[Code][Dataset][Kaggle Source]

Problem Formulation

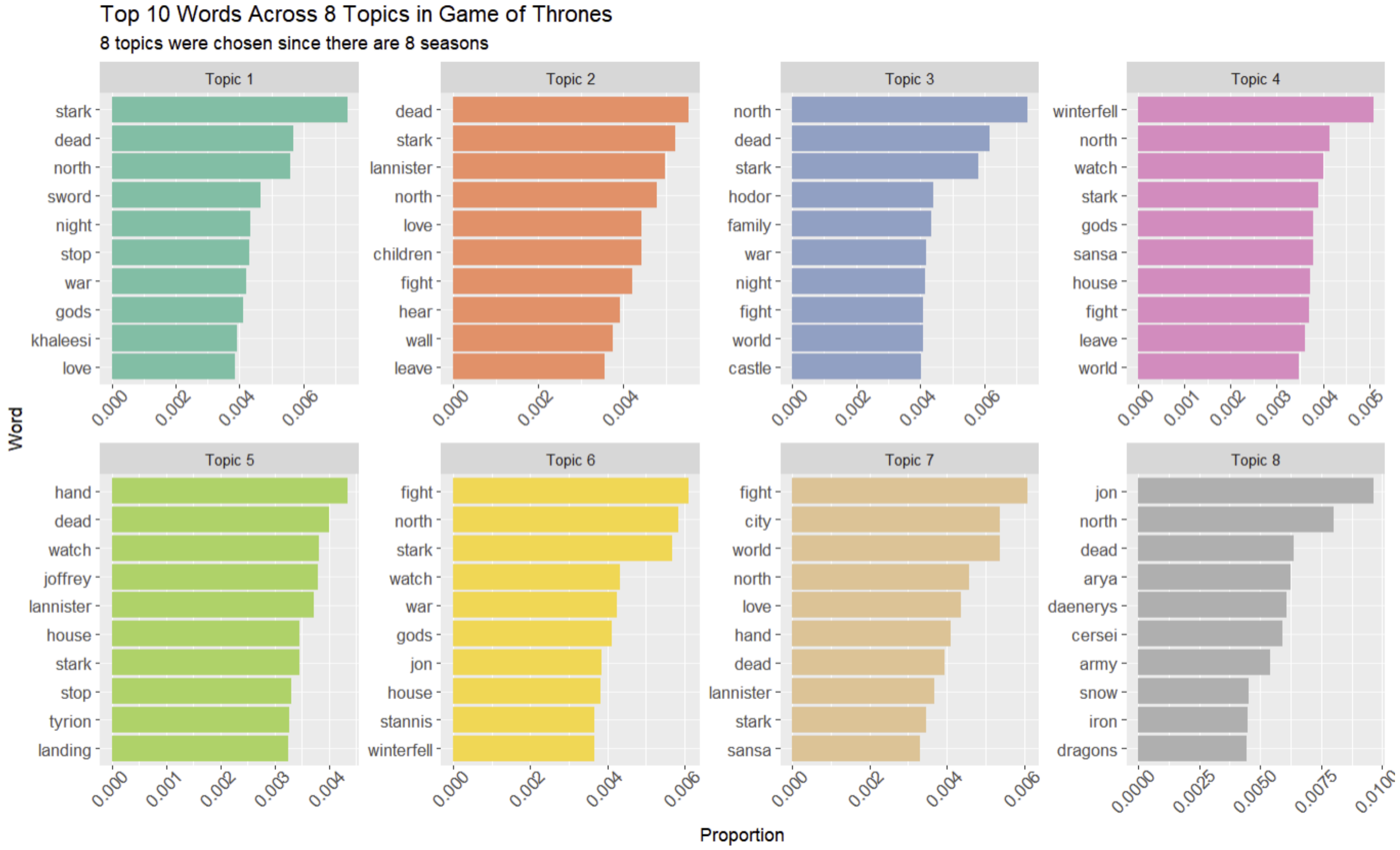
Data background and discussion

The database I used consists of the Game of Thrones script across all eight seasons. The dataset includes columns like release date, season, episode, character speaking, and sentence, where each row represents a character in the episode speaking the sentence. This dataset is interesting to use for analytical purposes such as vocabulary usage, tones, or common words. There are countless ways to analyze this script, such as how various characters have different vocabulary usage influenced by their origin or social status. My analysis focuses on creating a topic model by creating a document term matrix (DTM) with a count for each word contained within a given season and episode combination, then running it through LDA.

