Homework 2: Building Wrapper

INF 558 BUILDING KNOWLEDGE GRAPH

Runqi Shao 9418534943

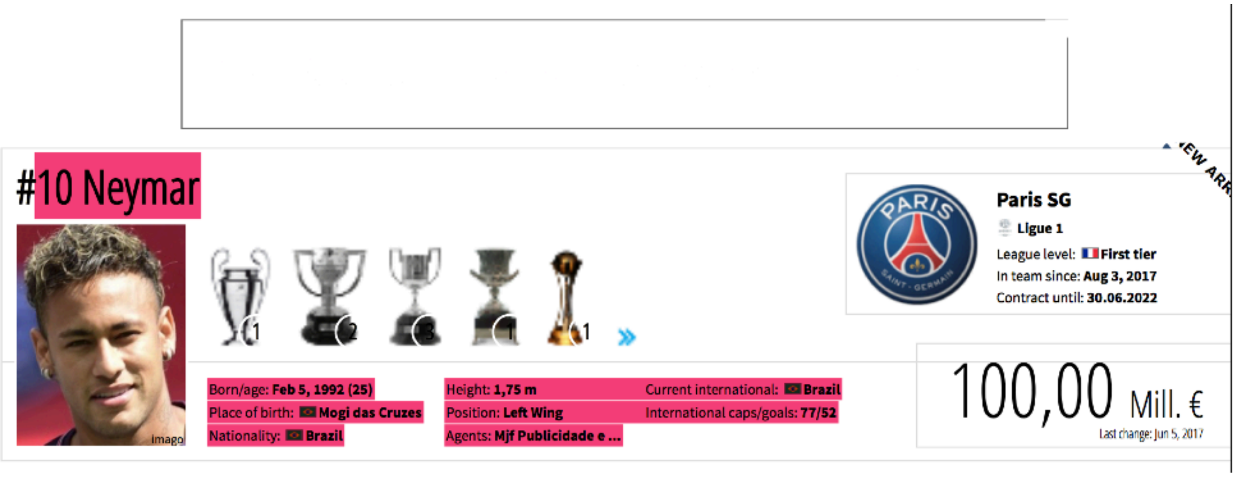
Task 1

Interesting fields from transfermarkt webpage:

- Fields about player:

Born/age
Nationality
Current international
Current club
Complete name
Player's agent
Outfitter
Main position
Current market value

Place of birth
Height
International caps/goals
Contract until
Foot
In the team since
Shoe model
Other position(s)
Last change



