Final Project

Csenge Szabó
c.szabo@student.vu.nl

1 Introduction

This project investigates the portrayal of immigration in online news articles from August to October 2023, a period marked by various geopolitical conflicts, such as the Russia-Ukraine war and the Israel-Hamas conflict, leading to waves of immigration affecting thousands of people (Eurostat, 2023). The objective is to compare the terminological and thematic aspects of English and German articles published in the U.S. and in Germany, assuming that English articles primarily focus on U.S.-related immigration matters, while German articles report on immigration in Germany and Europe. KMeans clustering will be used to identify thematic links between grouped articles and explore if there are recurring topics in the clusters, such as migration legislation, social matters and border crossings.

2 Dataset Description

A dataset of 200 articles per language was crawled using Media Stack API within the defined timeframe and divided into training and test sets in an 80-20% ratio. Keywords were 'immigration' and 'immigrant' for English, and 'Einwanderung,' 'Einwanderer,' 'Flüchtlinge,' and 'Migration'[1] for German. Specific steps were implemented to filter out articles with identical content or URL, articles shorter than 100 characters and to restrict the data to web domains present in the U.S. and Germany.

Preprocessing involved extracting author information from the URLs, and replacing it with "Unknown" for instances where the author information mirrored the domain name or was not available. Moreover, URLs, tags, newline characters, tabs, italics and bold markup, numbers, emojis, a set of special characters, and repetitive subscription-related phrases were removed from the textual content before the analysis.

[1] translate as: immigration, immigrant, refugee, migration

Regarding the temporal distribution of articles, most articles were published in September 2023, shown in Figure 1. Most English articles were published between 18:00 and 24:00, German articles were prevalent between 12:00 and 18:00. The ratio of unknown to known author ratio exhibited 40-60% in English and 56.25-43.75% in German articles, with unique authors reaching 70 in English training data and 63 in German training data.

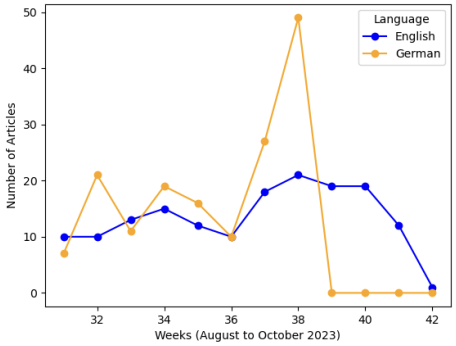

Figure 1: Weekly publication counts of English and German articles between August and October 2023

The most frequent domain in the German data is *Der Tagesspiegel*, while the least represented is *Süddeutsche Zeitung*. In the English data, the most prominent source is *Los Angeles Times*, while *Arizona Capitol Times* is the least common. 9 and 24 different domains hosted the German and English articles, respectively, see Figure 2 and Figure 3.

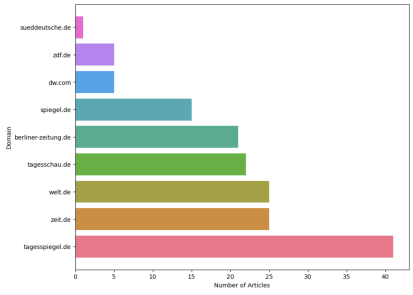

Figure 2: Distribution of German articles across web domains

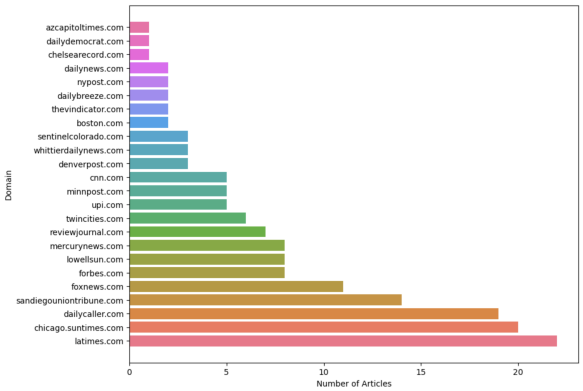

Figure 3: Distribution of English articles across web domains

The cleaned articles in the training data were processed with a Stanza pipeline. Words defined in the NLTK stop word list extended with additional function words were removed from the textual content. On average, English articles show a higher variability in sentence and token counts, with larger means and standard derivation values compared to their German counterparts, see Table 1. When focusing on the distribution of tokens in the articles, an identical pattern can be recognised across the data, illustrated in Figure 4.

