An Analysis of Ice and Fire

Pushkar Godbole, Paritosh Gote, Vinit Deodhar

Team: Brotherhood Without Banners

Report

for

CSE 6242: Data & Visual Analytics

Apr 24, 2015

1 Introduction

A Song of Ice and Fire (ASOIAF) is a is a series of epic fantasy novels written by novelist George R. R. Martin (GRRM), having sold more than 24 million copies in North America alone, as of September 2013 [1]. It is famously known for its television series adaptation *A Game of Thrones* averaged at over 18.6 million viewers by its latest season 4 [2]. The series follows the lives of a plethora of multi-dimensional characters in a medieval-ish setting in the fictional kingdom of Westeros, mainly centering on the story-line of a dynastic tryst among several families for control of the throne, and hence the name.

The universe of the series has been described in significant detail in the books, geographically, culturally and socially. The population of Westeros is primarily divided into seven houses spanning over 2000 named characters described in the book series in varying detail. With its huge and dedicated fanbase, the series has garnered an expansive online database that has yet to be explored from a data analytics standpoint.

On these lines, the aim of this project has been to accomplish a detailed statistical analysis of the data, starting from data extraction to evaluation, visualization and classification. Such statistical data and predictions about the series would be of great interest to all the fans of the books and television series.

2 Problem Description

The objectives of this project can be broadly classified into three parts in that order:

- **Feature extraction:** Although structured datasets of ASOIAF characters with various features are available online, these are limited only to basic features of major characters in the series. The aim of the project is therefore to collect more detailed data about more characters from un-formatted data sources, the primary source being the ASOIAF wiki [3].
- **Analysis & Visualization:** The next step is to analyze and visualize the data with the aim to draw objective conclusions regarding characters. The major part of this step comprises clustering of characters based on a group of features (such as age, affiliation etc) and rendering visual and statistical results to correlate them.
- **Classification and Prediction:** With basic statistical inferences drawn, these parameters are used to classify the characters based on their features. One classification of particular interest is the prediction of the risk of imminent death for the presently living characters. This classifier would be cross-validated to evaluate its performance using existing data (of dead characters).

3 Literature Survey

Due to the limited resources available on the statistical analysis of ASOIAF, we primarily concentrate on past research related to problems similar to the ones we anticipate such as, data scraping, Natural Language Processing, Complex Data Visualization and Predictive classification.

The work by Richard Vale [4] on prediction of events in The Winds of Winter (ASOIAF 6) using a Bayesian inference model is the closest to our objective. In this work, based on occurrence of events, posterior probabilities of future events are predicted. The data used for the same contains the number of Point Of View (POV) chapters per character per book in the series. [5] predict eruptions of old faithful geyser using logistical regression and predict eruptions within an accuracy of 30 minutes.

For generating the character features, unstructured data in text format had to be converted into a structured format before performing prediction. Perkins et al. [6] provide tools for data extraction using various Natural Language Processing techniques. In the data set about characters in ASOIAF, different features can contribute with a different factors of influence in prediction of events. To that end, Bishop et al. [7] describe techniques for identifying

most important features and reducing the dimensionality of data using Principal Component analysis.

A work on Game of Thrones Visualization by Jerome Cukier [8] creates a classification of characters in various classes based on a timeline of book chapters. The timeline helps to effectively interpret the progress of characters over time. However, the clustering in this case is non uniform with Houses Stark, Lannister, Antagonists, Protagonists used as groups.

Many current works on event prediction either do not present data and results using a visualization or use a static non interactive visualization that makes interpretation of results challenging. The ones with visualizations lack detailed statistical and predictive analysis. To that end, we wish to perform a holistic study and presentation of the results from a data analytics standpoint.

