Questions to be answered

1. What universities/countries/regions/continents are the most consistent through the years, and which regions are least consistent? For this we can consider both the rankings and the scores.

2. Are there any noteworthy differences/anomalies in the top 200 universities of the world? Find and analyze the ones you consider most interesting.

3. What are the patterns you discovered that you suspect could be interesting? Does the data contain any unusual patterns through the years that you did not expect? Although incomplete you could also use more data of the years before 2016.

Introduction

We hebben gekozen voor het University Ranking dataset. Voor het eerste college begonnen we al de pagina's te scrapen die we nodig zouden hebben. En hier liepen we al tegen mijn eerste probleem aan, het HTML-bestand zoals werd opgevraagd met de *requests* package van Python had niet de benodigde tabel erin, de website laadde deze interactief op door middel van JavaScript of iets dergelijks. Hierdoor moesten we een webdriver gebruiken, deze webdriver zou de pagina openen, de door JavaScript gecreëerde tabel uitlezen en zo konden we door met de relevante data te scrapen. Door de tabel heen gaan was toen simpel, een paar for-loops maakte alle benodigde data opzoeken geen probleem.

Hierna besefte we ons dat er data misten in de tabel, veel waardes uit de kolom "overall score" en "percent international students" miste. Hierop hebben we gekozen om een machine learning algoritme op de missende data punten los te laten. Deze zou als training data 80% van de wel ingevulde datapunten nemen, en hiermee een zo goed mogelijke voorspelling maken voor alle waardes die misten. Als testing data namen we de resterende 20%. Voor de kolommen "overall score" en 'industry_score' had de classifier het 98.9% en 50%, respectievelijk, van de tijd precies goed, wanneer het algoritme getest werd tegenover de training data.

Voor de kolom "female/male ratio" had de classifier, met een accuratie van maar 10-15%, moeite. Hierdoor kozen we om het gemiddelde te nemen van het land waar de universiteit met het missende ' female/male ratio'-datapunt zich in bevond. Hierna miste alleen nog datapunten in de kolom "percent international students". Hier vulden we de waarde 0 in voor alle missende datapunten, dit omdat we vermoeden dat een missend datapunt betekende dat er 0 buitenlandse studenten waren in het gegeven jaar.

Informatie dataset, opgehaald met functie: *df.info()*

Int64Index: 2884 entries, 0 to 2883
Data columns (total 15 columns):

university_name	2884 non-null object
country	2884 non-null object
ranking	2884 non-null int64
no_student	2884 non-null int64
no_student_p_staff	2884 non-null float64
pct_intl_student	2884 non-null float64
year	2884 non-null int64
score_overall	2884 non-null float64
score_teaching	2884 non-null float64
score_research	2884 non-null float64
score_citation	2884 non-null float64
score_industry	2884 non-null float64
score_int_outlook	2884 non-null float64
male	2884 non-null int32
female	2884 non-null int32