Taak "Machine Learning"

Project “Weblogs”

HO
GENT

Project “Weblogs”: situering

- De beheerders van de zoekmachine van een bibliotheekcatalogus zouden graag weten welke bezoekers mensen “van vlees en bloed” zijn en welke bezoekers “webrobots” zijn, die hun catalogus proberen te scrapen, te indexeren voor een andere zoekmachine (zoals Google) of misschien nog andere (minder eerlijke) bedoelingen hebben.

HO GENT

Project “Weblogs”: situering (vervolg)

- Ze beschikken over een dataset van een deel van de logs van de webserver (de maand maart van het jaar 2018). Er is één instantie per sessie met telkens verschillende attributen.
- Elke sessie werd manueel of via een rule-based systeem gelabeled als ROBOT = 0 of 1.

HO GENT

Features

ID	Unieke, geanonimiseerde ID van de sessie
NUMBER_OF_REQUESTS	Aantal requests tijdens de sessie
TOTAL_DURATION	Totale duurtijd van de sessie
AVERAGE_TIME	Gemiddelde tijd tussen twee opeenvolgende requests
STANDARD_DEVIATION	Standaarddeviatie van de tijd tussen twee requests
REPEATED_REQUESTS	Percentage van de herhaalde requests (t.o.v. totaal aantal requests)
HTTP_RESPONSE_2XX	Percentage van de requests met een succesvolle http status code (2xx)
HTTP_RESPONSE_3XX	Percentage van de requests met een redirection http status code (3xx)
HTTP_RESPONSE_4XX	Percentage van de met een client error http status code (4xx)
HTTP_RESPONSE_5XX	Percentage van de met een server error http status code (5xx)
GET_METHOD	Percentage van de requests met HTTP method GET
POST_METHOD	Percentage van de requests met HTTP method POST
HEAD_METHOD	Percentage van de requests met HTTP method HEAD
OTHER_METHOD	Percentage van de requests met een andere HTTP method
NIGHT	Percentage van de requests tussen middernacht en 7u
UNASSIGNED	Percentage van de requests met niet-toegekende referrer ("-")
IMAGES	Percentage van de image file requests
TOTAL_HTML	
HTML_TO_IMAGE	
HTML_TO_CSS	
HTML_TO_JS	
WIDTH	Breedte van de afgelegde "boom" in de URL-ruimte
DEPTH	Diepte van de afgelegde "boom" in de URL-ruimte
STD_DEPTH	Standaarddeviatie van de pagina-diepte tussen twee requests
CONSECUTIVE	Percentage opeenvolgende sequentiële HTTP requests
DATA	Totaal aantal getransfereerde bytes
PPI	(Popularity Index) Gemiddelde populariteitsindex van elke pagina gevonden in de sessie
SF_REFERRER	Switching Factor on unassigned referer field.
SF_FILETYPE	Switching Factor of file type.
MAX_BARRAGE	Maximum number of embedded resources in a web page.
PENALTY	Penalty for each backward and forward navigation or loop.
ROBOT	0 = mens, 1 = robot

Dataset

- https://github.com/HOGENT-ML/course/blob/main/datasets/weblogs.csv

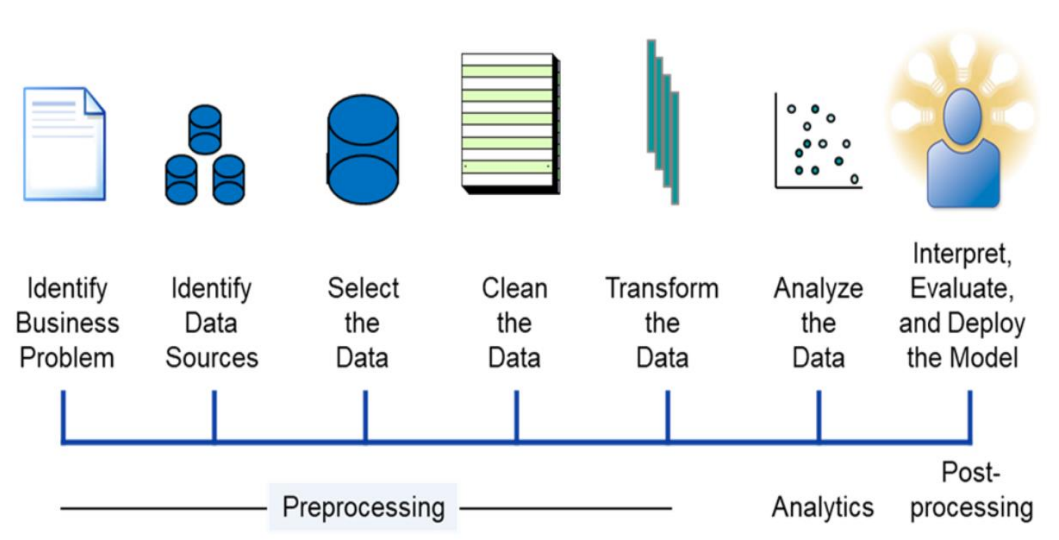
HO GENT

Taak

- Verken, zowel visueel als cijfermatig, de data met de bedoeling inzicht te verwerven in de natuur van de data.
- Probeer de data op te schonen (data cleaning!).
- Stel een classificatiemodel op voor de trainingsdata.
  - Kies een optimaal classificatiemodel door alle in de les geziene modellen uit te proberen en te vergelijken volgens de "best practices".
  - Kies verschillende maatstaven voor de nauwkeurigheid van je model en bespreek de voor- en nadelen van elke maatstaf.
  - Stel de “confusion matrix” op voor de 2 klassen en bespreek.
  - De website-beheerders willen het model gebruiken om robotsessies te cancelen, maar ze willen daarbij zo weinig mogelijk menselijke sessies per vergissing cancelen. Welke maatstaf is hierbij belangrijk?
- Mogelijkheid tot vraagstelling en werken aan het project wekelijks tijdens het vierde lesuur (vanaf week 2).
- Maak een Python-console-applicatie (script, geen notebook, zonder GUI) met als functionaliteit
  - Input = gegevens van nieuwe reeks sessies.
  - Output = klasse (mens, robot) met waarschijnlijkheidspercentage (vb. 60% mens, 40% robot) per sessie
  - De mogelijkheid tot hertrainen van het model bij aangevulde trainingsdata.

HO
GENT

Houd rekening met: het data mining proces



± 80-90 % van de totale projecttijd gaat naar preprocessing wegens:

- Bad veracity!
- Data en business probleem begrijpen is meestal niet eenvoudig.
- Algoritmes zelf zijn off-the-shelf beschikbaar (meestal geen PhD nodig)
- 80 % transpiratie, 20 % inspiratie

HO GENT

Evaluatie

- U werkt alleen.
- Presentatie van project in week 13 gedurende een korte mondeling toelichting.
- 25 % van de evaluatie in eerste zittijd.
- Geen tweede examenkans → punten project worden overgenomen.

HO GENT