	Mean	Std	Median	Mode	Variance	Min	Max
EN Titles tok.	14	4	14	14	13	4	23
DE Titles tok.	12	4	11	10	13	4	26
EN Articles sent.	39	25	37	39	644	3	223
DE Articles sent.	29	24	24	7	576	2	176
EN Articles tok.	781	439	728	328	193066	62	3412
DE Articles tok.	462	407	400	99	165617	35	3303

Table 1: Statistical information about title lengths in terms of tokens and article lengths in terms of sentence and token counts for English (EN) and German (DE).

The type-token ratio (TTR), which is a metric to indicate lexical variation, is 0.189 for English, and 0.270 for German. The most common Named Entities (NEs) in the English data are organizations, persons and geographical entities; conversely, these are locations, organizations and persons in the German data. Nouns, verbs, proper nouns, adjectives and adverbs constituted the most prevalent part-of-speech tag categories in both languages, see Table 2[2].

[2]Top 10 German tokens translate as: people, said, Germany, more, EU, immigration, be, immigrants, Italy, be

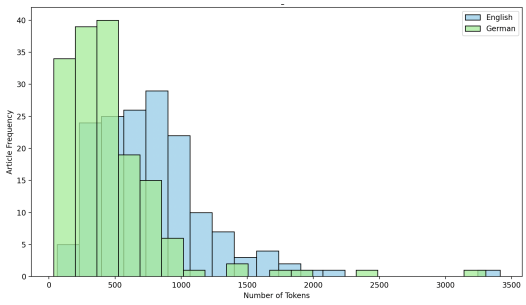

Figure 4: Distribution of tokens in English and German articles using a Stanza pipeline

3 Explorative Linguistic Analysis

This paper examines whether German articles discuss immigration related to different geographical regions than English articles. Secondly, it explores clustering articles based on their unique keywords, aiming to unveil connections between metadata and the thematic clusters.

Named Entity Recognition

SpaCy pipelines[3] trained on written media were used to process the articles and extract geographical locations (LOC and GPE) from the training data. Each location was counted once per article to account for varying article lengths and repeated mentions. By comparing the ten most frequent geographical locations in the datasets, a distinct pattern is noticeable: German articles mainly discuss immigration observed in Europe, Africa, Russia and in the U.S., whereas English articles center their narrative on North and South America, as shown in Figure 5 and Figure 6.

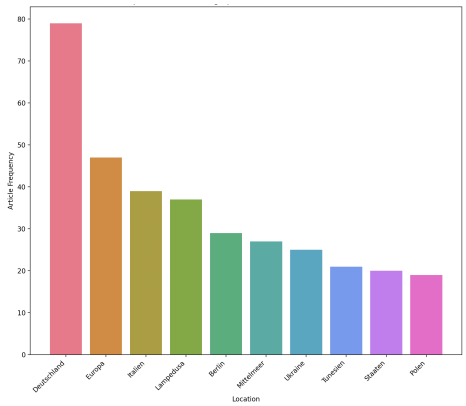

Figure 5: 10 most frequent geographical locations in German articles extracted with SpaCy. From left to right: Germany, Europe, Italy, Lampedusa, Berlin, Mediterranean Sea, Ukraine, Tunisia, States, Poland.

[3]'de_core_news_md' and 'en_core_web_sm'

Category	Frequency
NEs (EN)	Person: 4083, Location: 471, Ordinal: 164, Geo-Political Entity: 3075, Facility: 424, Date: 2282, Time: 140, Organization: 4753, Nationality or Religious or Political Group: 973, Quantity: 27
NEs (DE)	Person: 1969, Location: 3305, Organization: 2192, Miscellaneous: 141
POS-tags (EN)	Noun: 28455, Verb: 15541, Proper Noun: 13970, Adjective: 9002, Adverb: 2978, Adposition: 907, Auxiliary Verb: 749, Number: 725, Pronoun: 368, Subordinating Conjunction: 221
POS-tags (DE)	Noun: 16398, Proper Noun: 7143, Verb: 7068, Adjective: 5823, Adverb: 3106, Adposition: 592, Auxiliary Verb: 576, Determiner: 286, Number: 263, Other: 119
Top 10 tokens (EN)	said: 907, U.S.: 405, immigration: 394, immigrants: 351, border: 349, people: 252, also: 248, state: 237, migrants: 236, would: 232
Top 10 tokens (DE)	Menschen: 360, sagte: 348, Deutschland: 313, mehr: 300, EU: 295, Migration: 231, sei: 207, Migranten : 200, Italien: 157, seien: 141

Table 2: 10 most frequent Named Entity and POS-tag categories, 10 most frequent tokens in English and German articles after preprocessing using a Stanza pipeline.