#10 Neymar

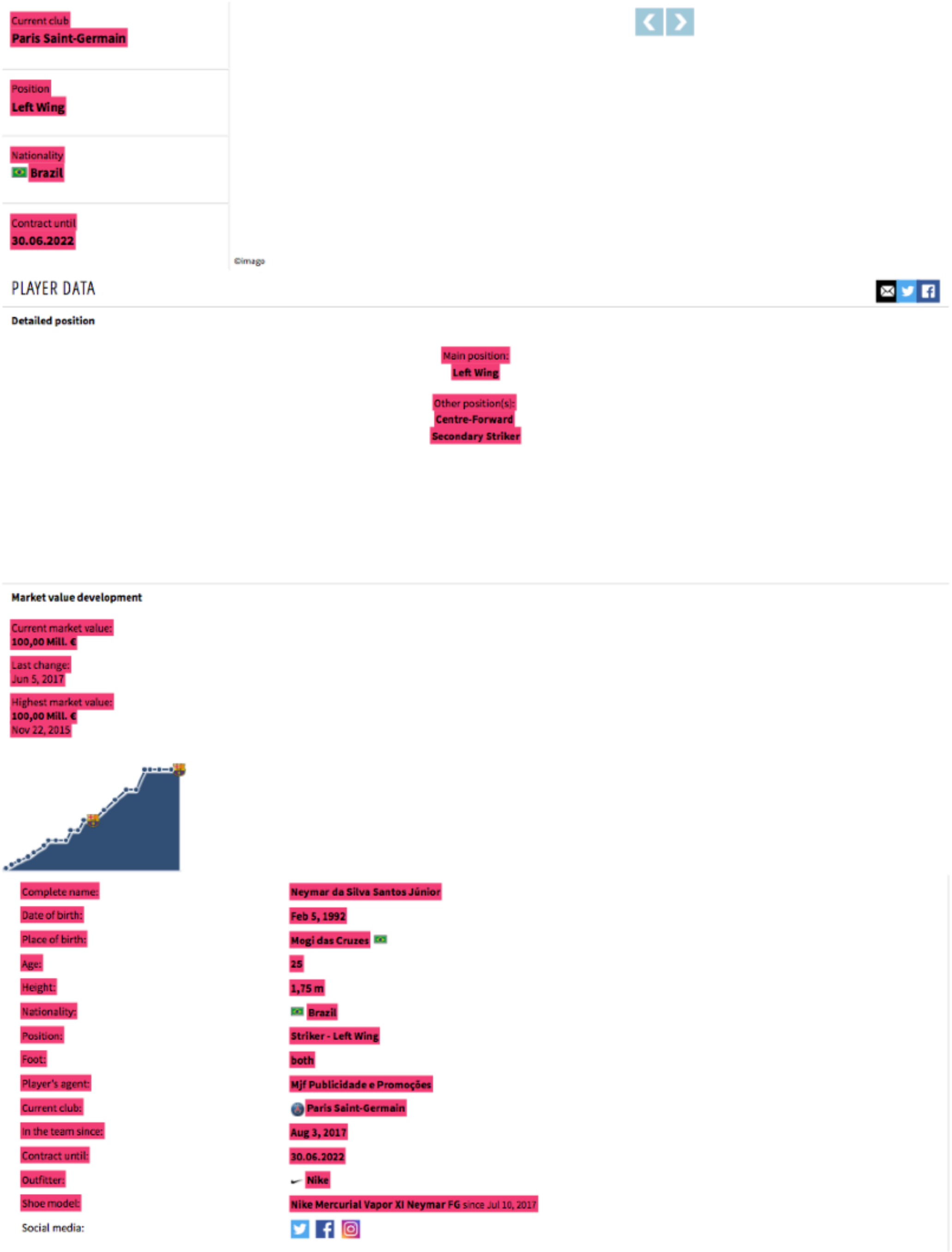
Born/age: **Feb 5, 1992 (25)**
Place of birth: **Mogi das Cruzes**
Nationality: **Brazil**

Height: **1,75 m**
Position: **Left Wing**
Agents: **Mjf Publicidade e ...**

Current international: **Brazil**
International caps/goals: **77/52**

Paris SG
Ligue 1
League level: **First tier**
In team since: **Aug 3, 2017**
Contract until: **30.06.2022**

