Word Frequency Analysis of Game of Thrones Season 1

SE_42 - Data Science Basics

Student: Noah Frank

1. Dataset and Objective

This report analyses the dialogue transcripts of Season 1 of the TV series *Game of Thrones*. The data consists of subtitle-like text for 10 episodes, stored as a JSON file with individual subtitle lines and their order. After loading the data into a pandas DataFrame, I used the Python data science stack (pandas, matplotlib, NLTK) to explore word frequencies, typical vocabulary, and statistical properties of the language used in the season.

The goal is to summarize the dataset for a non-technical audience and answer a set of concrete questions using quantitative analysis and visualizations.

2. Questions

The analysis is guided by the following questions:

1. Which words occur most frequently in the Game of Thrones Season 1 transcripts?
2. Which content words dominate the vocabulary after removing common English stopwords?
3. How do the frequencies of selected keywords (e.g. “lord”, “king”, “stark”, “father”, “ser”) differ between episodes?
4. What does the distribution of word lengths in the transcripts look like?
5. Do the word frequencies in Season 1 follow Zipf’s Law?

3. Methods

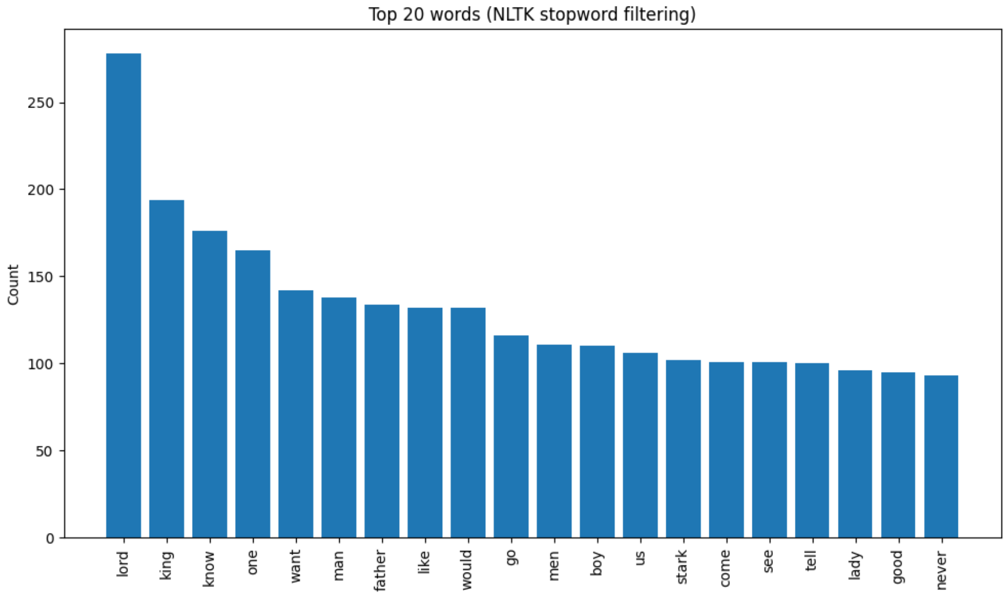
All analyses were conducted in Python using pandas and matplotlib. The subtitles were loaded from a JSON file into a table with three columns: episode identifier, subtitle line index, and text. Each episode name was normalized to a short code (S01E01-S01E10). The text was converted to lowercase and tokenised into words using a simple regular expression that keeps only alphabetic characters and apostrophes.

Word frequencies were computed over all tokens in the season. To focus on meaningful content words, I used the NLTK English stopword list to remove very common function words such as “the”, “and” or “to”. For keyword analyses per episode, I aggregated tokens per episode and counted the occurrences of selected terms. Finally, I visualized the distribution of word lengths and created a rank-frequency plot to check for Zipf’s Law.

