UkraineWordTrends_vignette

INTRODUCTION

This vignette provides a comprehensive analysis of word trends in Ukraine Conflict articles, offering insights into the evolving narrative and emphasizing key terms for each journal. It explores the trends in word usage in articles related to the Ukraine Conflict. We will cover the following aspects:

1. Data Collection and Joint Dataset.
2. Word Frequency Over Time.
3. TF-IDF Analysis.
4. Zipf's Law and Linear Regression

Data Collection and Joint Dataset

We obtained two original datasets from The New York Times "NYT_Russia_Ukraine" and The Guardian "Guardians_Russia_Ukraine" csv files. After importing and combining the data, we created a joint dataset named "Combined_Guardian_NYT.csv". This dataset contains information such as publication dates, headlines, articles, and respective journals.
The joint dataset is obtained by combining two original datasets: one from The New York Times (NYT) and the other from The Guardian. The **`read_csv`** function is used to read the data from each CSV file, and a new variable **`journal`** is created to identify the source of each article. The datasets are then combined using the **`bind_rows`** function, and the resulting combined dataset is written to a new CSV file named "Combined_Guardian_NYT.csv."

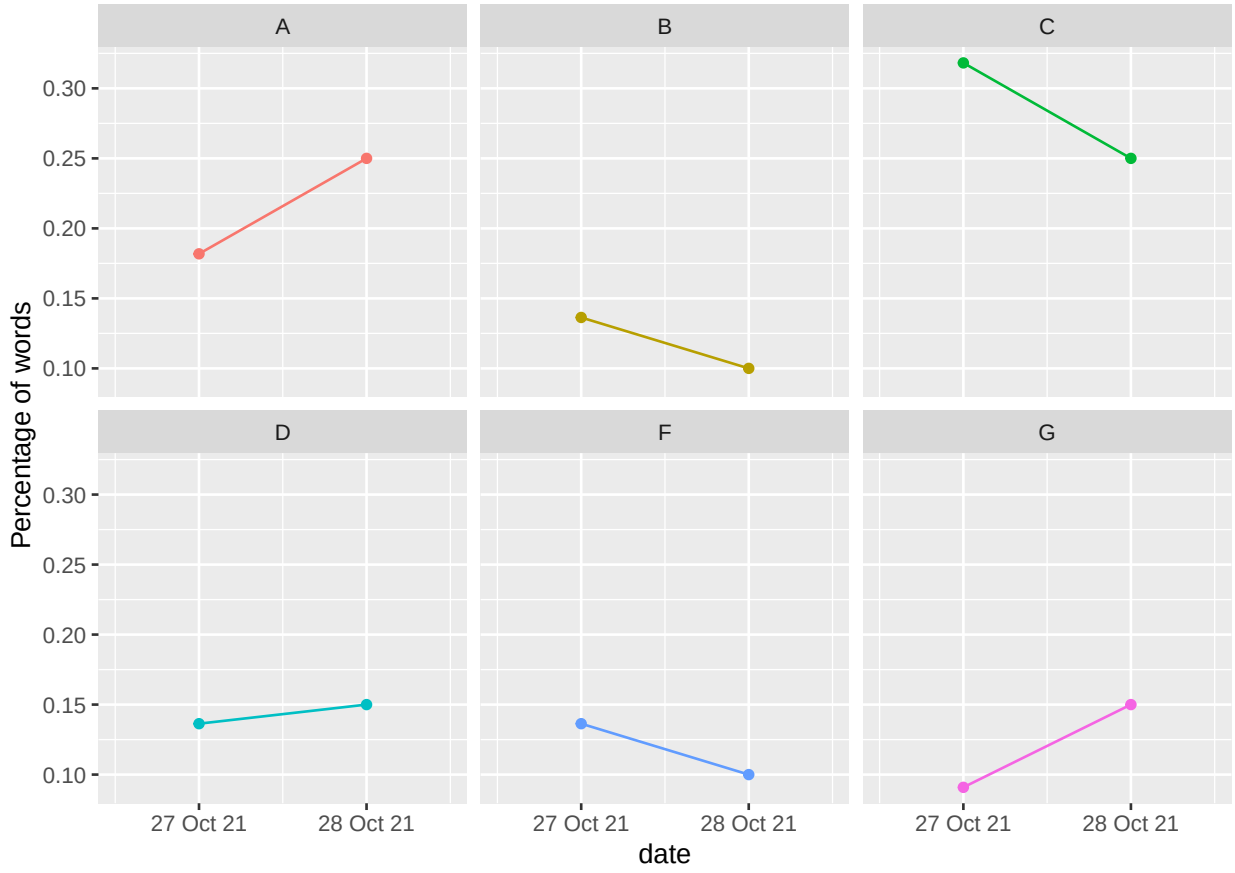
Data Collection and Joint Dataset

```
#> # A tibble: 407 x 4
#>    published           headlines                articles journal
#>    <dttm>              <chr>                    <chr>    <chr>
#>  1 2022-08-01 16:23:25 Sanctions against Russi~ "Simon ~ Guardi~
#>  2 2022-07-26 07:54:56 Can Ukrainian forces re~ "In the~ Guardi~
#>  3 2022-08-05 15:00:03 Nightlands review - tal~ "Who ex~ Guardi~
#>  4 2022-08-02 16:28:09 Russia claims US 'direc~ "The ro~ Guardi~
#>  5 2022-07-27 12:40:40 Is Russia killing off t~ "The In~ Guardi~
#>  6 2022-06-21 16:28:56 Russia blocks Telegraph~ "Russia~ Guardi~
#>  7 2022-07-28 11:59:51 Brittney Griner lawyers~ "Brittn~ Guardi~
#>  8 2022-05-22 14:41:11 Russia bans 963 America~ "Russia~ Guardi~
#>  9 2022-07-15 15:51:31 Russia escalating attac~ "A top ~ Guardi~
#> 10 2022-06-14 17:07:26 Russia bans 29 UK journ~ "Russia~ Guardi~
#> # i 397 more rows
```

