Regarding the data mining, we focused on the information freely (??) available on the ATP official website http://www.atpworldtour.com/. We decided to consider only matches from tournaments of the categories "ATP Challenger Tour", "ATP World Tour 250", "ATP World Tour 500", "ATP World Tour Masters 1000" and "Grand Slams". Besides we ignored all tournaments that started before 2000.

Main crawling steps:

1. Crawling Phase
    - Crawling all 45 seasons (2000 - 2014, category 1, 2 and 4)
        - Gathering basic data on tournaments
        - Following links to more than 3250 tournaments found
            - Gathering complete data on tournament
            - Creating tables of players hyperlinks
                - Following links to more than 110.000 matches found
2. Database generation
    - Merging the matches of all tournaments
    - Pre treatment of the player database
    - Following links to more than 3400 players found
        - Gathering general data
        - Using ranking history to compute actualized rankings
3. Post treatment, cleaning all bases

Temps total de crawling : 20h30 Nombre de pages visitées: 125.000 Poids de la base nettoyée : 97.7 Mo Poids total des bases (dont auxiliaires) : 207 Mo

Figure 1: Python example