4 Proposed method

The approach used for the project can be divided majorly into three categories:

4.1 Data collection:

We extracted features from the following two data sources

1. Structured data containing information of 2016 characters. It is available as a csv file and the size of data is 200 KB
2. Unstructured data is available as wiki pages [3]. Each html page contains data of one character. The number of html pages in the data set were 2026 and the total size of data set is 200 MB

The features extracted can be classified as objective features that are available directly in data and subjective features that could be inferred from the data set

1. **Objective feature extraction:** The objective feature extraction is performed to extract attributes including *date of birth*, *date of death*, *gender*, *allegiance*, *culture*, *Categories*, *Alive status*, *POV* and *chapter refrences* over time, for each character. These features are extracted from the wiki pages using a html parser implemented in python. Once the features for a character are extracted from the html pages, they are combined with the features extracted from the csv dataset. The features are combined by matching character names from both the dataset. Small variations in character names can result in the same charcter named differently in different datasets. eg *Blackwood* is named as *blackwood* in one of data set and *Lord Blackwood* in other. Some names contain aliases used as a part of name. We normalize the names by removing common prefixes such as *(*king), *queen*, *lord*, *Sir* as well as removing any aliases used as a part of name. After normalizing the names to a normal form, we match the names to combine objective features from both the data sets.
2. **Subjective feature extraction:** The subjective features include the physical appearances and abstract personality characteristics of characters. The extraction module scrapes the ASOIAF wiki [3] pages for each character and uses natural language processing techniques to generate a character map. We have used pattern.en [9] along with NLTK [10] for effectively performing web scraping and natural language processing. A parsetree is generated from the input data using the inbuilt statistical parser and english grammar library, which is further used to deduce which features are appropriate. The current model handles negative inferences and transferred epithets on predefined universal sets. The count of occurance of each feature is used as a weight for that particular feature.

4.2 Data Analysis:

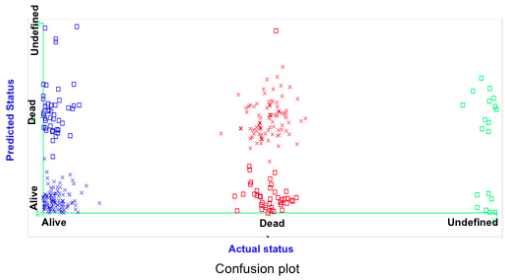
The analysis of the collected data primarily involves two tasks:

- **Clustering and Statistics:** With the availability of modular data, the characters have been clustered based on various attributes such as gender, allegiance, categories, personality. With this clustering, we generate statistics such as, male female ratios, occupation and status demographics, etc.
- **Visualization:** The visualizations generated are of two types:
  - *Static:* These include static plots and graphs of the above generated statistics. In particular, we generate bar plots for age distribution, gender, current dead-alive status, occupation, alligence and region demographic.
  - *Dynamic:* These are interactive visualizations that aggregate and render a broader scope of information. In particular, we have generated two visualizations. One that gives a micro scale view of each character while another that gives a view of different communities.
    * Character bio: This visualization lists the objective and subjective features of each character based on an autocomplete menu. It also renders a histogram of the *important* chapter references for that character by binning the references from the AWOIAF into groups of 20 chapters.
    * Character clustering: This visualization renders the clustering of characters based on five attributes: House, Gender, Social status, Occupation, Region. To facilitate fast rendering, only the characters whose dead/alive status is known are included. The dead characters are marked in red and alive ones in green.

4.3 Classification & Prediction:

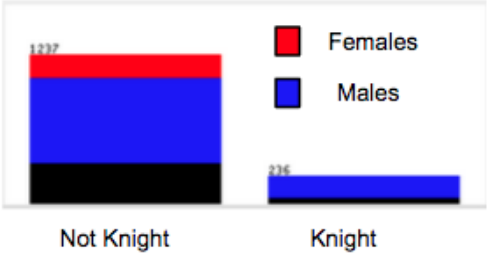
Once the objective and subjective features are generated, they are used to classify characters. Weka[11] is used to experiment with various classifiers and optimize the classification task. Various classifiers such as Decision Tree, Decision Table, Random Forest, Naive Bayes have been evaluated to predict deaths. Amongst these, BayesNet is empirically observed to be the best in terms of precision, recall and accuracy. Primary classification is performed to classify characters as "dead" or "alive". Classification to predict gender is also performed as a secondary classification. All these tasks have been performed with 10-fold cross-validation.

