(Short) Problem Set 7

The seventh and final problem set requires you to solve two small questions on text analysis and mapping.

Grading will reflect your performance on both the coding as well as the documentation and interpretation tasks – however, there is no fixed grading scheme between the two categories. Ideally you **submit an R markdown document** as discussed in the lecture.

This problem set is **due on July 31st by 12.00** through the wuecampus submission functionality.

Cooperation with other groups is not permitted and will lead to severe credit deductions.

1. Game of Thrones has emerged as an immensely popular TV series building on top of a powerful book series by George R. R. Martin. Here we are interested in analyzing the language across the different seasons and episodes. In absence of a readily available book corpus we will rely on subtitles extracted from the TV series. You find subtitle files for each episode in the provided zip archive. Note that the files are not curated and may slightly differ in their style.

    a. Read in the files and create a tidytext data.frame across all files while retaining season and episode information. [Hint: scan(f, what = "character", quiet = T, quote = "") is a good start for parsing the files] Apply stop words and subsequently identify the top 10 words per season. Compare the results and try to eliminate wrongly parsed information.
    b. Perform a word-level sentiment analysis with both the bing and the AFINN lexicons – watch out that the AFINN lexicon uses numeric scores instead of binary classifications. Provide plots of the evolution of aggregate episode sentiment across each season for the two scoring systems and compare the results. Use insider knowledge or a little lookup on IMDB to verify two particular noteworthy patterns.
    c. Calculate inverse term document frequency (with seasons serving as documents) to identify terms that are particularly identifying for some seasons. Visualize and discuss your results.

2. The Zeit magazine features a special category with funny maps of Germany. A case in point is this one: http://www.zeit.de/zeit-magazin/2016/17/rtl-shows-der-bachelor-bachelorette-kandidaten-deutschlandkarte

   Using the information from https://de.wikipedia.org/wiki/Der_Bachelor#Die_Bachelorette replicate the map in an interactive fashion using google maps location lookups and the leaflet package.

   Bonus task: Include some interactivity with respect to filtering the series (Bachelor vs. Bachelorette) or the season.