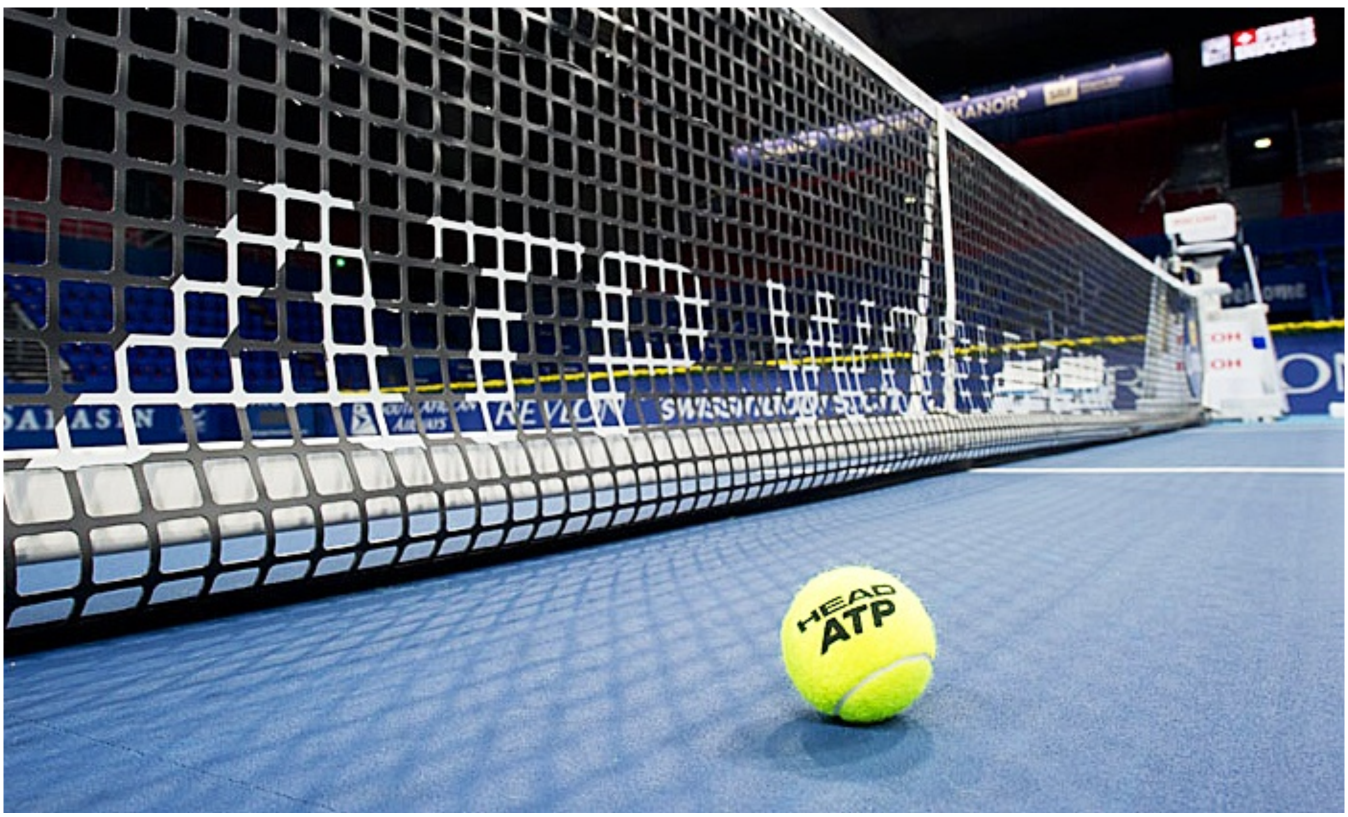
# Data Management and Database Design in Association of Tennis Professionals




Xinying Shi

Wendi Yu

2018 April

# Introduction

Our project files include:

- FinalProjectMysql.ipynb, which is  answering all the 12 questions about social media by using MySQL;
- FinalProjectNoSQL.ipynb, that file is answering 8 of 12 questions on NoSQL;
- FinalProject(OtherTables).ipynb, that jupyter notebook conclude how we created Synonyms table, Mis-spelling table and Semantic Table.

You could see all of this raw data on this file folder.

# Data Source

1. This is the master data I'm going to use, which is collected by someone else by Kevin Lin on GitHub.
   https://github.com/serve-and-volley/atp-world-tour-tennis-data/tree/master/csv

## ATP World Tour tennis data

This repository contains Python scripts that scrape tennis data from the ATP World Tour website, as of October 2017. Note that if the site layout is subsequently redesigned, then these scripts will no longer work.

### License



### Contents



2. Coach data is scrapped by ourselves from ATP official website.
   http://www.atpworldtour.com/

| coach_id | coach_name | coaching_player_id |
|---|---|---|
| 1 | Ivan-Ljubicic | f324 |
| 2 | Daniel-Vallverdu | d875 |
| 3 | Ricardo-Acioly | sg64 |

3. Twitter Data are scrapped by using Tweepy API.

4. Instagram Data are scrapped by using Instagram-scraper API.

5. Data of Synonyms table is using wordnet to get the synonyms words.

6. To get the data of  mis-spellings table, we use levenshtein distance to find words that are off by one or two letters from words. Here is the function that we use.

```
# Define a function to check two words
def levenshtein(s, t):
        ''' From Wikipedia article; Iterative with two matrix rows. '''
        if s == t: return 0
        elif len(s) == 0: return len(t)
        elif len(t) == 0: return len(s)
        v0 = [None] * (len(t) + 1)
        v1 = [None] * (len(t) + 1)
        for i in range(len(v0)):
            v0[i] = i
        for i in range(len(s)):
            v1[0] = i + 1
            for j in range(len(t)):
                cost = 0 if s[i] == t[j] else 1
                v1[j + 1] = min(v1[j] + 1, v0[j + 1] + 1, v0[j] + cost)
            for j in range(len(v0)):
                v0[j] = v1[j]

        return v1[len(t)]
```

7. Data of  semantic information table, we using the loop to get hashtag category data by using nltk library and load it to table semantic.

# Reference

- https://en.wikipedia.org/wiki/List_of_acronyms:_B

  Forms that we use to get synonyms table

- https://en.wikibooks.org/wiki/Algorithm_Implementation/Strings/Levenshtein_distance

  Levenshtein distance Algorithm

- https://github.com/rarcega/instagram-scraper

  Instagram API

- http://docs.tweepy.org/en/v3.5.0/getting_started.html

  Twitter API