- Classification as dead or alive: The area of the confusion matrix where predicted value is "dead" and actual value is "alive" can be viewed as the characters that are presently alive but predicted by the classifier to be killed in the future books.



Confusion plot

- Classification as male of female: Gender based classification is performed with a higher accuracy of 85.2%. The high accuracy of this classification could be attributed to high gender disparity among occupations and subjective features.



As the above example displays that almost all knights are males. Similarly many other professions are male dominated and only few others are female dominated.

5 Experiments & Evaluation

5.1 Data Extraction

The data scraping from the ASOIAF wiki and NLP based personality and appearance feature extraction gave a holistic picture of the character. The following two profiles of Cersei Lannister and Eddard Stark reflect the same.


```
{
    "Name" : "Cersei_Lannister",
    "Appearance":   {
        "graceful figure" : 1,
        "slender figure" : 1,
        "brilliant eyes" : 1,
        "green eyes" : 1,
        "fair skin" : 1,
        "blonde hair" : 1
    },
    "Characteristics": {
        "Beautiful" : 1,
        "Cunning" : 1,
        "Incompetent" : 1,
        "Astute" : 1,
        "Complex" : 1,
        "Graceful" : 1,
        "Willful" : 1,
        "Ambitious" : 1
    },
    "Allegiance": {
        "House Lannister" : 1
    },
    "Culture": {
        "Westerlands" : 1
    },
    "Age" : {
        "Born" : 266
        "Alive" : 1
    }
    "ref" : [67, 162, 185, 175, 45, 8, 16, 168, 30,
    39, 49, 65, 69, 73, 99, 111, 124, 130, 132, 135,
    183, 224, 230, 234, 141, 145, 146, 155, 166, 330,
    181, 346, 357, 122, 90, 314, 154, 171, 232, 349]
    "categories" : ["House Lannister",
    "House Baratheon", "Regents", "Noblewomen",
    "Characters from the Westerlands",
    "Members of the small council", "POV characters"]
}
```


Cersei Lannister Profile


```
{
    "Name" : "Eddard_Stark",
    "Appearance":   {
        "long face" : 1,
        "solemn face" : 1,
        "dark hair" : 1,
        "grey eyes" : 1,
        "dark eyes" : 1,
        "closely-trimmed beard" : 1
    },
    "Characteristics": {
        "Honorable" : 2,
        "Kind" : 1,
        "Good" : 1,
        "Protective" : 1,
        "Disdainful" : 1,
        "Cold" : 1
    },
    "Allegiance": {
        "House Stark" : 1
    },
    "Culture": {
        "Northmen" : 1
    },
    "Age" : {
        "Born" : 263,
        "Dead" : 299,
        "Alive" : 0
    }
    "ref : [ 1, 109, 97, 209, 6, 58, 188, 4, 301,
    2, 12, 200, 207, 20, 27, 33, 35, 39, 43, 44, 45,
    47, 49, 65, 67, 326, 333, 22, 30, 125, 240, 144]
    "categories" : ["POV characters", "House Stark",
    "Nobles", "Characters from the North",
    "Hands of the King", "Wardens", "Regents"]
}
```


Eddard Stark Profile

As we can see the character maps closely model the personalities and physical appearances of Cersei and Eddard. Each word is weighted according to the number of occurrences, in this case Eddard has weight 2 for “Honorable”.

Such detailed data however is not available for all of the 2016 characters. Out of these, 1495 have basic objective attributes such as gender, alligence, region etc. And only 557 of all characters have subjective personality attributes. Roughly 40% of all the characters have missing dead/alive status. This has made normalizing this dataset a challenge in itself. Below we present the statistical evaluation based on all available data.

5.2 Stats of Westeros

It may be noted that, all the statistics generated here is only for the named characters described in the book series. Since the series mainly centers around the elite classes of Westeros, the statistics does not represent the actual demographic of the populace, especially the lower majority.

Based on the data of the 1495 characters: Male : Female = 5:1

Also, majority of occupations and positions are held by males with most females assuming secondary roles. The Land of the seven isn’t big on gender equality.