4. Results

4.1 Overall frequent words

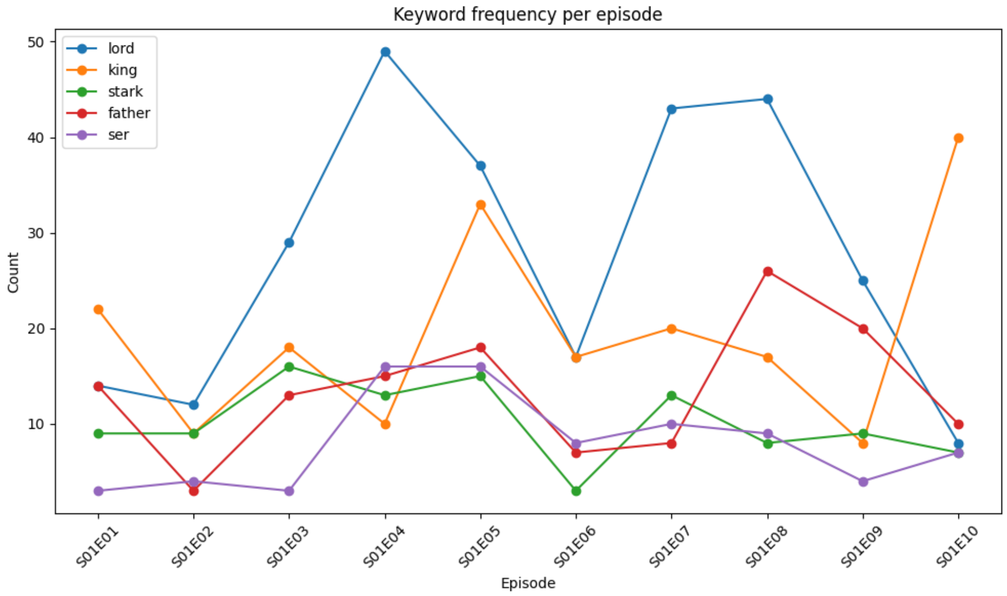
The full frequency table shows that a small set of words dominates the transcripts. Before stopword filtering, very common words like “the”, “and”, “to”, and “you” appear extremely often. After removing English stopwords with NLTK, the most frequent remaining words include “lord” (278 occurrences), “king” (194), “father” (134), “stark” (102), “lady” and “brother”. These results are summarized in a bar chart of the top 20 words after stopword filtering.



Answer to Question 1 and 2: The most frequent words in the Season 1 transcripts are function words like “the” and “and”. Once these are removed, the vocabulary is dominated by content words such as “lord”, “king”, “father”, “stark”, “lady” and “brother”, which reflect the political and family-centered themes of the series.

4.2 Keyword frequencies per episode

To compare episodes, I focused on a small set of domain-specific keywords: “lord”, “king”, “stark”, “father” and “ser”. For each episode, I counted how often these words occur and plotted their frequencies across S01E01-S01E10.

