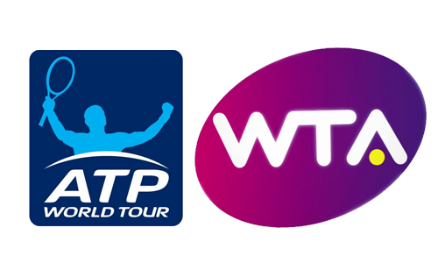
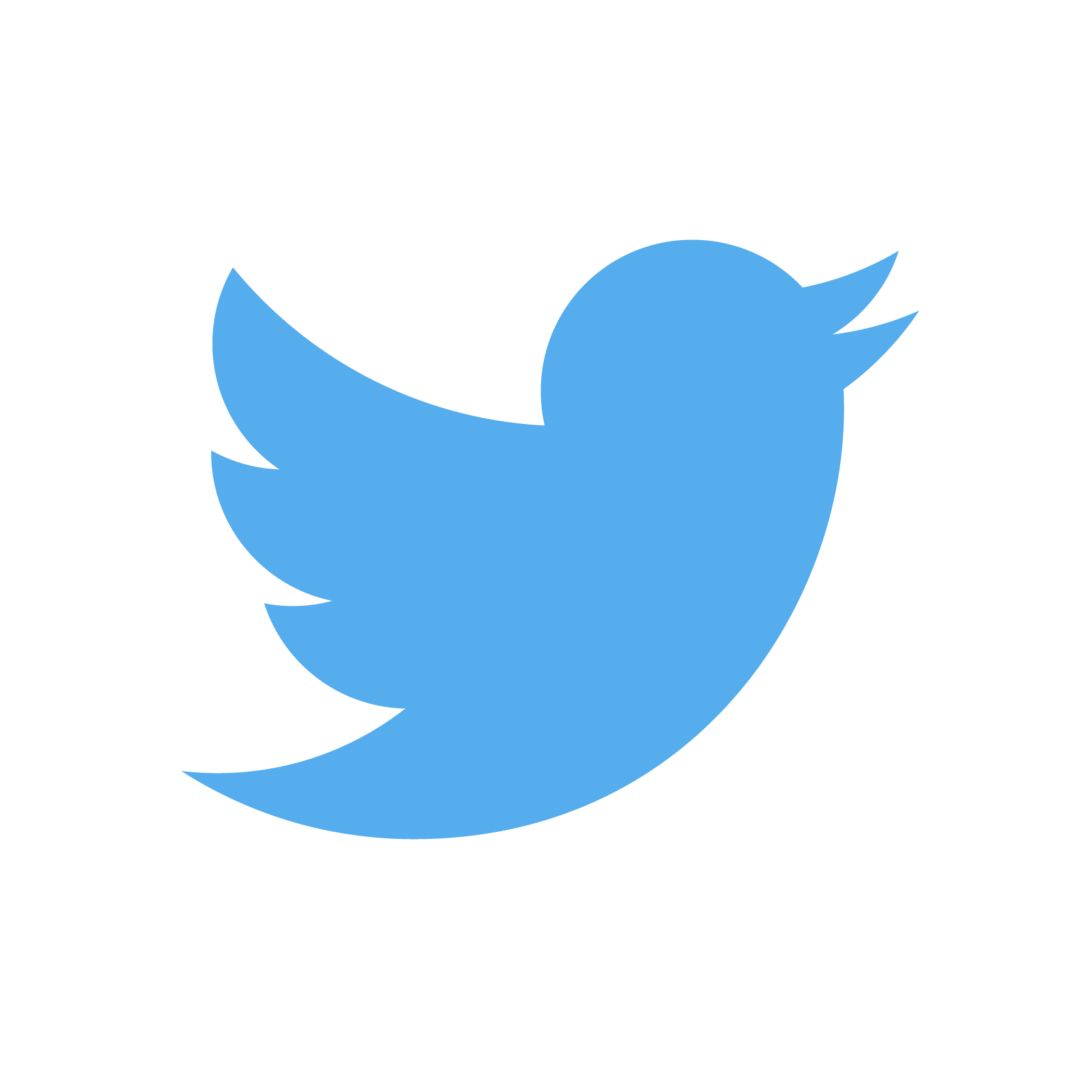