The following graphs plot the occupation and social status demographic of characters partitioned based on the dead/alive status.

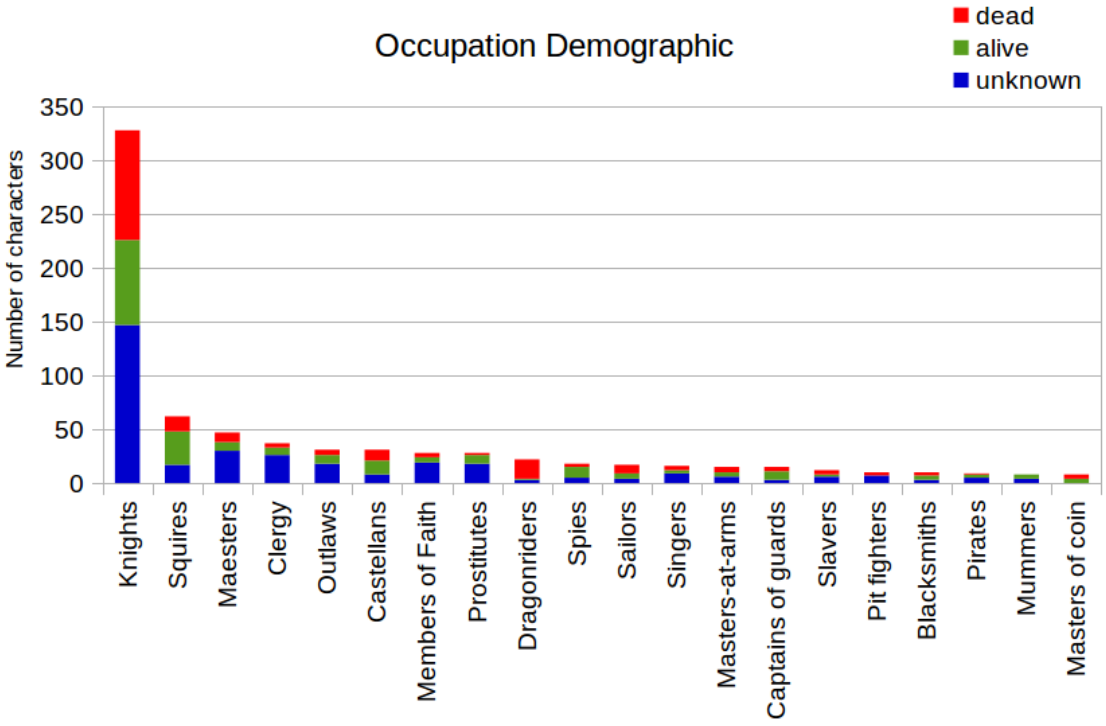


Figure 1: Occupation Demographic

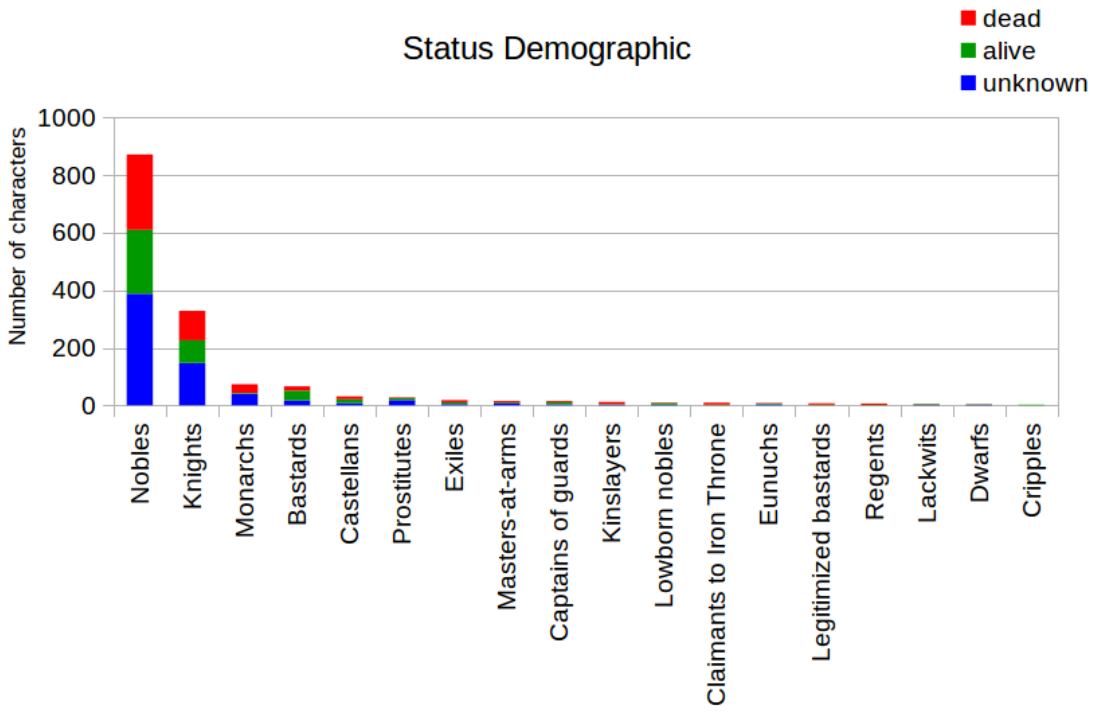


Figure 2: Status Demographic

Occupation	% Dead
Dragonriders	100
Knights	56.35
Masters-at-arms	55.56
Maesters	52.94
Members of Faith	44.44
Castellans	43.48
Outlaws	38.46

Status	% Dead
Monarchs	90.91
Knights	56.35
Nobles	54.13
Exiles	50
Castellans	43.48
Bastards	30.61
Prostitutes	20

Quite as expected, Dragon-riding turns out to be the riskiest profession and being a Monarch sounds like the bell of death. As an interesting caveat, being an outlaw is safer than being a Maester. Who knew education could be a bane!

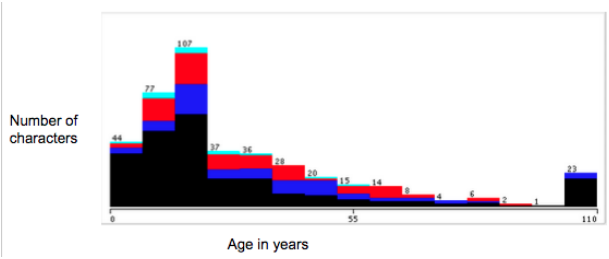


Figure 3: Age Distribution

The above histogram shows the age distribution in the given dataset. The mean age over all characters comes out to be 31 years while the standard deviation stands at 27 years. Maester Aemon of the Nights Watch is the oldest person alive at an age of 102 years. While Aegon (son of Rhaegar) Targaryen died the youngest at an age of 1.

5.3 Visualizations

A live visualization of the character bio, listing all attributes and plotting the histogram of AWOIAF references for a character as time progresses has been developed and is represented in the figure below for the character "Jon Snow".