Invented Dataset

```
#> # A tibble: 12 x 4
#> # Groups:   date [2]
#>    date       word      n      p
#>    <date>     <chr> <int>  <dbl>
#>  1 2021-10-27 A         4 0.182
#>  2 2021-10-27 B         3 0.136
#>  3 2021-10-27 C         7 0.318
#>  4 2021-10-27 D         3 0.136
#>  5 2021-10-27 F         3 0.136
#>  6 2021-10-27 G         2 0.0909
#>  7 2021-10-28 A         5 0.25
#>  8 2021-10-28 B         2 0.1
#>  9 2021-10-28 C         5 0.25
#> 10 2021-10-28 D         3 0.15
#> 11 2021-10-28 F         2 0.1
#> 12 2021-10-28 G         3 0.15
```

Invented Dataset



Word Frequency Over Time

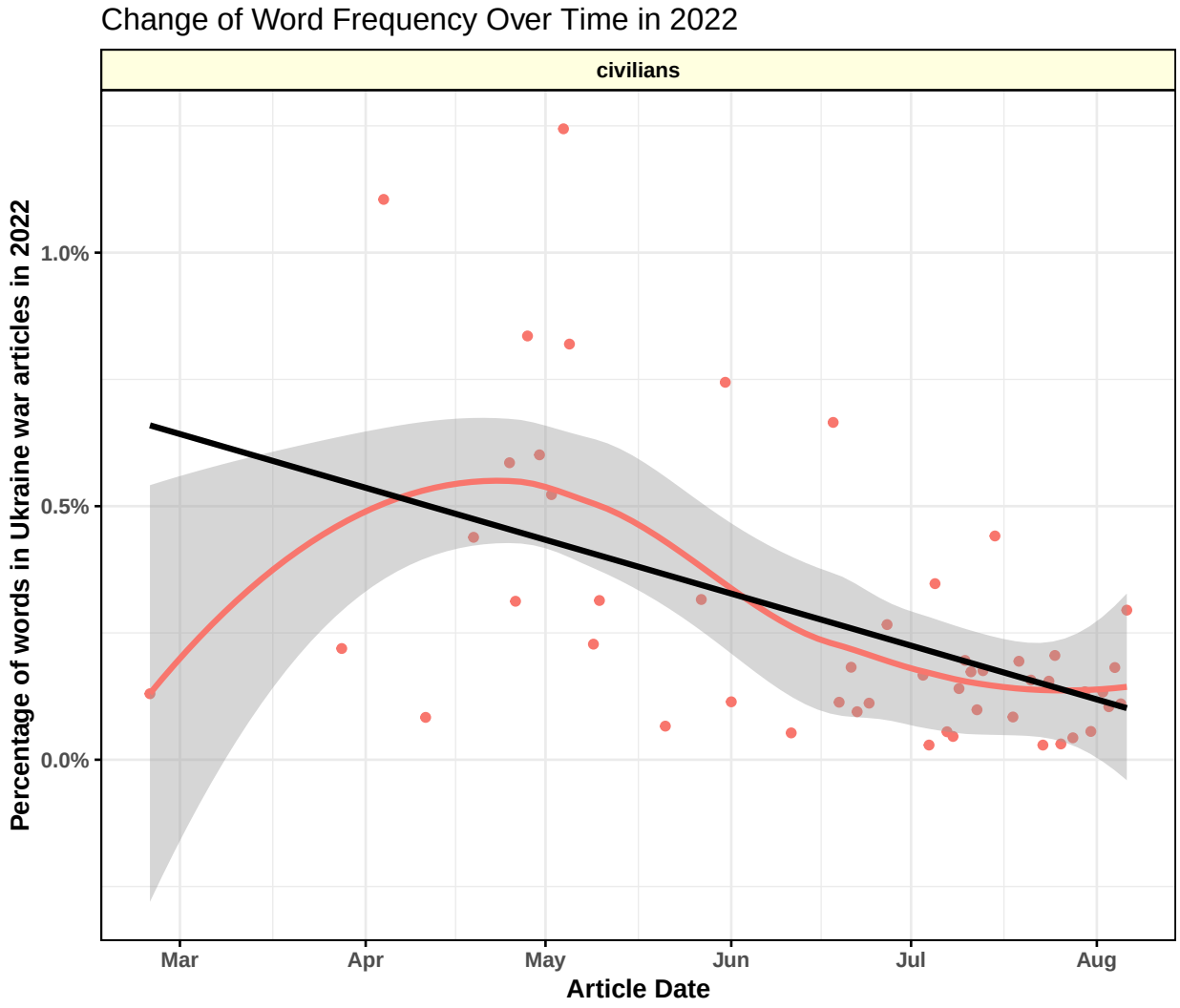
We created a graph showing the change in the frequency of selected words over time in 2022. The selected words include “war”, “Russia”, “Ukraine”, “weapons”, “Biden”, and “NATO”. The graph provides insights into how the usage of these words evolved throughout the year. When examining words such as “war”, “Russia”, “Ukraine”, “weapons”, “Biden”, and “NATO”, we may observe the following:

