Web Mining Project Proporsal

Group Members: Dürdane Türkmen, Şeyma Şen

We are planning to do analysis of the most popular top 10 series.Within the Project; publicity of the series, actors, seasons will be taken on a seperate page and the data will be processed.Each series will be compared each other.For instance;

What is the most popular season of "Game of Thronus"?

What is the highest and the lowest rated episode of "Game of Thronus"?

What is the name of the series with the highest rating ?

We will show the answers to above question in the visualization.

Web Scraping:

- Reference page :We'll be mining this site http://www.tv.com/shows/
- In intro page will include names of the series that ranked by the scores.
- Each directory will have a page of its own and publicity of series , information of cast/crew will take place in this page.
- Also this page will include seasons and will be redirected to a different page for each season.
- Every season will have data processing and visualizaton that means it will be visited multiple pages.
- Cast/crew, publicity of the series, information of episode, rating of episode will take place in item page.
- We will use scrapy and to scrape data as dynamically.The information needs to be updated when the episode is added.

Data Processing:

- The data will be stored as Json.
- According to rating scores will be visualization by using cleaning of the data.
- The seasons will be compared their own ;and then between the seasons , then between the series.

Data Visualization:

- This data will be present as web-based.Users will enable to present our findings.
- Users will be able to see the series by ranking with rating scores.
- Users will be able to see the episode by ranking with the highest rating scores.
- Users will be able to see the seasons by ranking with rating scores.
- We will use bar chart.

API Usage:

- We will use twitter and tumblr for api usage.
- It will be taken for each series of important hash tags and will be used for visualization part.
- For example, with the hashtag #arrow how many tweets or Tumblr in blogging to identify numerically.
- Also by identifying thrown tweets related index of the character, we think the characters according to the ranking popularity.