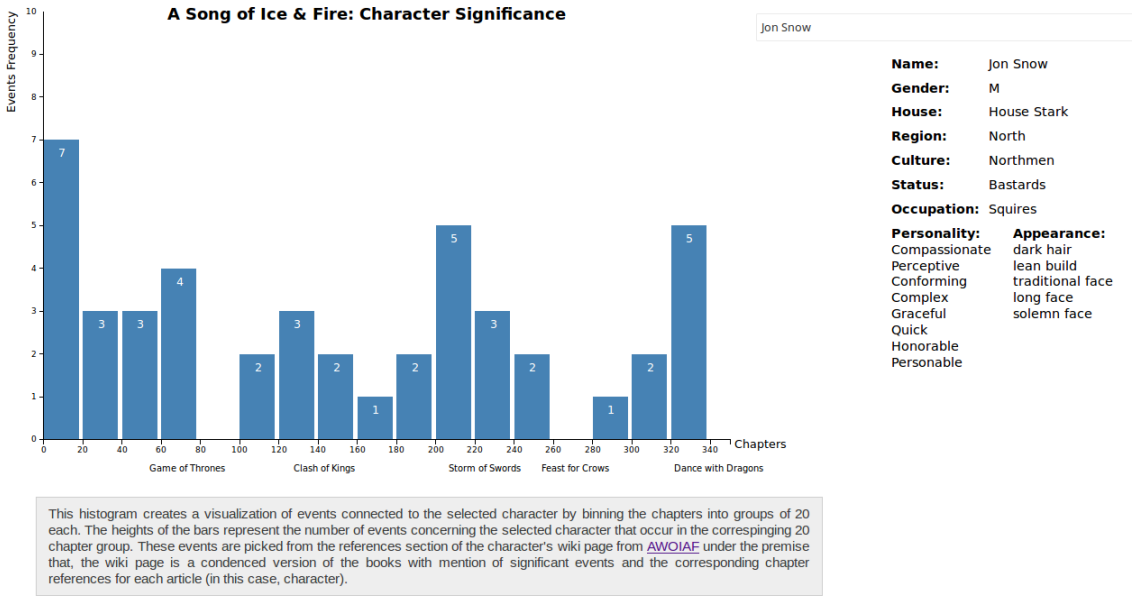


Figure 4: Character Significance: Jon Snow

This visualization can be accessed by following the link at:

https://www.prism.gatech.edu/ pgodbole7/GoT/character_bio.html

The following image shows a snapshot of the clustering visualization that groups characters based on attributes. In particular, the characters can be clustered based on Region, House, Gender, Occupation and Status. Since it is a live visualization, the transition from one clustering to another happens dynamically.

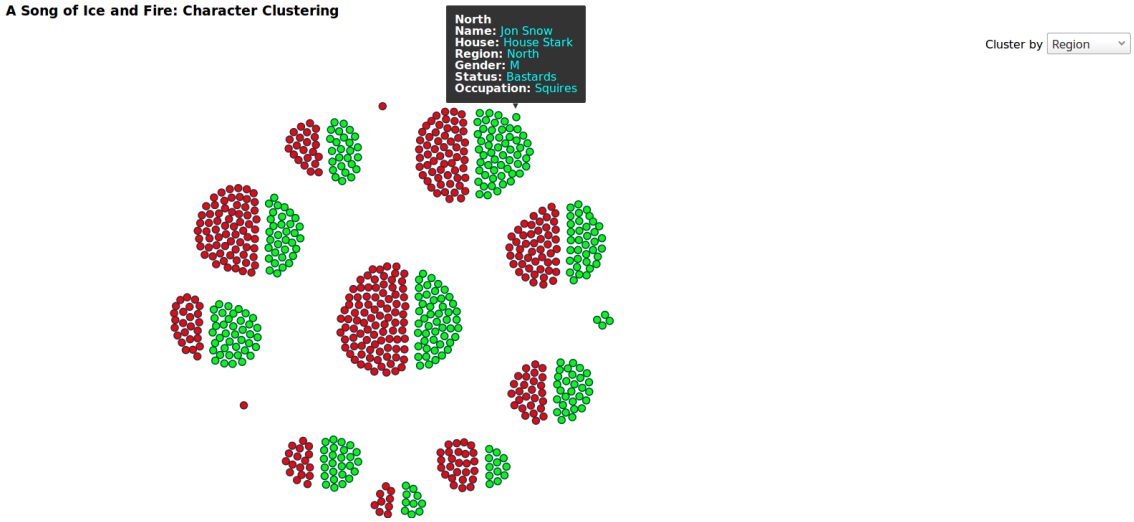


Figure 5: Character Clustering: Region

This visualization can be accessed by following the link at:
https://www.prism.gatech.edu/ pgodbole7/GoT/clustering.html

5.4 Classification

The results for the classification to predict the dead/alive status can be seen in the figure below. It can be seen from the results that adding subjective features to the feature space increased the classification accuracy.