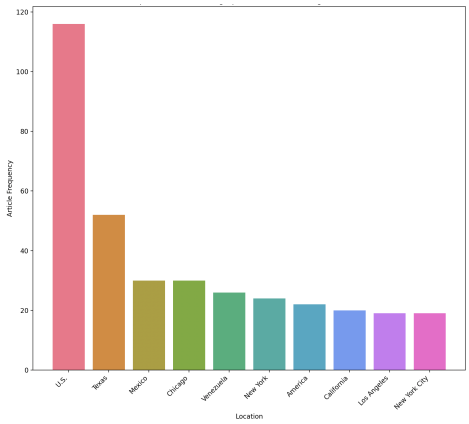

Figure 6: 10 most frequent geographical locations in English articles extracted with SpaCy

Examining location entity distributions across domains reveals that certain web sources mention more NE locations than others. *Tagesschau*, which is the fourth most represented domain in the dataset mentions locations the most frequently, followed by *Tagesspiegel* and *Die Zeit*. In the English data, the distribution of location entities across web domains approximately mirrors the domain distribution in the dataset, see Figure 10.

Thematic Clustering

The second research question aims to explore whether articles could be clustered based on their keywords, and if discernible trends are identifiable in the domain distributions of each cluster. Using the lemmas of the article contents, TF-IDF was utilised to extract relevant unique key words from each article. Additionally, word embeddings from pre-trained FastText models were utilized to capture semantic connections between keywords [4].

[4]For English, the 'wiki-news-300d-1M.vec' model was used, while German utilized 'wiki.de.zip', which contains pre-trained word vectors obtained using the skip-gram model (Bojanowski et al., 2016).

Silhouette Scores[5] were calculated to identify optimal KMeans clustering for the datasets. Three clusters were created per language, which is an optimal clustering point based on the scores, see Figure 7. Unique keywords were identified, and word clouds were generated to represent the find out about the distinguishing terms in each cluster.

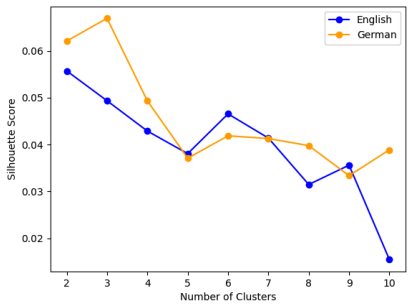

Figure 7: Silhouette Scores for optimal KMeans clustering of English and German articles

The article distribution of clusters reveals an imbalance in the German data, with one cluster only containing 4.4% of the articles, whereas English articles are fairly evenly distributed across clusters, as depicted in Figure 8. Standard deviation was used to quantify the balance, which returns a value of 10.69 for the English data. The German data exhibits a higher standard deviation of 44.19, which quantified the degree of imbalance in distribution.

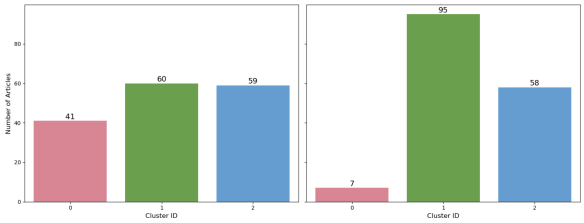

Figure 8: The distribution of English (left) and German (right) articles per KMeans clusters

[5]https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

The examination of **keywords** within the established clusters provides an insight into their thematic orientation. The majority of English articles discuss U.S. immigration legislation and politics (cluster 1), described by keywords like "program", "law" and "congress". The second largest cluster (cluster 2) discusses regional immigration issues, characterised by city names and words related to social matters. Articles related to Mexican and international border crossings belong to the smallest cluster (cluster 0).

In the German data, 59.4% of articles belong to cluster 1, grouped around political discourse connected to Germany's role in European migration policy. Articles in cluster 2 describe immigration in Italy and other hotspots, such as the UK and Tunisia, while cluster 0 centres on the Ukraine-Russia conflict-related immigration, see Figure 9.