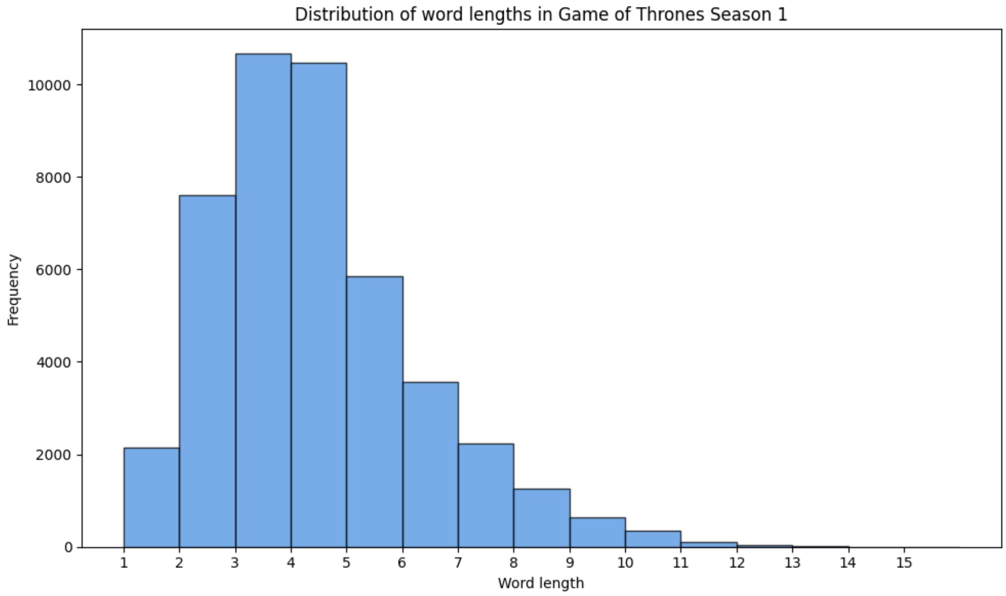


The plot shows that “lord” is consistently frequent in almost all episodes, with a notable peak in some mid-season episodes that contain many court and council scenes. “King” appears less often overall, but its frequency increases towards the later episodes, which is consistent with the focus on succession and the Iron Throne. The word “stark” is particularly common in episodes that emphasize the Stark family and their story-lines, while “father” appears in episodes with strong family and honor themes.

Answer to Question 3: The selected keywords show distinct patterns across episodes. “Lord” is prominent throughout the season, “king” becomes more frequent towards the end, and “stark” peaks in episodes centered on the Stark family. This suggests that word frequencies capture shifts in narrative focus between locations, families and political conflicts.

4.3 Word length distribution

I computed the length of every token and plotted a histogram of word lengths (Figure 3). Most words in the transcripts are relatively short, with a peak around 3-5 characters. Longer words are increasingly rare, and very long words occur only occasionally.

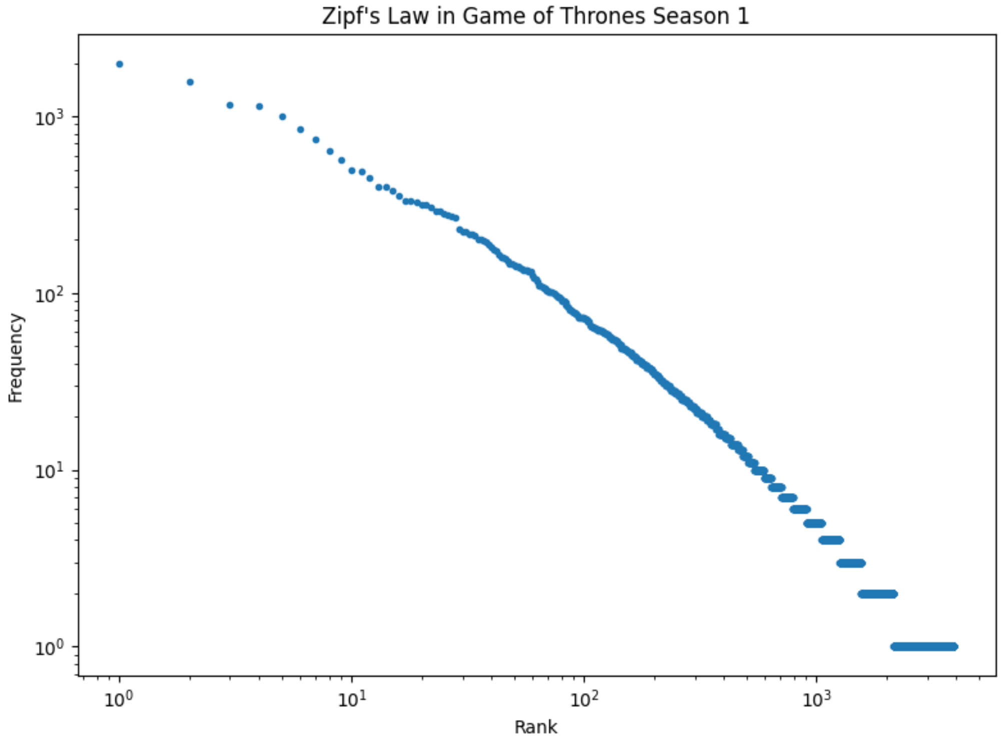


This distribution is typical for English dialogue: short function words and common verbs dominate, while long nouns and names are less frequent.