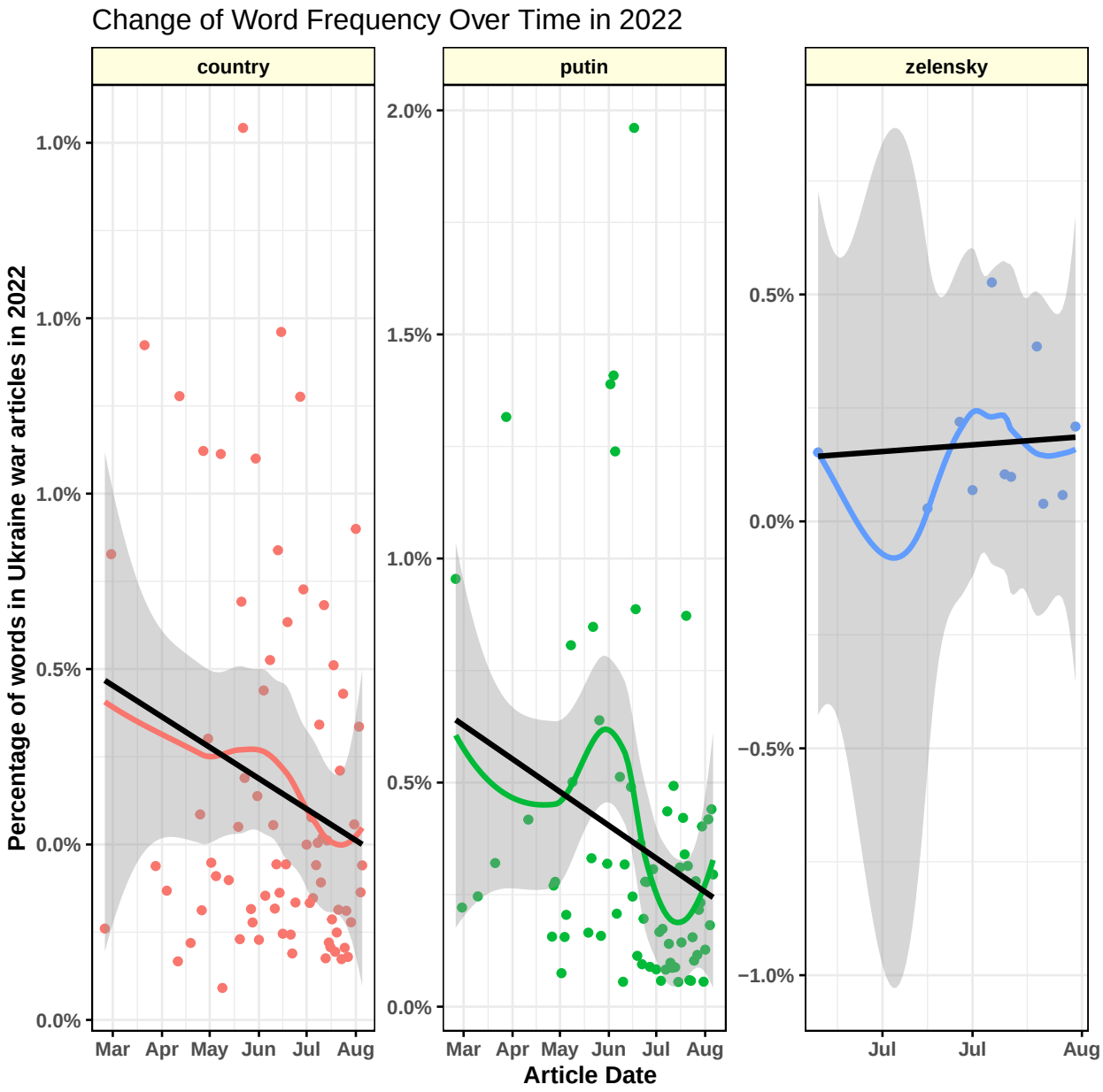
- Peaks in the usage of certain words may correspond with key events in the conflict. For instance, an increase in the mention of “weapons” might coincide with international arms sales or discussions about military aid.
- The frequency of “Ukraine” and “Russia” is likely high throughout, as they are the primary subjects of the articles, but significant fluctuations could indicate shifts in the conflict or international focus.
- The use of the word “war” may increase around escalations in the conflict or significant military engagements.
- Political figures such as “Biden” or references to “NATO” might peak during diplomatic talks, sanctions discussions, or when these entities play a significant role in events related to the conflict.