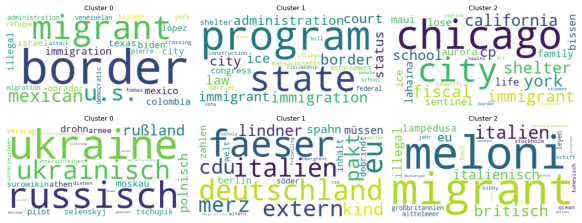

Figure 9: 25 most frequent keywords across the English (top row) and German clusters (bottom row)

Exploring the relationship between clusters and publication time revealed no discernible trends due to the limited timeframe. The connection between **clusters and domains** however, allows us to identify the prevalent thematic preferences of news platforms. The majority of domains in the German data have a shared presence across all clusters. Examples include *Tagesspiegel* and *Welt*, each appears in all clusters, shown in Figure 11.

Conversely, this pattern differs in the English dataset, where only domains with the largest representation, such as the *Los Angeles Times*, are present in all clusters. Interestingly, some domains with average representation, such as *CNN*, are only present in one cluster, suggesting that such domains may represent immigration from a single thematic perspective.

4 Discussion

NER results revealed that English articles predominantly focus on immigration present in the Americas, whereas NEs in German articles encompass a broader range of continents. Observing the distribution of articles across thematic clusters shows a significant imbalance in the German dataset. Conversely, the distribution among English clusters is more balanced. The unique keywords in these clusters underscored different thematic emphases, with German articles focusing on German and European migration policy, immigration in Italy and the UK, and issues related to the Ukraine-Russia war. In contrast, English articles predominantly address U.S. immigration legislation, domestic immigration matters and border crossings in the Americas.

Hypotheses

1. The majority(>50%) of the 10 most common NE locations in the English test data are expected to be places in the Americas, while the majority(>50%) of German locations will be parts of Europe, Asia, Africa or the Americas.

2. German clusters will show a larger cluster imbalance compared to the English clusters in test data when applying the same clustering methods, with a minimum double standard deviation value.

3. TTR will be higher for German than English in the test data, indicating a sustained linguistic diversity, with a minimum difference of 0.08. The code includes calculating confidence intervals and conducting Randomized Test for TTR.

Alternative explanations

The geographical orientation observed in the data could be impacted by events that happened within the selected timeframe rather than inherent regional preferences. Moreover, the imbalance in German clusters might be the result of the chosen clustering algorithm and crawled data, rather than a true thematic concentration. Finally, linguistic diversity could be influenced by the writing style of the selected articles, and not be a true reflection of language characteristics. Extending the analysis for a longer period may provide a more comprehensive understanding of the discourse on immigration. Considering alternative clustering methods might offer different perspectives on the thematic distinctions among the clusters, which could be topic of further research. The findings presented in this paper are constrained to the timeframe, selected sources and languages.