Universität St. Gallen  
Hochschule für Wirtschafts-, Rechts- und Sozialwissen-

schaften sowie Internationale Beziehungen

**Sportler Case  
Datenbeschaffung / Datenverarbeitung / Datenaufbereitung**



**8,045,1.00: IC: Technology und Market Intelligence**

**21.05.19**

**Regina Arocha || Elias Böni || Kevin Elliott || Manuela Huber || Fabian Schmid**

Inhaltsverzeichnis

[1. Einführung 3](#_Toc9327524)

[2. Datenbeschaffung 4](#_Toc9327525)

[2.1 Tweets 4](#_Toc9327526)

[2.2 Instagram 4](#_Toc9327527)

[2.3 ATP / WTA 5](#_Toc9327528)

[3. Datenverarbeitung 6](#_Toc9327529)

[4. Tableau 9](#_Toc9327530)

# Einführung

Dieses Dokument bietet eine zusätzliche Erklärung und Übersicht zu den Kommentaren, welche in den jeweiligen Codes und Skripts vorzufinden sind.

Für die Ausführung der einzelnen Codes ist der GitHub Repository „*regina21/golfis*“ relevant.

Wahl der Tennisspieler  
Der Fokus liegt bei den Top 10 der Welt Männer Tennisspieler, die auf beiden Social-Media-Kanälen (Twitter und Instagram) aktiv sind. Aufgrund von Inaktivität gegenüber Social Media wird der Nummer 3 Top-Player (Alexander Zverev) aus der Probe entnommen. Dafür wurde Marin Cilic (Nummer 11) hinzugenommen. Die ausgewählten Top-Player basieren auf dem ATP-Ranking vom 29.03.2019 auf https://www.atptour.com/en/rankings/singles. Der Schwerpunkt der Arbeit liegt auf dem Einzelranking.

Auch bei den Frauen wurden die 10 Besten der Welt Spielerinnen, die auf beiden Social-Media-Kanälen (Twitter und Instagram) aktiv sind ausgewählt. Aufgrund von Inaktivität gegenüber Social Media ist die Top 9, Ashley Barty, nicht Teil der Stichprobe. Deshalb wurde stattdessen die Top 11, Serena Williams, hinzugenommen.

Die ausgewählten Top-Frauenspieler basieren auf dem WTA-Ranking vom 24.04.2019.

auf https://www.wtatennis.com/rankings. Der Schwerpunkt der Arbeit liegt auf dem Einzelranking.

Die Datei „*Tennis\_Twitter\_Instagram.xlsx*“ bietet eine Übersicht der ausgewählten Spieler(-innen) inkl. deren Social Media Name.

# Datenbeschaffung

## 2.1 Tweets

|  |  |  |  |
| --- | --- | --- | --- |
| Skript | Setup.py | Automate.py | Datenbeschaffung.R |
| Input | - | - | Tweets.csv |
| Output | - | Tweets.csv | Twitter\_complete\_v1 |

Die Tweets der ausgewählten Spieler und Spielerinnen wurden für den Zeitraum 01.01.2017 bis 31.12.2018 mit dem Skript „automate.py“ gecrawlt.   
Diese CSV Datei wurde zunächst ins R-Markdown „*Datenbeschaffung.R*“ importiert (wichtig, damit der Code funktioniert, muss diese Datei in einem separaten Ordner als die instagram \*.csv Dateien gespeichert werden). Die Datei wurde weiterverarbeitet und als Excel-Datei „*twitter\_complete\_v1.xlsx*“ exportiert.   
In dieser Excel-Datei wurde eine neue Spalte hinzugefügt, in welcher die einzelne Tweets anhand der nachfolgenden Excel – Formel übersetzt worden sind: *=GOOGLETRANSLATE(Tweets;DETECTLANGUAGE(Tweets);"en")*

Die Datei wurde anschliessend als „*twitter\_complete\_v2.xlsx*“ gespeichert.

## Instagram

|  |  |  |  |
| --- | --- | --- | --- |
| Skript | instaloader.py | Instagram.py | Datenbeschaffung.R |
| Input | - | - | Die einzelnen \*.csv pro Spieler(-in) |
| Output | Mehrere \*.txt pro Spieler(-in) | Eine \*.csv pro Spieler(-in) | Instagram\_complete\_v1.xlsx  Import\_list.csv |

Mit dem Skript „*instaloader.py*“ wurde in der command Zeile den folgenden Befehl eingegeben:   
*„instaloader USERNAME [USERNAME] --post-filter="date\_utc >= datetime(2017, 1, 1)"*

Dies hat für jeden Spieler einen Ordner erstellt, in der die einzelnen Instagram Posts auf separate Textdateien gespeichert wurden. Der Titel der Textdatei entsprach jeweils dem Datum, an dem es gepostet worden ist.

