A close-up photograph of a tennis court. In the foreground, a bright yellow-green tennis ball sits on the green artificial turf surface, near a white boundary line. Behind the ball, a black tennis net with a rope mesh extends across the frame. The background is slightly blurred, showing a clear blue sky and a building in the distance.

Tennis sentiment analysis

By Louis Grillon
27/01/2021

Table of content

1. Context
2. Approach
3. Methodology
4. Results
5. Improvements
6. To broaden things out



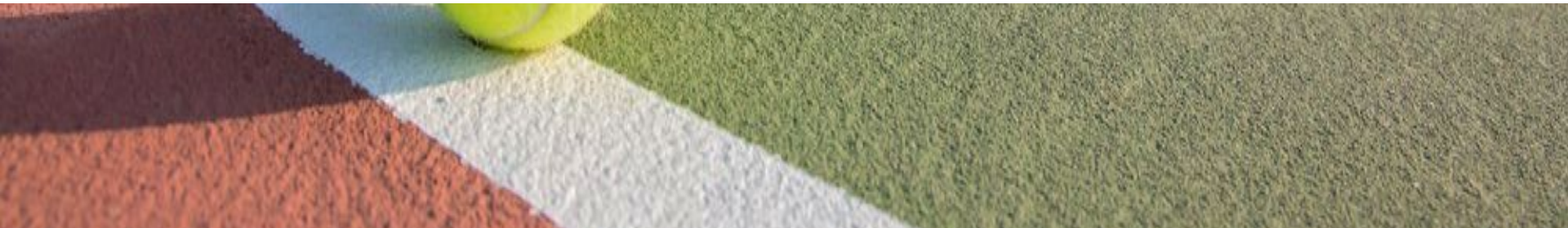
Context

01 | I am a very big tennis fan (and player), and love data!

02 | There are different ways to compare and rank players (tournaments, points, ...). But those don't take into account the behaviour of the players, on and off the court !



How can we measure and compare the behaviour of the players?



Approach: Sentiment Analysis

What is a Sentiment Analysis?

A sentiment analysis (or opinion mining) is a recent machine learning technique used to identify, extract, quantify, and study opinion and information about a chosen topic, without having a proper dataset to analyze.

How does it work?

- 1 Gather unstructured text data about what people write (usually on social media) about the topic you want to analyze
- 2 Apply Natural Language Processing on that data, in order to quantify how positive or negative the reviews about the topic is

Methodology

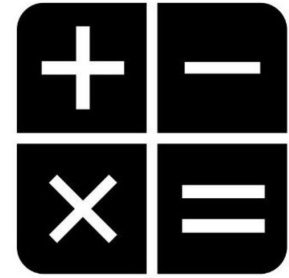


Step 1

Web Scrape wikipedia in order to automatically get who are the best players, and their ranks

Step 2

Get relevant tweets about those players and clean them



Step 3

Quantify how positive the tweets are (based on NLP), and aggregate them by player



Step 1: Web Scraping to get the best players and their rank

1. Automatically download wikipedia page about top tennis players
2. Find the information we need in that html file, and transform it into a table



Step 2: Retrieve tweets about the players

1. Retrieve tweets in python (tweepy) containing the name of the player, preceded by a hashtag. (#rogerfederer, #federerroger, #rfederer, etc...)
2. Clean the tweets (remove punctuation, lemmatize, remove unnecessary words)



Step 3: Natural Language Processing



- 1 For each tweet, we will use a Natural Language Processing library (which works like a dictionary).
- 2 Split each tweet in words. Each word will be attributed a coefficient. Then, the total polarity (positivity) of each tweet is computed (a value between -1 and 1).
- 3 Aggregate all the tweets per player and per day, while averaging the polarity of the tweets

