STAT7016 Final Project Proposal

Admittedly, the TV series *Game of Thrones* has been a great success since released. Lots of data lovers took an insight of this great story and made a ton of nicely designed visualizations. For example, a character relation network, a detailed mortality report, or even the color palette for each episode. But really, not many people decide to investigate further into the origin of all those greatness, *A Song of Ice and Fire*, written by our beloved George R. R. Martin. As we know, one of the reason people love to read such stories is because its magnificent view, not chronically, but geologically as well. Consequently, thousands of characters emerged on the continent of Westeros.

- What are the main issues or problems to be addressed?
  - So one might wonder, how long can these character lives, or crucially speaking, lasts? As we all know, GRRM is famous for being a killer in his novel.
- Which data set are you using?
  - So I want to scrape the awoiaf Wiki, which is a fan-made community for the novel series ASOIAF, for the complete list of characters appeared in the story so far. Luckily, a good samaritan on Kaggle uploaded his own cleaned data set for the books. (Thanks to GRRM again, his infamous procrastination delayed the release of the six volume, so that this data set is not out-of-date.)
- What variables in the data set will you use?
  - In the file `character-deaths.csv`, there are 13 variables for each row of entry. We only need the following **some** of them:
    - `Name`. Of course each character needs a name, especially in this case, we don't have person id.
    - `Death Year`. Not sure if I should keep variable. I'll think about it later.
    - `Book of Death`, `Death Chapter`, `Book Intro Chapter`. These three variables generally marked the "life span" of a character, which is the core of our problem.
    - `Gender`. During the war time, male tend to have higher mortality rate than females.
    - `Nobility`. The variable marks the social class of a character.
    - `GoT`, `CoK`, `SoS`, `FfC`, `DwD`, the five sets of initials of the books. I think those variables are necessary for alive characters as they can reflect that if he or she is a recurring character or just a temporary one, i.e. the importance of such a character.
- What are your initial thoughts on appropriate models/distributions?
  - So far, the only thing has come to my mind is that it is a survival analysis with Weibull distribution.
- What questions and/or concerns do you have about the project?
  - What kind of writing style is preferred? As formal as an essay, or can it be somewhat casual like a popular science report?
  - Is there any space for naïve bayesian classifier, although it might be covered in this course? If yes, how should we implement it in this problem?