100,00 Mill. €
Last change: Jun 5, 2017



Current club
Paris Saint-Germain

Position
Left Wing

Nationality
Brazil

Contract until
30.06.2022

©imago

PLAYER DATA

Detailed position

Main position:
Left Wing

Other position(s):
Centre-Forward
Secondary Striker

Market value development

Current market value:
100,00 Mill. €

Last change:
Jun 5, 2017

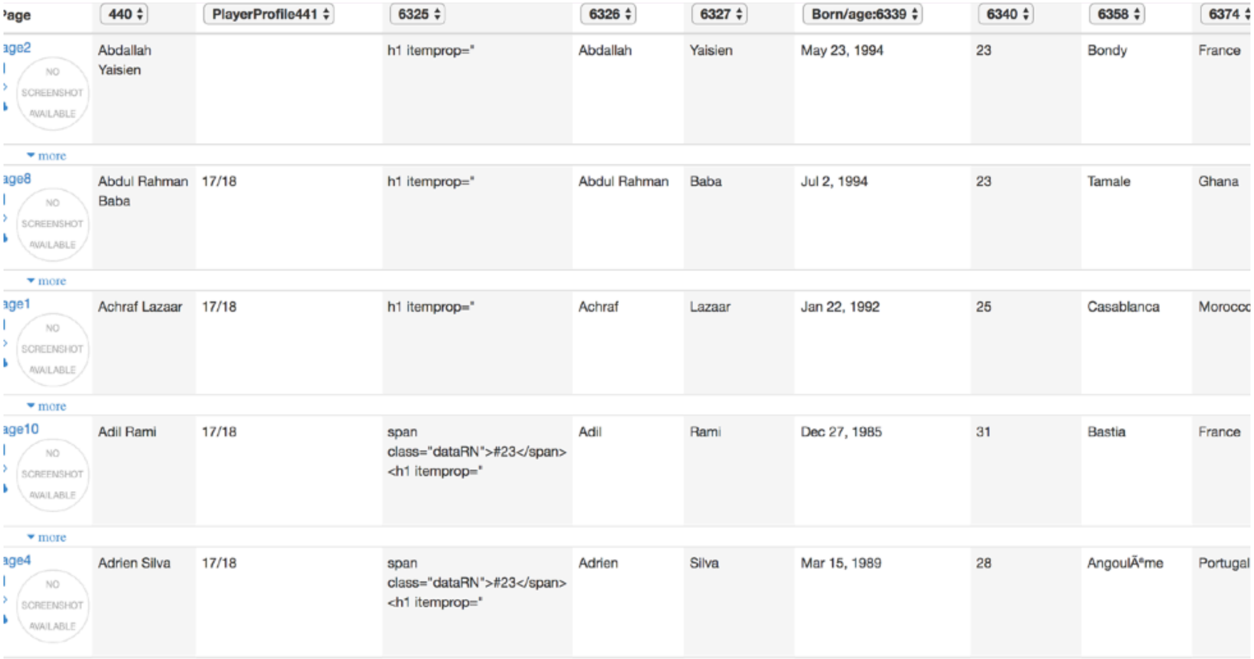
Highest market value:
100,00 Mill. €
Nov 22, 2015

Complete name:	Neymar da Silva Santos Júnior
Date of birth:	Feb 5, 1992
Place of birth:	Mogi das Cruzes
Age:	25
Height:	1,75 m
Nationality:	Brazil
Position:	Striker - Left Wing
Foot:	both
Player's agent:	Mjf Publicidade e Promoções
Current club:	Paris Saint-Germain
In the team since:	Aug 3, 2017
Contract until:	30.06.2022
Outfitter:	Nike
Shoe model:	Nike Mercurial Vapor XI Neymar FG since Jul 10, 2017
Social media:	

Task 2

No screen from inferlink available.

Screen shots of verify page of inferlink:



'age	440	PlayerProfile441	6325	6326	6327	Born/age:6339	6340	6358	6374
age2 NO SCREENSHOT AVAILABLE	Abdallah Yaisien		h1 itemprop="	Abdallah	Yaisien	May 23, 1994	23	Bondy	France
▾ more									
age8 NO SCREENSHOT AVAILABLE	Abdul Rahman Baba	17/18	h1 itemprop="	Abdul Rahman	Baba	Jul 2, 1994	23	Tamale	Ghana
▾ more									
age1 NO SCREENSHOT AVAILABLE	Achraf Lazaar	17/18	h1 itemprop="	Achraf	Lazaar	Jan 22, 1992	25	Casablanca	Morocco
▾ more									
age10 NO SCREENSHOT AVAILABLE	Adil Rami	17/18	span class="dataRN">#23</span> <h1 itemprop="	Adil	Rami	Dec 27, 1985	31	Bastia	France
▾ more									
age4 NO SCREENSHOT AVAILABLE	Adrien Silva	17/18	span class="dataRN">#23</span> <h1 itemprop="	Adrien	Silva	Mar 15, 1989	28	AngoulÃªme	Portugal