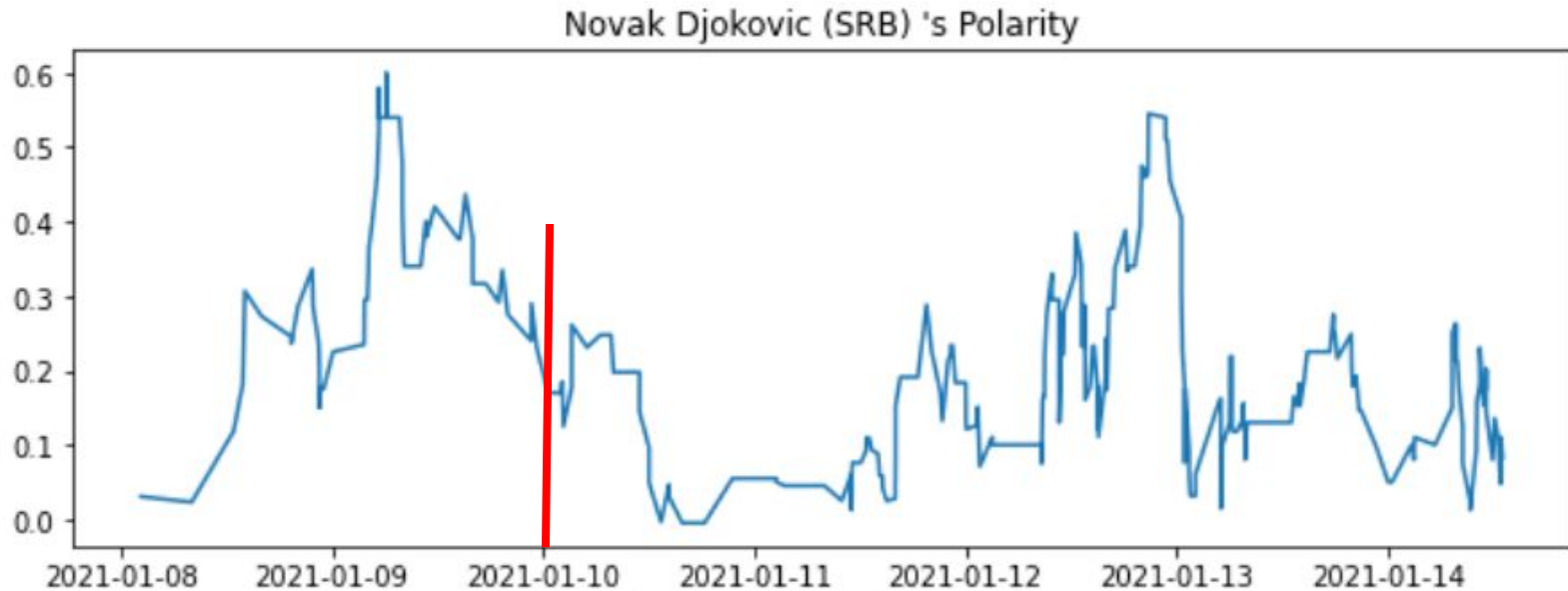
“What a great match ! Nice play from #RFederer”

great
+0.3

Nice
+0.2

Total polarity of that tweet: $0.3+0.2=0.5$

Results: Rolling Moving Average (L7D)



Djokovic tweets about his disagreement with the official new health restrictions for a major tournament

Answer to the problem:



Player	ATP points
Novak Djokovic	12 030
Rafael Nadal	9 850
Dominic Thiem	9 215
Daniil Medvedev	8 470
Roger Federer	6 630



Player	Polarity (L7D)
Roger Federer	0.23
Rafael Nadal	0.21
Novak Djokovic	0.20
Daniil Medvedev	0.20
Dominic Thiem	0.19





Improvements

- 01 | Bias due to the language (english) or the seasonality (birthday)
- 02 | Subjectivity and irony are not taken into account
- 03 | Only tweets from the last 7 days, in a limited amount

To broaden things out

Sentiment Analysis is a quite powerful tool to understand how the public feels about a topic, and especially efficient when coupled with social media.


In a business context, this is what it can be used for:



- Review customer comments about products with Sentiment Analysis, in order to improve or cut them out



- Evaluate popularity of politicians before an election



Thank you for your attention !
Feel free to ask me anything !



To view other personal projects about various data topics (business intelligence dashboards, price forecasting, multiplayer market analysis, computer vision, ...), please feel free to go to my portfolio website:
https://grillon6u.github.io/Data_Science_projects/