Football Player Data Analysis

Part 3

Oleksandr Solovei

126784

https://github.com/s126784/fcd/

universidade de aveiro

Project Evolution

Previous Parts

- Part 1: Data Collection & Initial Analysis
- Part 2: Text Analysis & Historical Data

Part 3 Goals

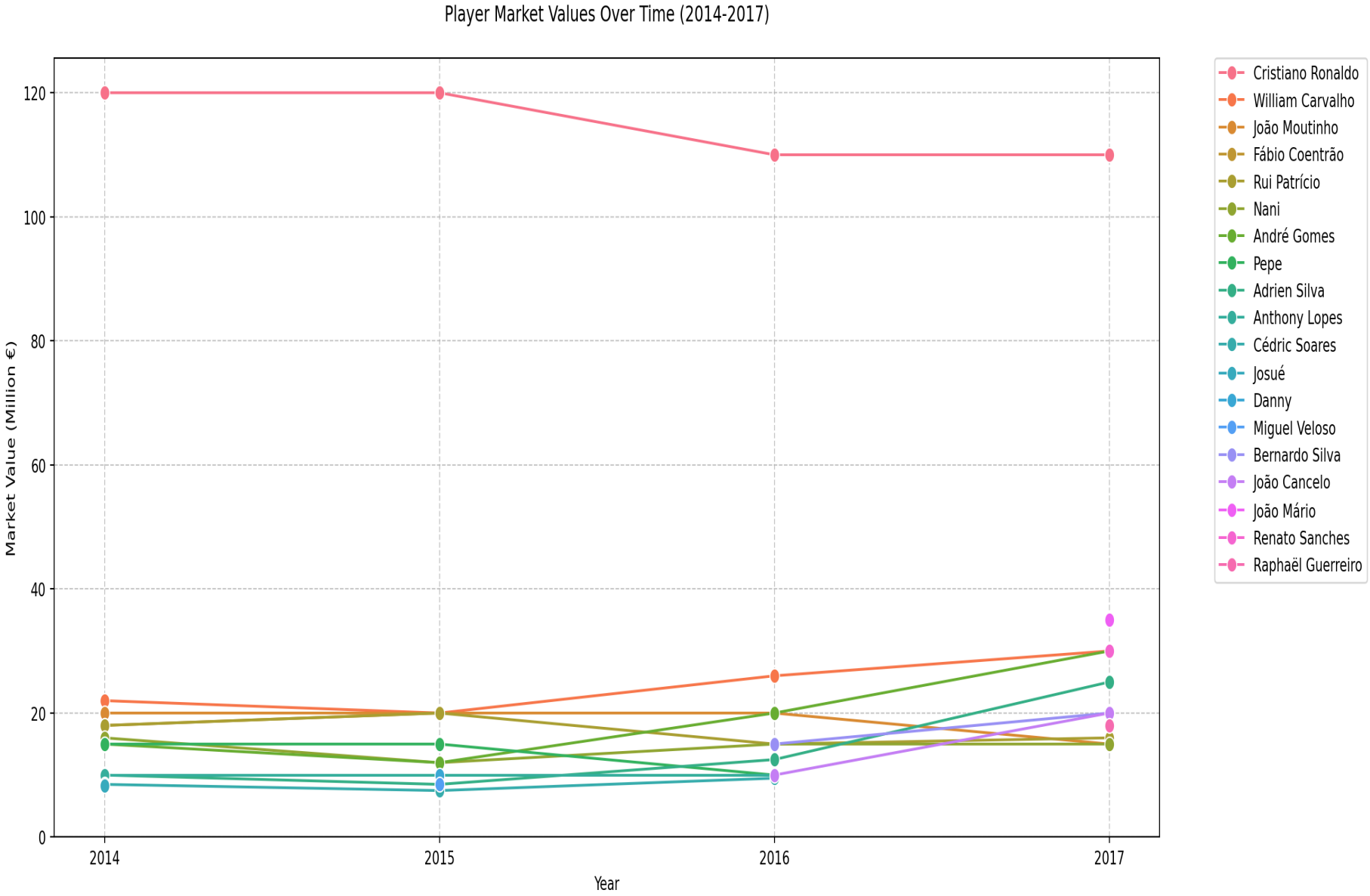
- Advanced text processing & sentiment analysis
- Time series prediction for market values
- Player clustering and network visualization
- Market value trend prediction

Dataset

```
1 years = [2014, 2015, 2016, 2017]
2 data = list(map(lambda year: pd.read_csv(f'data/portugal_{year}_plus.csv'),
3 data[0].head()
```