Head and first 2 samples data of downloaded csv file:

```
“,
7286”,"1,6383","30.06.6396","440","6325","6326","6327","6340","6358","6374","6405","6
412","6428","6440","6445","6459","6487","6498","6518","6520","6702","6844","7282","7
314","7525","Allteamsoftheplayer'scarreer.7231","Born/age:6339","Currentmarketvalue:
6874","Foot:6964","Informationandfacts6920","Inteamsince:6504","Lastchange:
6524","Lastchange:6896","PlayerProfile441","Position:6389","Position:
6960","Totaltransferproceeds:7135","__PAGENAME__","until7285","â¬6900"

"2007","82","2019","Aaron Ramsey","span class=""dataRN"">#8</span> <h1
itemprop=""","Aaron","Ramsey","26","Caerphilly","Wales","<span class=""cp""
title=""Base Soccer Agency Ltd"">Base Soccer Agency ...</span>","Current
international","Wales","47","12","Getty Images","League level","First
tier","35,00","Mill","Arsenal FC","""><span>Pronunciation</span></div> <div
class=""aussprache""> <audio id=""audio_aussprache"" controls preload=""none""
name=""audio_50057""> <source src=""https://tmssl.akamaized.net/static/audio/spieler/
50057.MP3"" type=""audio/mpeg""> </audio> <span> <a href=""https://
internationalassociationfootball.wordpress.com/nameproject-uefa-euro-2016/""
```

target=""_blank"" title=""Infos on the pronunciation tool""> <img src=""/images/info.png""/> </a> </span> <br class=""clearer""/> </div> </div> <div class=""weitere-daten-spielerprofil""> <div class=""blauerbalken","Cardiff U18","span>Premier League</span></div> </","Shkodran Mustafi","","Dec 26, 1990","35,00 Mill","both","> <th>Name in home country:</th> <td>Aaron James Ramsey</td> </tr> <tr","Jul 1, 2008","28, 2017","35,00 Mill","17/18","Central Midfield","Midfield - Central Midfield","6,40 Mill. â¬","page5","","Aug 1, 2016”

"2011","73","2018","Abdallah Yaisien","h1 itemprop="""","Abdallah","Yaisien","23","Bondy","France","EW Management GmbH","Former international","France U20","2","0","Getty Images","Country","France","250","Th","Paris Saint-Germain","","Paris SG U19","span>Serie C - Girone A</span><span>Campionato Primavera Girone A</span><span>Serie B</span><span>Coppa Italia</span><span>Serie A</span></div> </","Yohan Demoncy","","May 23, 1994","250 Th","left","","Jul 27, 2017","19, 2015","400 Th","","Attacking Midfield","Midfield - Attacking Midfield","0","page2","","Jul 11, 2013”

Q. Can Inferlink tool extract your highlighted fields?

It can extrac many of my highlighted fields. But not all.

Q. If not, list up to 3 fields that cannot be extracted and explain why the tool cannot extract these fields.

Inferlink tool can not extract all transfer records.

The reason, by looking at the rules that inferlink generated, I think is because it is over-specific. It tends to find the name of data every time, otherwise, the data will not be extracted. This is bad for my page because the transfer records is in a table and only have one head line. So inferlink only extract the first transfer record.

Rules:

```
{
      "begin_regex": "\\.\\s+\\</p\\>\\s+\\<div\\>.*?\\<div\\s+class=\"spielerstationen\\-headline\"\\>\\s+\\<span\\s+class=\"spielerstationen\\-",
      "end_regex": "headline\\-text\"\\>Transfer",
      "id": "810c5062-6698-4ebc-94b5-c69dcfb84f77",
      "include_end_regex": true,
      "name": "Allteamsoftheplayer'scarreer.7231",
      "removehtml": false,
      "rule_type": "ItemRule",
      "strip_end_regex": "headline\\-text\"\\>Transfer",
      "visible_chunk_after": "Transfer",
      "visible_chunk_before": "All teams of the player's carreer."
    },
    {
      "begin_regex": "class=\"station\\-saison\"\\>\\s+\\-.*?/\\>\\<b\\>\\<.*?a\\s+class=\"vereinprofil_.*?/saison_id/.*?\"\\>",
      "end_regex": "\\<",
      "id": "4088a02b-e77e-481b-864c-d54070e0932d",
      "include_end_regex": true,
      "name": "7282",
      "removehtml": false,
      "rule_type": "ItemRule",
      "strip_end_regex": "\\<"
    }
```