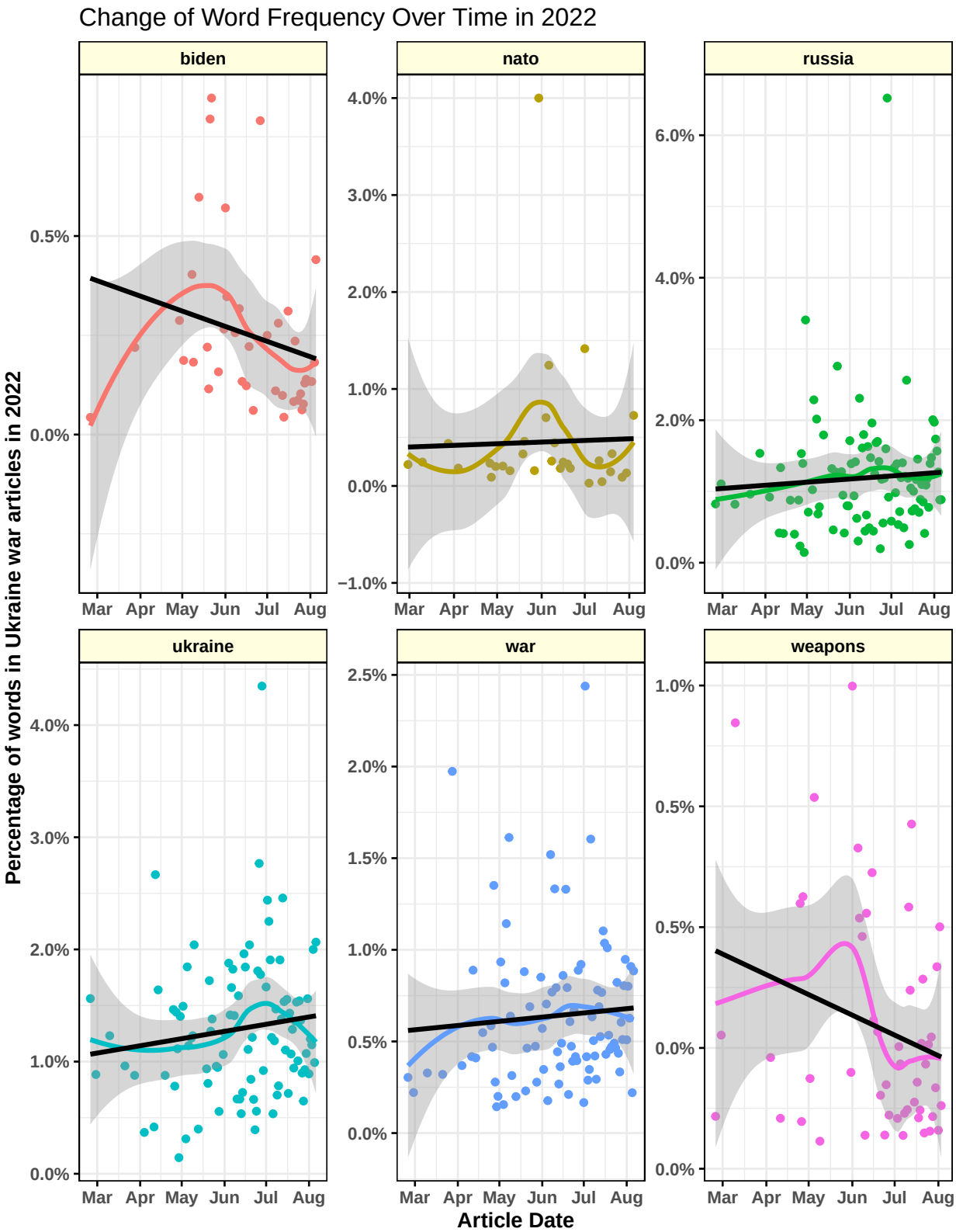
Observations:

- The graph illustrates the dynamic shifts in word usage over time.
- Peaks in certain words may correspond to significant events or development in the conflict.

Word Frequency Over Time




Change of Word Frequency Over Time in 2022
country
putin
zelensky
Percentage of words in Ukraine war articles in 2022
1.0%
1.0%
1.0%
0.5%
0.0%
0.0%
2.0%
1.5%
1.0%
0.5%
0.0%
0.5%
0.0%
−0.5%
−1.0%
Mar
Apr
May
Jun
Jul
Aug
Article Date
Jul
Jul
Aug


Change of Word Frequency Over Time in 2022
biden
nato
russia
ukraine
war
weapons
Percentage of words in Ukraine war articles in 2022
Article Date
Mar
Apr
May
Jun
Jul
Aug
0.0%
0.5%
1.0%
2.0%
3.0%
4.0%
−1.0%
1.5%
2.5%
6.0%