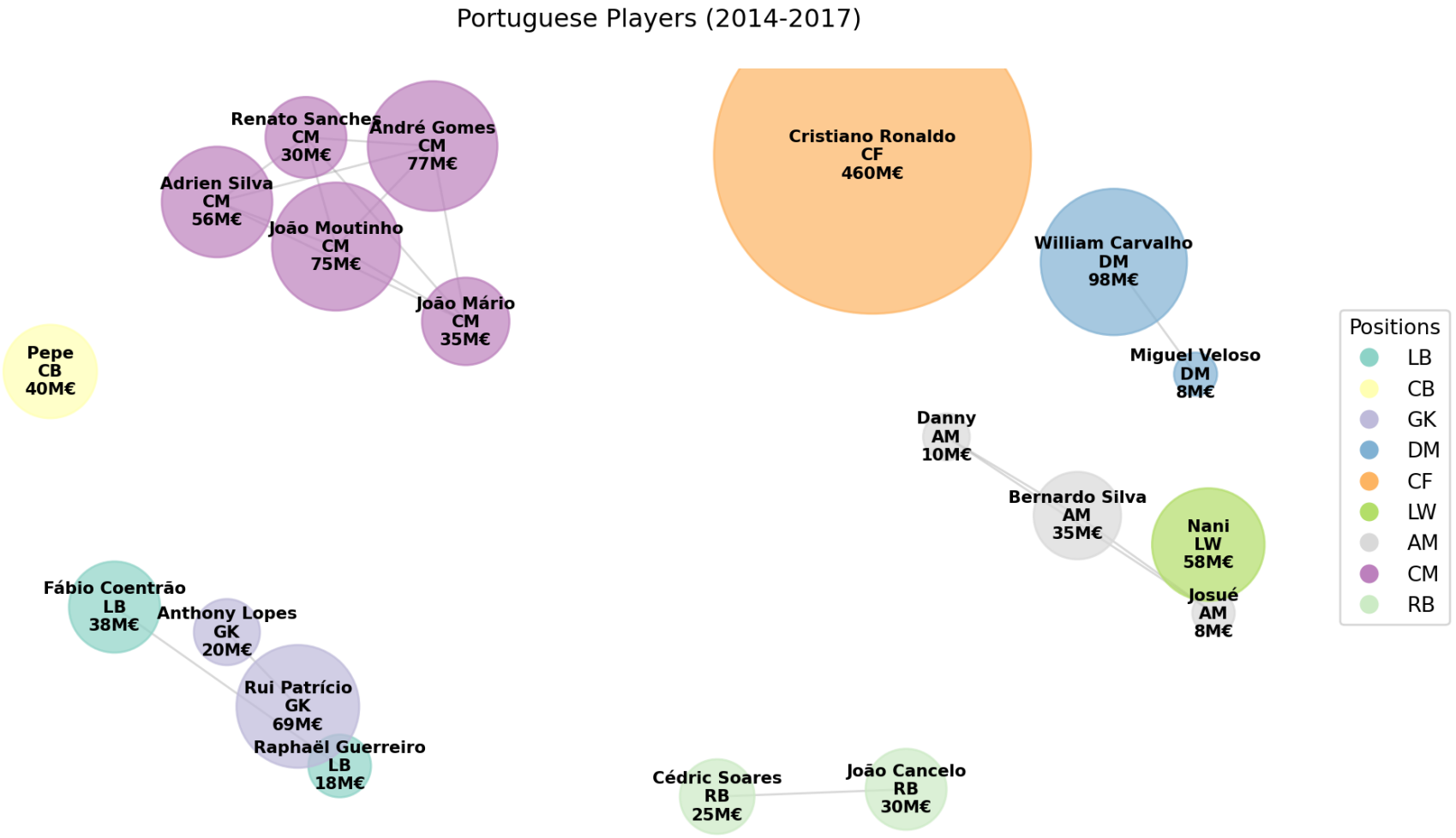
	#	Player	Age	Market value	Name	Position	search_results
0	7	Cristiano Ronaldo Centre-Forward	30.0	120000000	Cristiano Ronaldo	CF	8631215
1	6	William Carvalho Defensive Midfield	23.0	22000000	William Carvalho	DM	2809567
2	8	João Moutinho Central Midfield	28.0	20000000	João Moutinho	CM	1431291
3	5	Fábio Coentrão Left-Back	27.0	18000000	Fábio Coentrão	LB	503646

Graphical Representation (Matplotlib)



universidade de aveiro

Graphical Representation (NetworkX)