Q. Choose one extracted rule and explain how the rules can be used to extract field from webpages.

One sample rule:

```
{
    "begin_regex": "\\>\\s+\\<span\\s+itemprop=\"birthDate.*?\"\\s+class=\"dataValue\".*?
\\>",
    "end_regex": "\\(",
    "id": "61fd2f12-e0fa-4d65-ba7b-4e7dcb929ba9",
    "include_end_regex": true,
    "name": "Born/age:6339",
    "removehtml": false,
    "rule_type": "ItemRule",
    "strip_end_regex": "\\(",
    "visible_chunk_after": "(",
    "visible_chunk_before": “Born/age:"
}
```

This rule can be used to extract “Born/age” field from webpages because the Regex rule matches the following example element in html. “begin_regex” matches blue pattern, “end_regex" matches red “pattern”.

<span class=\"dataItem\">Born/age:</span>\r\n\t\t\t\t\t\t\t\t<span itemprop=\"birthDate\" class=\"dataValue\">\r\n\t\t\t\t\t\t\t\t\tSep 11, 1997\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t(19)\r\n\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t\t</span>

Task 3

I use BeautifulSoup to write my own wrapper.

Basically, I use functions provided by BeautifulSoup to find and locate tags in html. Such as “table”. Or use regex matching provided by BeautifulSoup to find the location where contains the “text” I want. Then do some cleaning job to form the output json object for each player.

The libraries I used:

