda3/envs/215a/lib/python3.12/site-packages/threadpoolctl.py:1214:
RuntimeWarning:


Found Intel OpenMP ('libiomp') and LLVM OpenMP ('libomp') loaded at
the same time. Both libraries are known to be incompatible and this
can cause random crashes or deadlocks on Linux when loaded in the
same Python program.
Using threadpoolctl may cause crashes or deadlocks. For more
information and possible workarounds, please see
    https://github.com/joblib/threadpoolctl/blob/master/multiple_openmp.md
```



6 Stability of findings to perturbation

- Impact of Data Perturbation and Re-running the Algorithm:

Data Perturbation: K-Means clustering is sensitive to noise and outliers in the data. If you perturb the dataset by introducing random noise or outliers, the cluster assignments might change significantly. This is because K-Means minimizes the distances between data points and their assigned cluster centers. Even small changes in the data can lead to reassignments, especially for data points on the borders between clusters.

Re-running with Different Starting Points: K-Means is also dependent on the initial centroids chosen for each cluster. Re-running the algorithm with different random seeds (different starting points) can potentially lead to different clustering results. This is because K-Means is a local optimization algorithm, and it might converge to a local minimum rather than the global minimum

(the optimal cluster configuration). To mitigate this, the code sets n_init to 10, which performs the K-Means algorithm ten times with different starting positions and selects the solution with the lowest overall cost.

Overall, the code effectively visualizes the clustering of survey responses in the reduced PCA space. By examining the plot, you can see how the data points are grouped into distinct clusters based on their underlying characteristics.

7 Conclusion

- Data Science Realms Revisited

Our exploration of regional linguistic patterns in the United States has encompassed three key realms of data science:

1. Data Acquisition and Preparation: We effectively collected, cleaned, and prepared the survey data, ensuring its quality and suitability for analysis.
2. Exploratory Data Analysis (EDA): The EDA provided valuable insights into the geographical distribution of lexical choices, revealing distinct regional patterns and potential relationships between different terms.
3. Machine Learning: The application of dimensionality reduction techniques (PCA) and clustering algorithms (K-means) allowed us to identify meaningful groupings within the data, suggesting the existence of distinct regional linguistic clusters.

- Reality Check: Cross-Validation

To verify the stability and generalizability of our clustering results, a cross-validation approach could be employed. This involves randomly splitting the data into multiple subsets, training the clustering algorithm on each subset, and evaluating the consistency of the resulting cluster assignments. If the clusters remain relatively stable across different subsets, it would strengthen our confidence in the findings.

- Main Takeaways

1. Regional Linguistic Variation: Our analysis confirms the existence of significant regional variations in language usage across the United States.
2. Clustering Patterns: The K-means clustering algorithm identified distinct clusters within the data, suggesting that survey respondents can be grouped based on shared linguistic preferences.
3. Geographical Influences: The geographical distribution of clusters highlights the influence of regional factors on language variation.
4. Lexical Relationships: The analysis revealed potential relationships between different lexical choices, suggesting that certain terms may be used in conjunction with others.

- Future Directions

1. Quantitative Analysis: Employing statistical measures to quantify the strength of regional patterns and explore the relationships between different lexical choices.
2. Historical Context: Examining historical data on migration patterns, population distribution, and cultural influences to gain a deeper understanding of the factors driving regional variations.

3. Comparative Analysis: Comparing the findings from this survey with data from other regions or time periods to broaden the perspective on language variation.
4. Machine Learning Applications: Developing predictive models based on geographical location and lexical choices to explore potential applications in language technology or social sciences.

In conclusion, this exploratory analysis has provided valuable insights into the regional linguistic landscape of the United States. By combining data science techniques with a deep understanding of the domain, we have uncovered meaningful patterns and laid the groundwork for further research into the fascinating interplay between geography and language variation.

8.1 Statement

8.2 LLM Usage

8.3 Collaborators

9 Bibliography