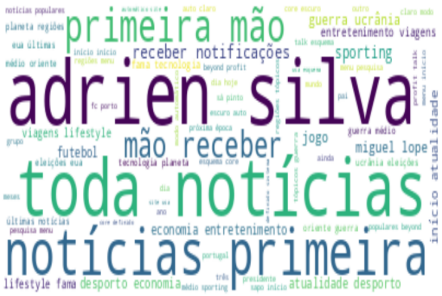
universidade de aveiro

Text Processing

```
 1 from nltk.tokenize import word_tokenize
 2 from nltk.corpus import stopwords
 3 from nltk.stem import WordNetLemmatizer
 4
 5 def advanced_tokenization(text):
 6     lemmatizer = WordNetLemmatizer()
 7     if not isinstance(text, str):
 8         return []
 9     tokens = word_tokenize(text.lower())
10     stop_words = set(stopwords.words('portuguese'))
11     # Remove non-alphabetic and stopwords
12     tokens = [lemmatizer.lemmatize(t) for t in tokens
13               if t.isalpha() and t not in stop_words]
14     return tokens
15
16 # Apply advanced tokenization to content_df
17 content_df['tokens'] = content_df['extracted_text'].apply(advanced_tokeniza
```

Word Clouds

Adrien Silva Central Midfield



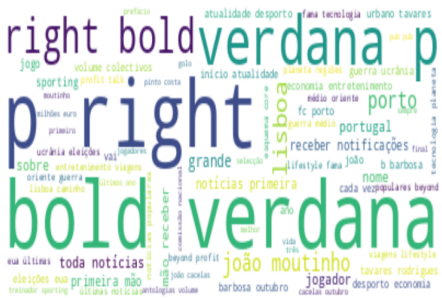
André Gomes Central Midfield

Bernardo Silva Attacking Midfield

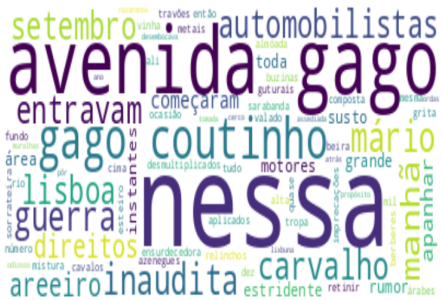
Cristiano Ronaldo Centre-Forward

Cédric Soares Right-Back

João Moutinho Central Midfield

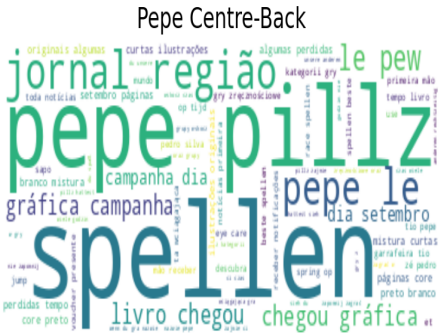


João Mário Central Midfield



Nani Left Winger

Pepe Centre-Back



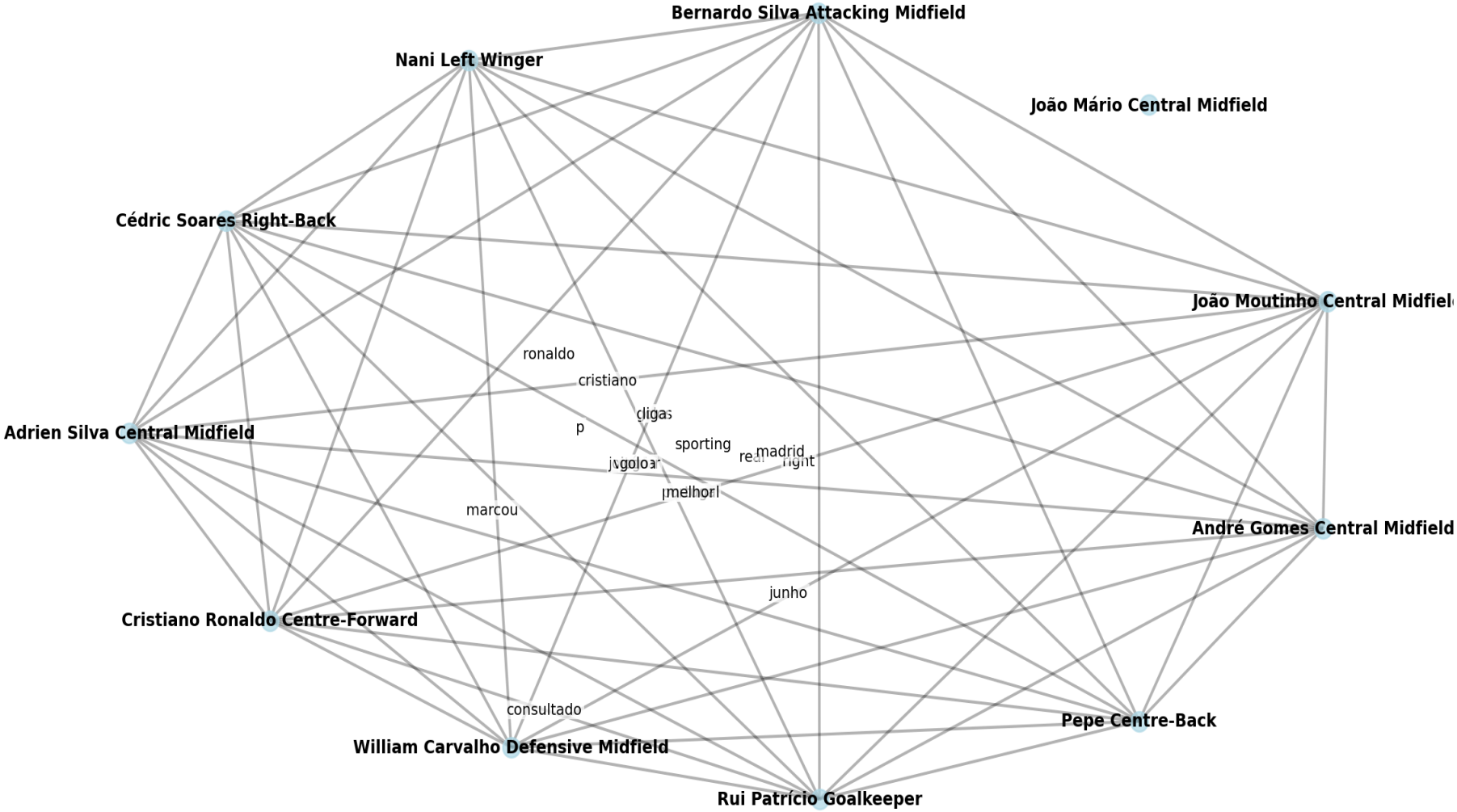
Rui Patrício Goalkeeper

William Carvalho Defensive Midfield

Summary

universidade de aveiro

Connections Between Players by Shared Keywords


Bernardo Silva Attacking Midfield
Nani Left Winger
João Mário Central Midfield
Cédric Soares Right-Back
João Moutinho Central Midfiel
ronaldo
cristiano
liga
p
sporting
madrid
goloa
melhorl
marcou
André Gomes Central Midfield
Adrien Silva Central Midfield
junho
Cristiano Ronaldo Centre-Forward
consultado
Pepe Centre-Back
William Carvalho Defensive Midfield
Rui Patrício Goalkeeper

Neural Network for Market Value Prediction

	Player	tokens	Market value
0	Adrien Silva Central Midfield	[sapo, início, início, atualidade, desporto, e...	56000000
1	André Gomes Central Midfield	[sapo, início, início, atualidade, desporto, e...	77000000