Answer to Question 4: The word length distribution is heavily skewed towards short words between 3 and 5 characters, with longer words becoming progressively rarer. This pattern is consistent with natural conversational English.

4.4 Zipf’s Law

To test Zipf’s Law, I ranked all words by their frequency and plotted frequency against rank on a double logarithmic scale.



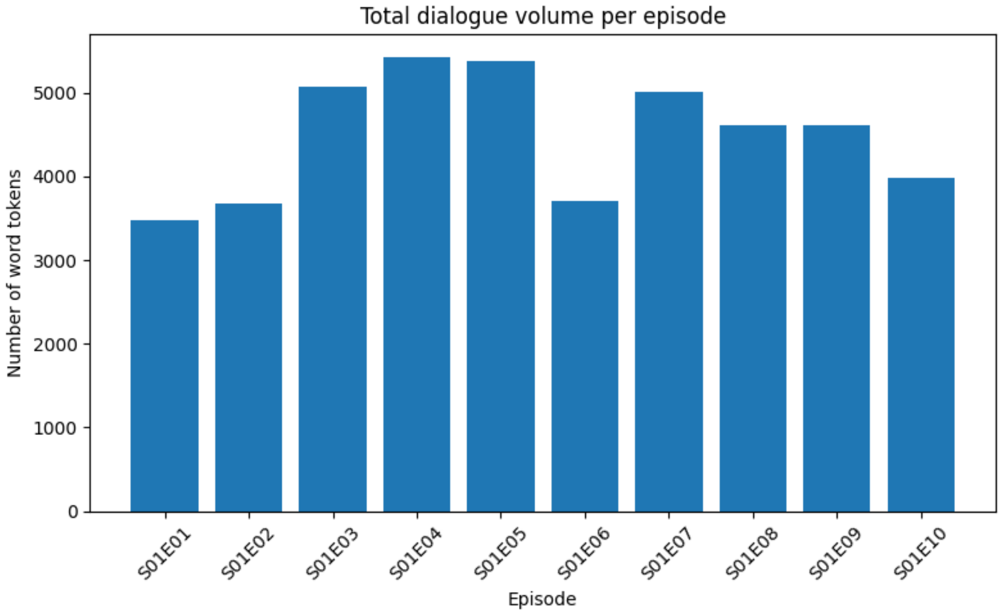
The resulting curve is close to a straight line over a wide range of ranks, with a small group of very frequent words and a long tail of rare words.

This is characteristic of Zipfian behavior and indicates that the Season 1 transcripts follow the same heavy-tailed distribution that has been observed for many natural language corpora.

Answer to Question 5: Yes, the word frequencies in Season 1 follow Zipf’s Law: the rank-frequency plot on a log-log scale approximates a straight line, showing a classic heavy-tailed distribution of word usage.

4.5 Dialogue volume per episode

As an additional view on the dataset, I summed the total number of word tokens per episode.



The episodes differ noticeably in dialogue volume: some have denser dialogue, others rely slightly more on non-verbal storytelling. However, all episodes contain several thousand word tokens, which confirms that the Season 1 transcripts form a substantial text corpus for analysis.

Additional observation: The variation in word counts per episode suggests differences in pacing and dialogue intensity but does not change the overall statistical patterns observed above.

5. Summary

This analysis shows that the subtitles of *Game of Thrones* Season 1 exhibit typical properties of natural language: a small number of very frequent words, a heavy-tailed rank-frequency distribution consistent with Zipf's Law, and a word length distribution centered around short words. After removing common English stopwords, the most frequent content words clearly reflect the core themes of the series, highlighting power structures ("lord", "king") and family relations ("father", "stark", "brother").

The keyword analysis across episodes reveals how narrative focus shifts between political centers and families, and the episode-level word counts provide a simple quantitative view on dialogue density. Overall, the dataset is well suited to demonstrate how relatively simple statistical tools and visualizations can be used to obtain interpretative insights from textual data.