Dead/Alive	Precision	Recall	Accuracy
Objective	61.6%	73.7%	54.7%
Object+Subjective	70.7%	72%	65.8%

Table 1: Character Classification: Dead/Alive

The BayesNets classifier predicts 42 of the characters that are presently alive in the story-line to die, including Roose Bolton, Rickon Stark, Cersei Lannister, Jeyne Poole, Samwell Tarley, Gilly and Osha.

The results for the classification to predict gender can be seen in the figure below. In the gender classification, objective features weighed in more in the classification task.

Gender	Precision	Recall	Accuracy
Objective	93.1%	88.8%	85.2%
Object+Subjective	91.2%	87.9%	83.6%

Table 2: Character Classification: Gender

6 Discussion & Conclusions

The scraping and parsing of data from the ASOIAF wiki has been challenging particularly due to the non-uniform nature of data for characters. This as necessitated the normalization of the data to fit into uniform features to be fed to the classifier. The extraction of personality traits of characters from the textual data has been another major challenge owing to the complex structure of compounding and contradicting statements.

We created a visualization to view number of chapter references over time. This visualization uncovers interesting patterns. The chapter references of a character are higher than the average before and after occurrence of significant events such as marriage and death in the life of a character. Thus an increase in chapter references could be an indication of an important event about to occur. We also constructed a visualization of number of dead and alive characters grouped according to various categories such as occupation, gender and allegiance. This visualization helped us to observe interesting facts such as some occupations or status like Dragonrider, monarch have more death probability (over 90%) compared to others.

We found Weka to be effective in carrying out the classification. After extensive testing we found that BayesNet to be most accurate classifier giving precision and recall values of 71% and 72% and an accuracy of 65.8%. On the other hand, classifying characters to predict genders turns out great with an accuracy of 85% mainly due to the stark differences between male and female characters (particularly in their occupations).

Since this is the first time such a detailed analysis of the ASOIAF data would be performed, this work would be of great interest to all the fans of the books and television series. Additionally, we believe that the dataset created by us is the most detailed in the realm of ASOIAF. This structured data can be considered as a valuable ASOIAF resource in itself, which may be used in further research by us and others.

And as always,

Valar Morghulis

References

[1] With swords and 'Game of Thrones' spurs sales for HBO suds. http://www.reuters.com/article/2013/09/20/us-emmys-gameofthrones-idusbre98j0w020130920. 2013.

[2] Game of Thrones has become more popular than 'The Sopranos'. http://www.hitfix.com/the-fien-print/game-of-thrones-has-become-more-popular-than-the-sopranos-sorta-kinda. 2014.

[3] A Song of Ice and Fire. http://awoiaf.westeros.org. 2015.

[4] Richard Vale. Bayesian prediction for the winds of winter. 2014.

[5] J. K. Raye. Using nonlinear dynamics to predict old faithful. *Math. Comput. Model.*, 41(6-7):679–687, March 2005.

[6] Jacob Perkins. *Python Text Processing with NLTK 2.0 Cookbook*. Packt Publishing, 2010.

[7] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics), chapter 12 : Continuous Latent Variables*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[8] Jerome Cukier. Game of thromes visualization. 2013.

[9] Pattern.en for Natural Language Processing and Data Mining. http://www.clips.ua.ac.be/pages/pattern-en. 2015.

[10] Natural Language Toolkit 3.0. http://www.nltk.org/. 2013.

[11] Weka. https://weka.wikispaces.com/. 2013.

Appendix

The following table represents the revised status of tasks completed and those in progress.

Task	**Members Responsible**	**Deadline**	**Status**
Data collection	Puskhar, Paritosh	week 1	Completed
Data cleaning	Paritosh, Vinit	week 1	Completed
Identifying various characteristics for feature creation	Vinit, Pushkar	week 2	Completed
Creating a character map using NLP	Pushkar, Paritosh	week 2	Completed
Statistical Analysis	Paritosh, Vinit	week 3	Completed
Experimenting various classifiers for predicting character deaths	Vinit, Pushkar	week 4	Completed
Designing visualizations from obtained results	Pushkar, Paritosh, Vinit	week 5	Completed

All three team members have done approximately equal amount of work.