2	Bernardo Silva Attacking Midfield	[vésperas, estreia, rúben, amorim, treinador, ...	35000000
3	Cristiano Ronaldo Centre-Forward	[cristiano, ronaldo, página, apresenta, trecho...	460000000
4	Cédric Soares Right-Back	[confimação, saída, nuno, coelho, surge, possi...	25500000

Predictor

```
class PlayerValuePredictor:
    def __init__(self):

        # Create pipeline with Portuguese-specific TF-IDF
        self.pipeline = Pipeline([
            ('tfidf', TfidfVectorizer(
                max_features=1000,
                ngram_range=(1, 2),
                min_df=2
            )),
            ('scaler', StandardScaler(with_mean=False)),
            ('model', RandomForestRegressor(
                n_estimators=500,
                max_depth=None,
                min_samples_split=3,
                min_samples_leaf=2,
                max_features='sqrt',
                random_state=126784
```

Usage Example

Radio Cadena Voces @RCVHonduras

Portugal se impuso este sábado por 3-0 a Turquía y se clasificó para los octavos de final de la Eurocopa 2024 como primera del Grupo F, gracias a un gol de Bernardo Silva, otro de Samet Akaydin en propia puerta y un tanto de Bruno Fernández. #RCVNoticias

rcv.hn pic.x.com/wK0Vn2PLAK

```
1 text = 'Portugal se impuso este sábado por 3–0 a Turquía y se clasificó par
2 new_keywords = predictor.preprocess_portuguese_text(text)
3 predicted_value = predictor.predict(new_keywords)
4 print(f"\nPredicted value for new player: ${predicted_value:,.2f}")
```

```
Predicted value for new player: $65,251,703.39
```

Conclusions

- Advanced text processing & sentiment analysis
- Visualizations of player connections
- Market value trend prediction

Future Research Directions

- Potential for real-time market value predictions
- Expansion to other football leagues and languages
- Integration with broader sports analytics systems

References

- Sozen, Y. (2023). Predicting Football Players Market Value Using Machine Learning
- Transfermarkt Documentation
- Arquivo.pt API Documentation