TF-TDF Analysis

We conducted a tf-idf analysis to identify words with the highest tf-idf indexes for each journal. TF_IDF helps highlight words that are important in a specific context.
A function named **plot_word_trends** is defined to plot the change in word frequency over time for a given set of words. The function utilizes the `tidytext` and `ggplot2` packages to tokenize articles, calculate word frequencies, and plot the trends. The function is then applied to three sets of selected words: "civilians," "country," "putin," "zelensky," "biden," "nato," "russia," "ukraine," "war," and "weapons."

From the graph, the following conclusions can be drawn:

- The terms with the highest tf-idf index are those that are characteristically relevant in the Ukraine war articles for each publication. For example, "zelenskiy" likely refers to the President of Ukraine, Volodymyr Zelensky, and is the most prominent term in the "Guardian" panel.
- The difference in the most relevant terms between the two panels suggests that each news outlet has a unique focus and context when covering the Ukraine war. While "Guardian" seems to focus more on political and geographical aspects ("uk," "eu," "un"), the "NYT" includes a variety of terms that might be related to personal stories or cultural elements ("courtney," "sashko," "gay").
- The use of tf-idf helps to identify significant keywords that may be unique to the coverage of a specific topic across different news sources, which is useful for content analysis and media comparison.

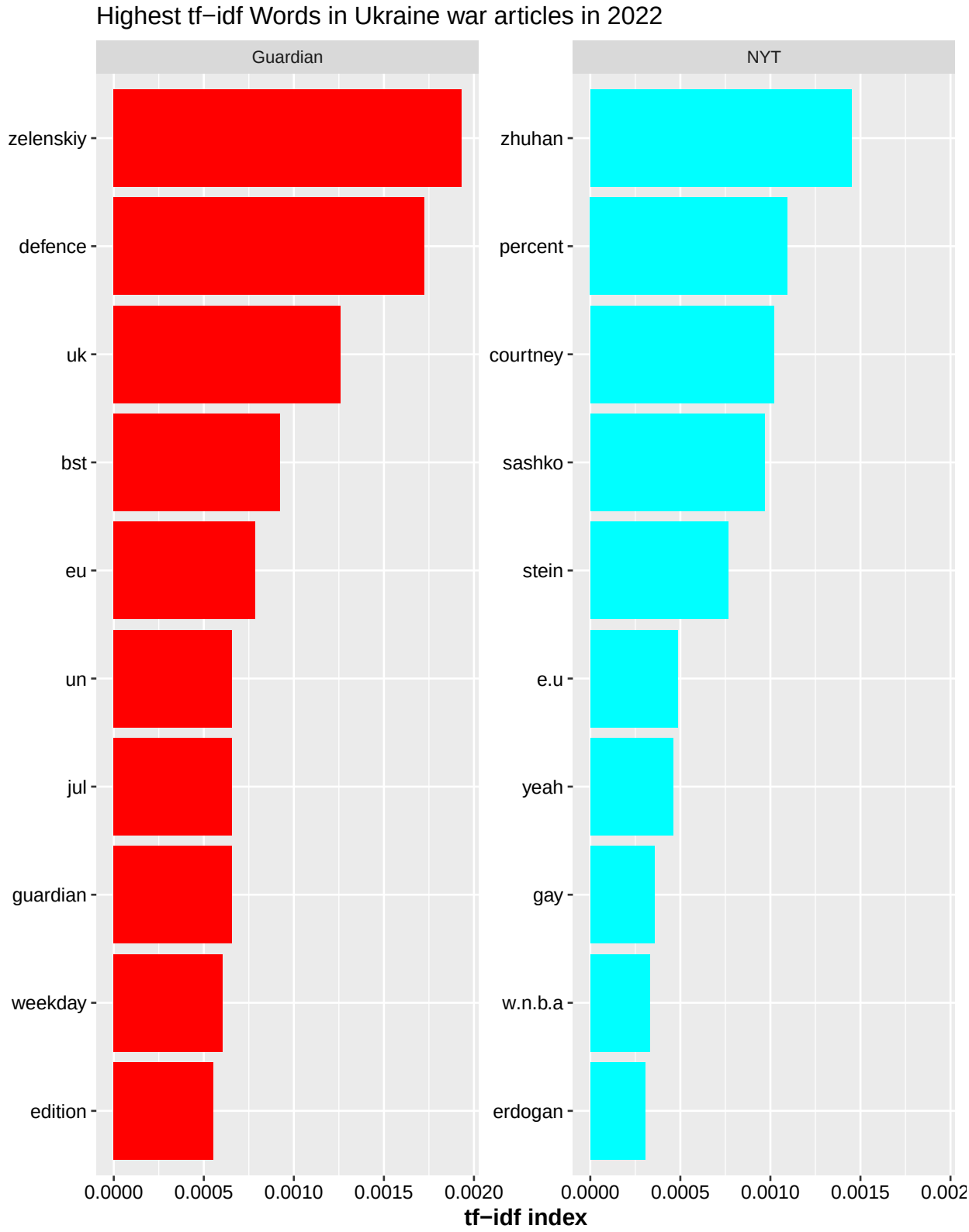
Observations:

