EuroSoccerViz

A Country-Based Soccer Data Visualization System

Motivation

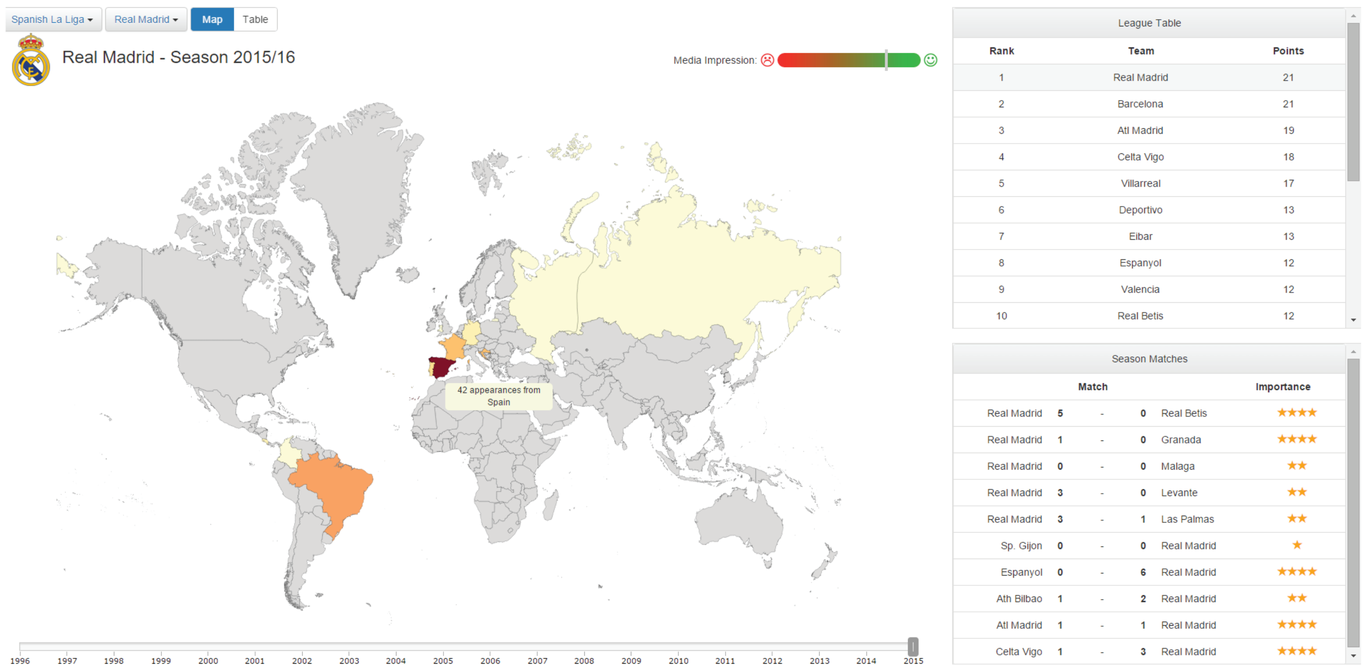
- Soccer is the most popular sport in the world
- Some under-developed countries can contribute talented players to championship teams
- Interest in tracking and visualizing the growing international presence in European soccer
- Analyze the countries that contribute most to a team's success
- Predict match results based on analysis of each country and match importance

Data

- **Sources:**
  - SoccerBase.com
    - 40 leagues from 1996 to present
    - 50,000 players and 90,000 matches
  - The Guardian Newspaper
    - 64,000 articles
- **Methods:**
  - Web Scraping with Java jsoup
  - The Guardian and Alchemy APIs
- **Size on disk**:
  - Raw data from soccerbase:150 MB
  - The Guardian articles: 290 MB
  - Processed data: 155 MB

Visualization - User Interface

- Choropleth map showing the total appearances by country with capability to switch to a table of the team's roster



- UI developed using D3 and HTML
- Sentiment analysis bar indicating media's impression of a team
- Current standings for the league and match results
- Match importance indicated by a star rating system
- Prediction of next match based on the developed model

Algorithm

- **Model:** Logistic Regression Model
  - Model generated independently for each league
- **Response Variable:** Win - Draw - Lose
- **Explanatory Variables:**
  - Contribution of a country/region -
    If # players < k, then group by geographic region
    else, select country of origin
  - Importance of the match considers:
    - Time within the season
    - Difference in the rank of teams on match day
    - Standing of each team in the current table
  - Average age of roster for current year
  - Results of last five matches

Experiments

- **Evaluation and Results:**
  - Cross validation of 10 folds
  - Compared accuracy with other Machine Learning algorithms