bs4, BeautifulSoup 4, which helps me find data I want

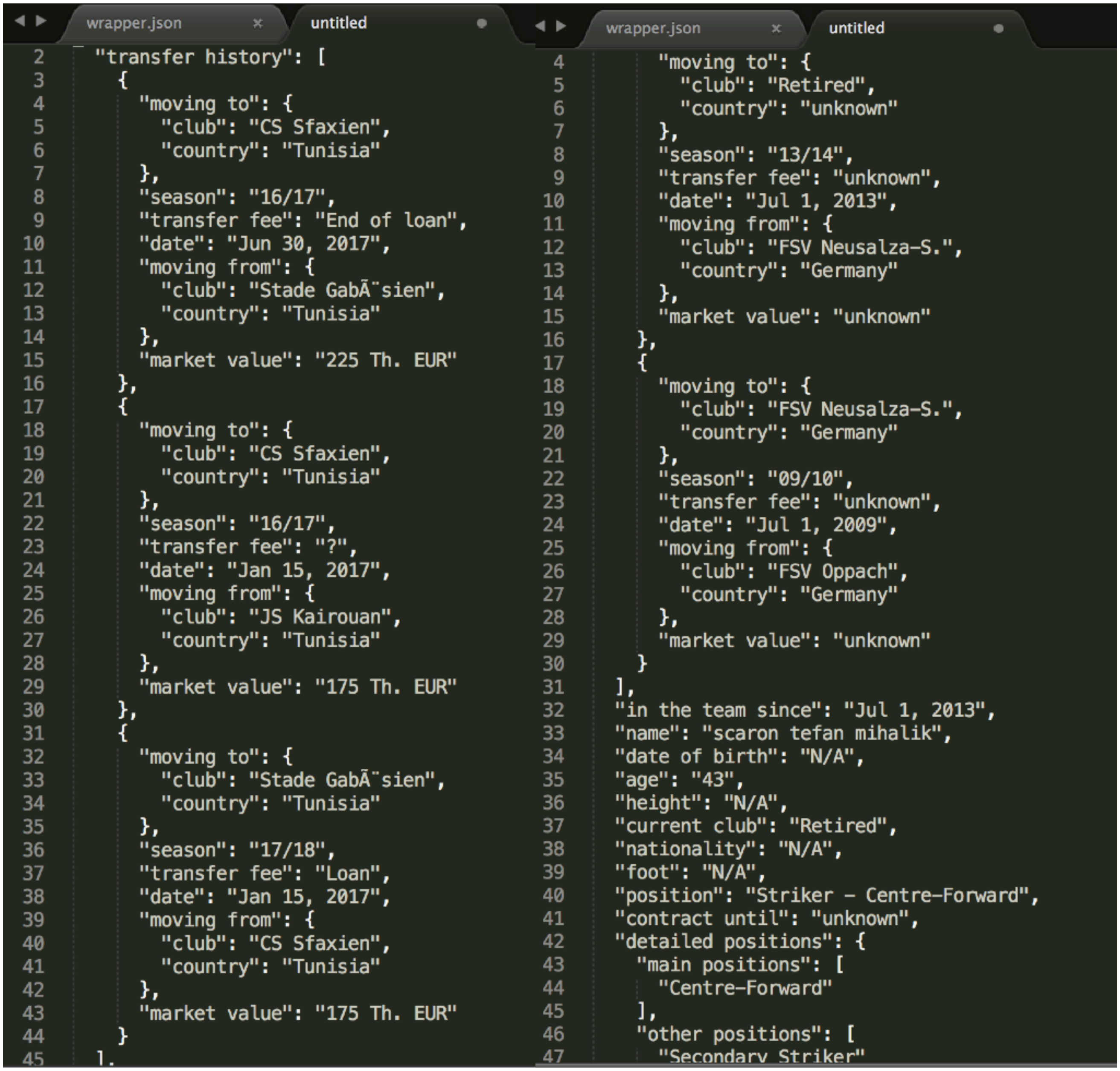
re, regex library

codecs, open jl file

sys, get input arguments

json, reading and writing json.

First two json objects:



```
  "transfer history": [
    {
      "moving to": {
        "club": "CS Sfaxien",
        "country": "Tunisia"
      },
      "season": "16/17",
      "transfer fee": "End of loan",
      "date": "Jun 30, 2017",
      "moving from": {
        "club": "Stade GabÃ¨sien",
        "country": "Tunisia"
      },
      "market value": "225 Th. EUR"
    },
    {
      "moving to": {
        "club": "CS Sfaxien",
        "country": "Tunisia"
      },
      "season": "16/17",
      "transfer fee": "?",
      "date": "Jan 15, 2017",
      "moving from": {
        "club": "JS Kairouan",
        "country": "Tunisia"
      },
      "market value": "175 Th. EUR"
    },
    {
      "moving to": {
        "club": "Stade GabÃ¨sien",
        "country": "Tunisia"
      },
      "season": "17/18",
      "transfer fee": "Loan",
      "date": "Jan 15, 2017",
      "moving from": {
        "club": "CS Sfaxien",
        "country": "Tunisia"
      },
      "market value": "175 Th. EUR"
    }
  ],
```

```
      "moving to": {
        "club": "Retired",
        "country": "unknown"
      },
      "season": "13/14",
      "transfer fee": "unknown",
      "date": "Jul 1, 2013",
      "moving from": {
        "club": "FSV Neusalza-S.",
        "country": "Germany"
      },
      "market value": "unknown"
    },
    {
      "moving to": {
        "club": "FSV Neusalza-S.",
        "country": "Germany"
      },
      "season": "09/10",
      "transfer fee": "unknown",
      "date": "Jul 1, 2009",
      "moving from": {
        "club": "FSV Oppach",
        "country": "Germany"
      },
      "market value": "unknown"
    }
  ],
  "in the team since": "Jul 1, 2013",
  "name": "scaron tefan mihalik",
  "date of birth": "N/A",
  "age": "43",
  "height": "N/A",
  "current club": "Retired",
  "nationality": "N/A",
  "foot": "N/A",
  "position": "Striker - Centre-Forward",
  "contract until": "unknown",
  "detailed positions": {
    "main positions": [
      "Centre-Forward"
    ],
    "other positions": [
      "Secondary Striker"
```