- The graph highlights words with the highest TF-IDF indexes for each journal.
- These words are indicative of the distinctiveness of each journal's coverage.

TF-TDF Analysis

```
#> # A tibble: 10 x 6
#>    location word      n     tf   idf tf_idf
#>    <chr>    <chr> <int>  <dbl> <dbl>  <dbl>
#>  1 Plymouth A         9 0.346  0.693  0.240
#>  2 Plymouth B         5 0.192  0.693  0.133
#>  3 Penzance C         7 0.438  0      0
#>  4 Penzance D         3 0.188  0      0
#>  5 Penzance F         2 0.125  0      0
#>  6 Penzance G         4 0.25   0      0
#>  7 Plymouth C         5 0.192  0      0
#>  8 Plymouth D         3 0.115  0      0
#>  9 Plymouth F         3 0.115  0      0
#> 10 Plymouth G         1 0.0385 0      0
```

TF-TDF Analysis



Zipf's Law and Linear Regression

We explored Zipf's Law by plotting the term frequency against word rank on a log-log scale. Linear regression was applied to understand the relationship between the two variables. The linear regression model applied to the log-log transformed data aims to describe this relationship. The black line in the plot represents the linear regression line fitted to the data points from both journals. The closeness of the data points to the regression line indicates how well Zipf's law describes the distribution of word frequencies in the articles.Interpreting the regression results involves examining the slope of the regression line. If the slope is close to -1, it suggests that the word frequency distribution follows Zipf's law closely. If the slope deviates significantly from -1, it could indicate that the distribution of word frequencies does not follow Zipf's law exactly, or that there are other factors influencing the frequency of word usage.