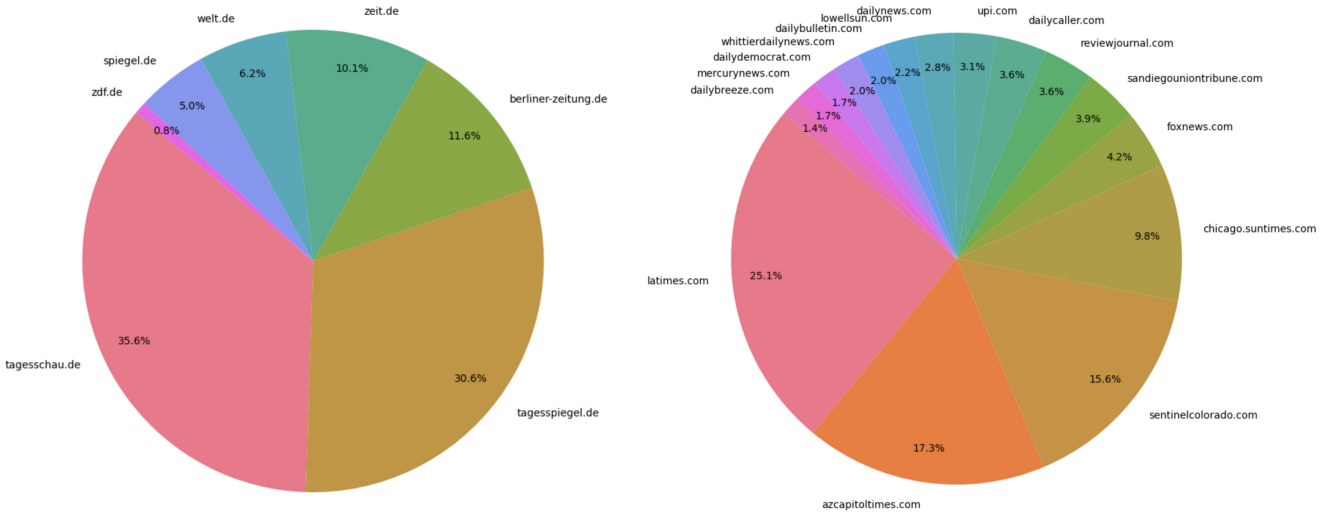

Figure 10: Distribution of location Named Entities for all web domains in the German (left) and English (right) training data extracted with SpaCy

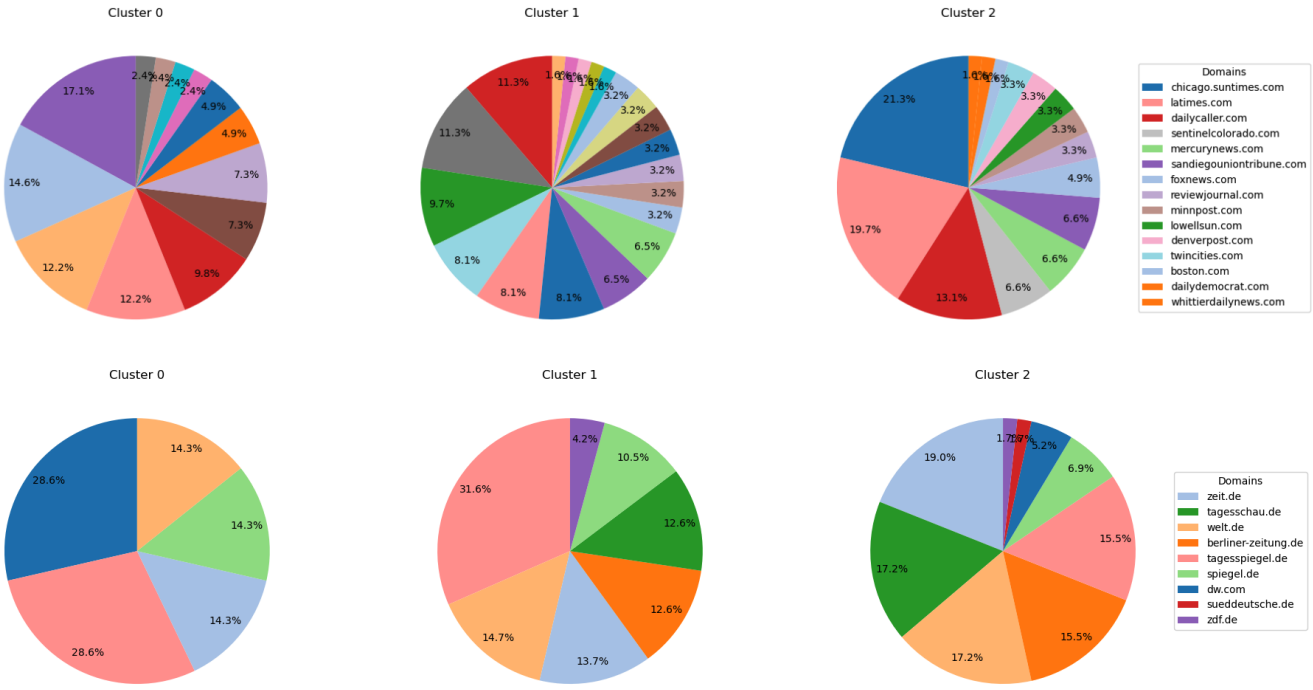

Figure 11: Web domain distribution across clusters in the English (top row) and German training data (bottom row)