Das Skript „*instagram.py*“ wurde in die einzelne Instagram-Spieler(-innen) Ordner gespeichert und ausgeführt. Dieses Skript hat jeweils alle Textdateien samt Titel in eine einzige \*.csv Datei gespeichert. Die einzelnen \*.csv Dateien wurden mit dem Skript „*Datenbeschaffung.R*“ miteinander kombiniert, sodass eine Datei „*instagram\_complete\_v1.xlsx*“ exportiert werden konnte. In dieser Excel-Datei wurde eine neue Spalte hinzugefügt, in welcher die einzelne Posts anhand der nachfolgenden Excel – Formel übersetzt worden sind: *=GOOGLETRANSLATE(Posts;DETECTLANGUAGE(Posts);"en")*

Die Datei wurde anschliessend als „*instagram\_complete\_v2.xlsx*“ gespeichert.

## ATP / WTA

|  |  |
| --- | --- |
| Skript | Die Match-Statistiken sind unter: <https://github.com/JeffSackmann/tennis_atp> und <https://github.com/JeffSackmann/tennis_wta> zu finden |
| Input | Atp\_matches\_2018.csv  Atp\_machtes\_2017.csv  Wta\_matches\_2018.csv  Wta:mathces\_2018.csv |
| Output | Atp\_tennis\_2017.xlsx  Atp\_tennis\_2018.xlsx  Wta\_tennis\_2017.xlsx  Wta\_tennis\_2018.xlsx |

Die CSV-Files von ATP wurden für die Jahre 2017 und 2018 aus dem Github Repository von Jeff Sackmann (<https://github.com/JeffSackmann/tennis_atp> & https://github.com/JeffSackmann/tennis\_wta) entnommen. Die CSV Dateien wurde mit einer Spalte ergänzt, in der der Winner’s und der Loser’s First Service (%) berechnet worden ist. Diese Datei wurde anschliessend als \*.xlsx gespeichert und im Skript „Datenverarbeitung.Rmd“ weiter verarbeitet.

Analog wurde dasselbe für die WTA Dateien gemacht.

# Datenverarbeitung

|  |  |  |
| --- | --- | --- |
| Skript | Datenverarbeitung.Rmd | Roger\_Federer.Rmd |
| Input | Twitter\_complete\_v2.xlsx  Instagram\_complete\_v2.xlsx  Import\_list.csv  Atp\_tennis\_2017.xlsx  Atp\_tennis\_2018.xlsx  Wta\_tennis\_2017.xlsx  Wta\_tennis\_2018.xlsx  Tennis\_Twitter\_Instagram.xlsx | Twitter\_complete.v2.xlsx |
| Output | Atp\_total.csv  Wta\_total.csv | - |

Das R-Markdown „*Datenverarbeitung.Rmd*“ beschreibt die Vorgehensweise der Datenverarbeitung. Das R-Markdown „*Roger\_Federer.Rmd*“ zeigt die verschiedenen Sentimentsanalyse Methoden und Auswertungen.

In diesem Dokument wird kurz darauf verwiesen, wie die Match-Datumsdaten approximiert worden sind (da die Datumsdaten nicht in den ATP und WTA Dateien vorhanden waren und diese Information nicht konsolidiert vorzufinden war).

Die zwei Output-Dateien dienten dann als Input für Tableau (siehe Abschnitt 4).

**Regeln für die Approximation der Match-Datumsdaten**

Die ATP und WTA Excel-Files enthalten jeweils das Startdatum des Turniers und die draw-size. Die draw-size gibt die Anzahl Spieler an. In unserem Datensatz wurde eine draw-size von 4 / 8 / 32 / 64 / 128 identifiziert. Somit wurde stichprobenartig untersucht, über wie viele Tage sich die Turniere ausstrecken, je nach draw-size. Ebenfalls wurde stichprobenartig untersucht, wie die Matches je nach Stadium vom Turnier stattfanden (also wie viele Spiele pro Tag). Daraus erhielten wir die Erkenntnis, wie viele Spiele pro Tag pro Turnier je nach draw-size in der Regel stattfanden. Da unser Datensatz die Spalte „match\_num“ erhielt, war die Reihenfolge der Matches klar. Jedoch ist die Nummerierung durchgehend. Deshalb wurde im Datensatz eine zusätzliche Spalte „seq“ eingefügt (siehe R-Markdown), welche die Matches innerhalb der Turniere durchnummerierte (siehe Tabelle):

|  |  |  |
| --- | --- | --- |
| Tourney ID | Match\_Num | Seq |
| 1234 | 1 | 1 |
| 1234 | 2 | 2 |
| 1234 | 3 | 3 |
| 345 | 4 | 1 |
| 345 | 5 | 2 |
| 233 | 6 | 1 |