Framing questions

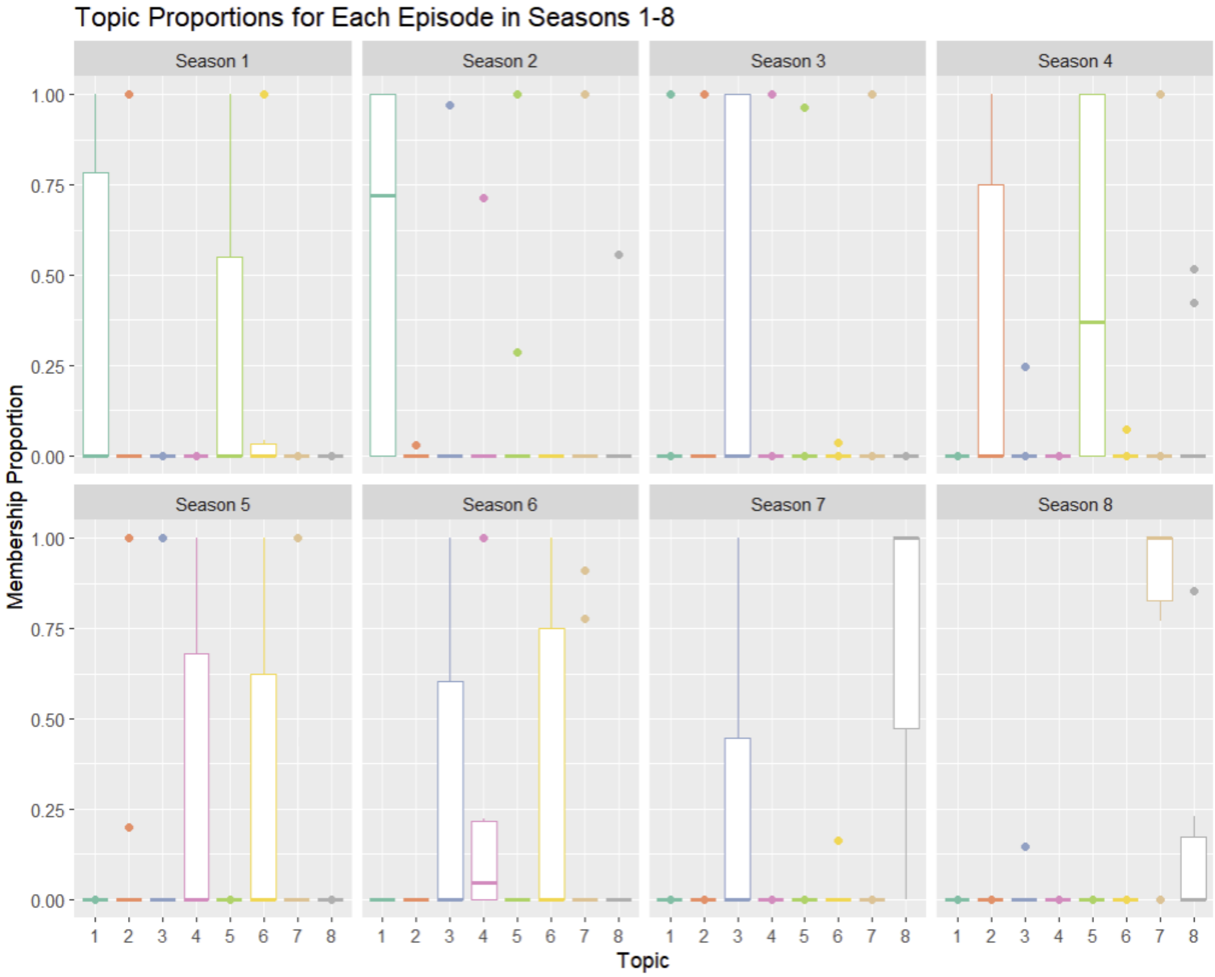
As alluded to above through the data background, I wish to create a topic model. The episodes in each season are the documents and there will be a count of the words in the episodes. The question I aim to answer is do the different seasons in Game of Thrones have unique vocabularies? If not, then what distinctions exist? I will answer this by highlighting the top words displayed in each topic fitted in the model to see similarities then looking at the topic membership of the episodes in each season.

Visualization 1 - Top Words per Topic



This visualization displays the top ten words across the topics in my topic model. It allows easy comparisons between the topics and helps determine frequently used words across topics. Eight topics were chosen because there are eight seasons in the show. One trade-off to create this visualization was I had to remove certain words in Game of Thrones that are used constantly throughout the entire show, such as King, Queen, Lady, Lord, Ser, etc., because these are all titles. One interesting finding from this visualization is some topics indicate which seasons are contained in it. For example, Topic 1 has the word “khaleesi” which I believe would represent Season 1 as the Dothraki call Daenerys Targaryen this. The key finding from this is that a lot of the topics have common words in their top ten, such as stark, war, north, and dead. The word “dead” is important in the context of Game of Thrones because the characters often say “the dead” to refer to the White Walkers. Due to this finding, we can begin to see that the topics do contain similar words and are not perfectly distinct. While they possess differences, we can begin to answer our proposed question which is that the seasons do contain similar vocabulary and we can further pursue our analysis to see what distinctions are present.

Visualization 2 - Topic Memberships per Episode in each Season



This visualization shows the topic proportion for each episode across the eight seasons. It allows us to see if a season is contained in a singular topic or not. It also allows us to compare two or more seasons to see if they belong to the same topic(s). The box plots contained within the small multiples show the distribution of episodes that belong to one or more topics. If there is no box plot, so just a colored line, then there are little to no episodes that belong to the topic. Therefore to make our comparisons, we must look at the box plots to gauge the topic membership for the episodes. One trade-off I had was adding color to thematically link the colors between my two static visualizations. Originally I had a black outline, but I wanted to create a connection between color and topic across the visualizations while not losing comparative features. The key finding in this visualization is that only two seasons primarily have episodes belonging to one topic, those being Season 2 and Season 3 belonging to Topic 1 and Topic 3 respectively. This helps answer my follow up question which is there are some distinctions between seasons because some have non-overlapping topic membership. However, we can note that some seasons, like Season 1, have membership in Topic 1 similar to Season 2, leading to similar vocabulary between those seasons. While there are clear distinctions between which topic(s) the episodes in the eight seasons belong to, many seasons have similar topic memberships. Through this and the prior visualization, we can conclude that Game of Thrones has similar vocabularies used across the eight seasons and, while some distinctions exist, there are few differences.