A Additional Results

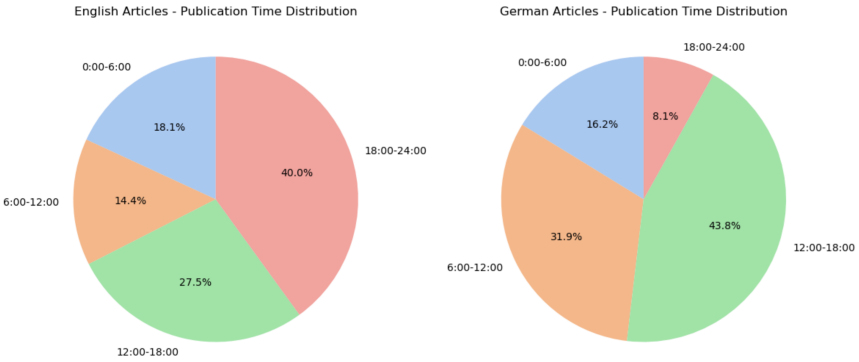

Figure 12: Daily publication time of English and German articles

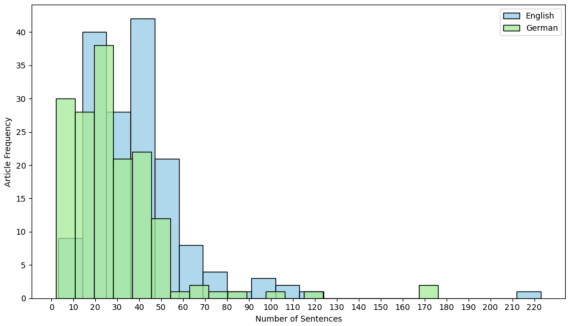

Figure 13: Distribution of sentences in English and German articles created with Stanza pipeline

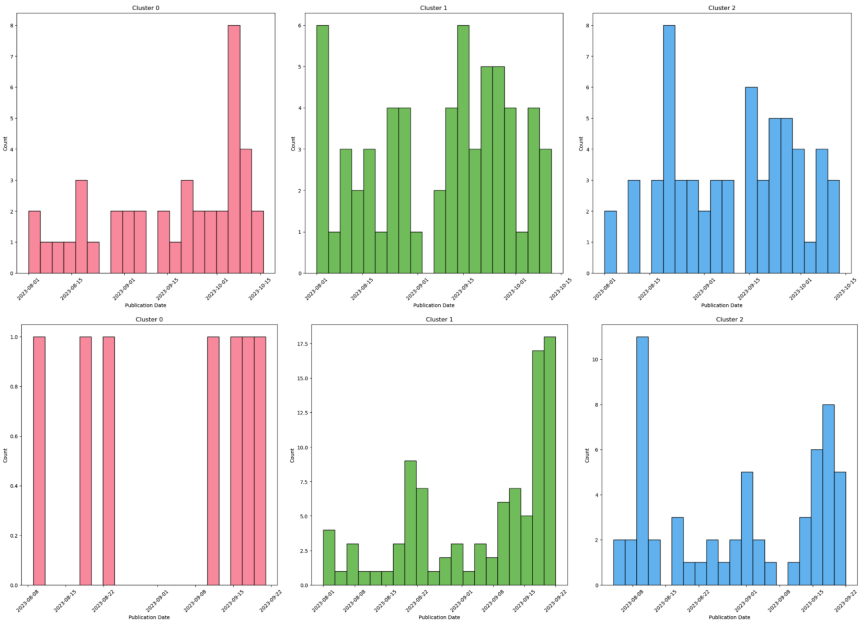

Figure 14: Publication date distribution across clusters in the English (top row) and German training data (bottom row)

B Software packages

- Stanza version 1.6.1. Documentation: https://stanfordnlp.github.io/stanza/
- Spacy version 3.6.1. Documentation: https://spacy.io/api/doc
- NLTK version 3.8.1. Documentation: https://www.nltk.org
- scikit-learn version 1.3.0. Documentation: https://scikit-learn.org/stable/
- Pandas version 2.1.1. Documentation: https://pandas.pydata.org/docs/
- Matplotlib version 3.7.2. Documentation: https://matplotlib.org/stable/contents.html
- Seaborn version 0.13.0. Documentation: https://seaborn.pydata.org/
- Wordcloud version 1.9.2. Documentation: https://pypi.org/project/wordcloud/
- Gensim version 4.3.0. Documentation: https://pypi.org/project/gensim/
- SciPy version 1.11.1. Documentation: https://docs.scipy.org/doc/scipy-1.11.1/

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information.

Eurostat. 2023. Returns of irregular migrants - quarterly statistics. Accessed: 20 November 2023.