Tabelle 1: Beispiel was die "seq" Spalte darstellt

Nachdem die Reihenfolge der Matches klar war, konnten die folgenden Regeln aufgestellt werden:

Draw-size = 4

Das Spiel erstreckt sich über 3 Tage. Insgesamt finden 4 Matches statt. Das heisst Matches mit der Sequenznummer 1 -2 finden am Datum vom Tournament statt, Matches mit der Sequenznummer 3 – 4 finden am Datum vom Tournament + 2 Tage statt.

Draw – size = 8

|  |  |
| --- | --- |
| Sequenznummer | Match\_dat (=tatsächliches Datum) |
| 1-2 | Tourney\_date |
| 3-4 | Tourney\_date + 1 |
| 5-6 | Tourney\_date + 2 |
| 7-8 | Tourney\_date + 3 |
| 9-11 | Tourney\_date + 4 |
| 12-13 | Tourney\_date + 5 |
| 15 | Tourney\_date + 6 |

Draw-size = 32

|  |  |
| --- | --- |
| Sequenznummer | Match\_dat (=tatsächliches Datum) |
| 1 - 7 | Tourney\_date |
| 8 - 15 | Tourney\_date + 1 |
| 16 - 19 | Tourney\_date + 2 |
| 20 - 24 | Tourney\_date + 3 |
| 25 – 28 | Tourney\_date + 4 |
| 29 – 30 | Tourney\_date + 5 |
| 31 | Tourney\_date + 6 |

Draw Size = 64

|  |  |
| --- | --- |
| Sequenznummer | Match\_dat (=tatsächliches Datum) |
| 1-3 | Tourney\_date |
| 4 -13 | Tourney\_date + 1 |
| 14 -28 | Tourney\_date + 2 |
| 29 – 40 | Tourney\_date + 3 |
| 41 – 48 | Tourney\_date + 4 |
| 49 – 52 | Tourney\_date + 5 |
| 53 – 54 | Tourney\_date + 6 |
| 55 | Tourney\_date + 7 |

Draw Size = 128

|  |  |
| --- | --- |
| Sequenznummer | Match\_dat (=tatsächliches Datum) |
| 1 - 22 | Tourney\_date |
| 23 - 43 | Tourney\_date + 1 |
| 44 – 64 | Tourney\_date + 2 |
| 65 – 80 | Tourney\_date + 3 |
| 81 – 96 | Tourney\_date + 4 |
| 97 – 104 | Tourney\_date + 5 |
| 105 – 112 | Tourney\_date + 6 |
| 113 – 116 | Tourney\_date + 7 |
| 117 – 120 | Tourney\_date + 8 |
| 121 – 122 | Tourney\_date + 9 |
| 123 – 124 | Tourney\_date + 10 |
| 125 – 126 | Tourney\_date + 12 |
| 128 | Tourney\_date + 14 |

# Tableau

|  |  |
| --- | --- |
| Skript |  |
| Input | Atp\_total.csv  Wta\_total.csv  Instagram\_count.xlsx  Twitter\_count.xlsx  Twitter\_tableau.csv |
| Output | Visualisierung von Sentiment in Relation zu verschiedenen Dimensionen: Turnierarten, Datum (Zeitverlauf), Spieluntergrund, geographischer Standort, spezifisches Turnier (Wimbledon, US Open etc.)  Visualisierung deskriptiver Statistiken zum Datensatz (Anzahl Posts etc.)  Visualisierung anderer Performance Indikatoren und deren Zusammenhang zu sowohl Sentiment als auch First Service % (im Appendix der Präsentation) |

Die Dateien, welche aus dem Skript „Datenverarbeitung.Rmd“ exportiert worden sind, dienten als Grundlage für die Erstellung der Visualisierung im Tableau.