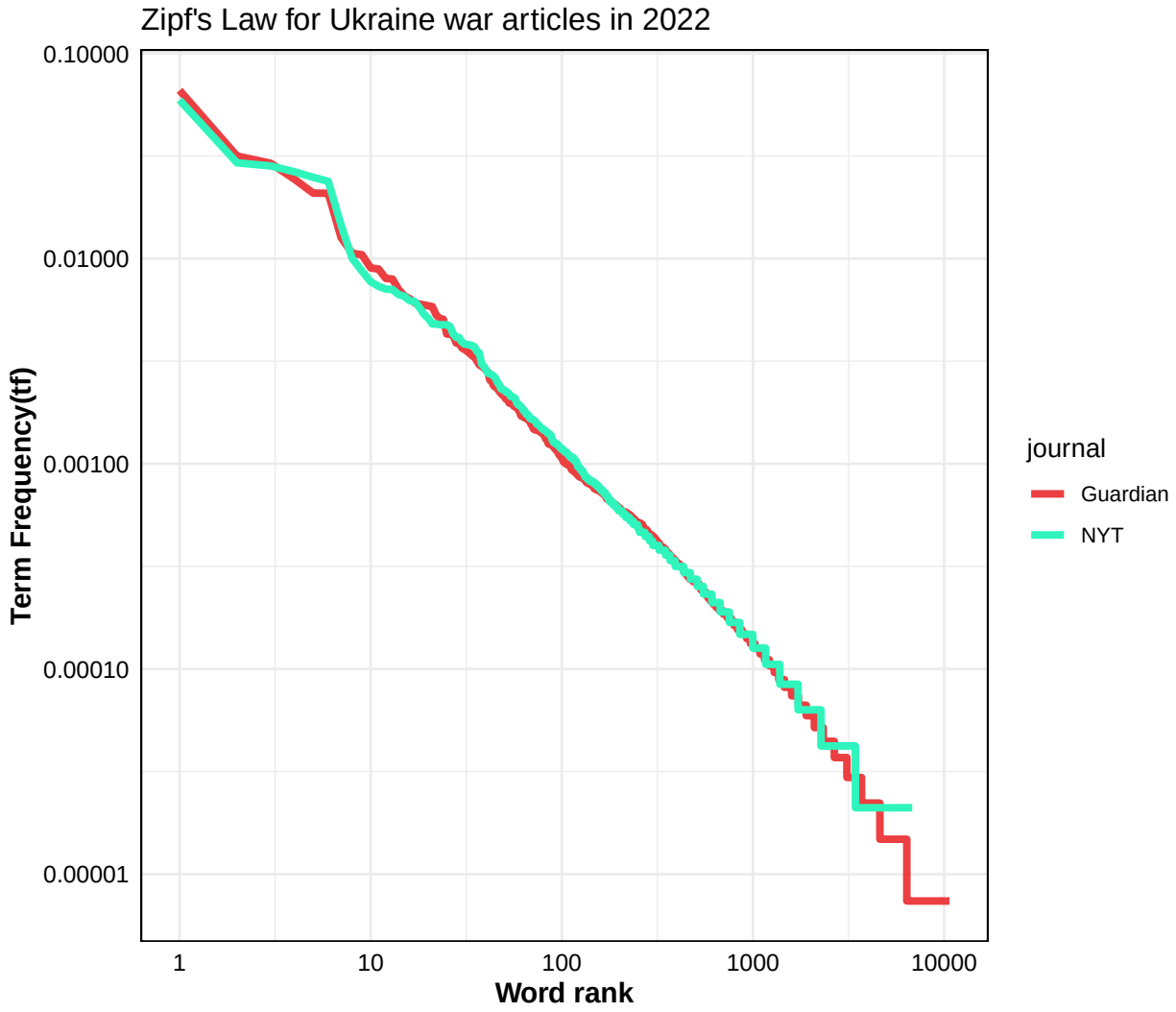
Looking at the graph, we see that the lines for both The Guardian and NYT closely follow the regression line, which suggests that the word frequencies in the Ukraine conflict articles from these journals have a distribution that is consistent with Zipf's law. However, there are some deviations, especially in the tail end of the distribution, which may be due to less frequent words or idiosyncratic uses of language in the articles.From this graph, we can conclude that the articles analyzed from both The Guardian and the NYT show a pattern of word frequency distribution that is in alignment with Zipf's law.

The coding start with tokenizing the dataset without removing stopwords, and term frequency-inverse document frequency (tf-idf) values are calculated. The top tf-idf words for each journal are plotted, showing the words with the highest tf-idf index in Ukraine war articles in 2022. The articles are tokenized, and term frequency (tf) is calculated using the `bind_tf_idf` function. The data is then arranged by journal and term frequency in descending order. The rank of each word within each journal is calculated, and Zipf's law is visualized on a log-log scale. Additionally, a linear regression line is added to the plot to describe the relationship between log(rank) and log(term frequency).

Observations:

- The log-log plot aligns with Zipf's Law, indicating a power-law distribution.
- The linear regression line helps visualize the relationship, with the negative slope supporting Zipf's Law.

Zipf's Law and Linear Regression



Zipf's Law and Linear Regression