Subject: Fwd: EINF-156 Automated Systematic Reviewing with active learning
Date: Thursday, 16 January 2020 at 10:52:24 Central European Standard Time
From: Schoot, A.G.J. van de (Rens)
To: Ferdinands, G. (Gerbrich), Schram, R.D. (Raoul), Ellen de Bruin

Begin doorgestuurd bericht:

Van: SURF Access Request Portal <NOREPLY@surfsara.nl>
Datum: 16 januari 2020 om 10:08:05 CET
Aan: "Schoot, A.G.J. van de (Rens)" <A.G.J.vandeSchoot@uu.nl>
Onderwerp: EINF-156 Automated Systematic Reviewing with active learning

Hello,

We would like to confirm that Your case had been recorded at SURF Access Request Portal with following references:

ID: EINF-156
Requestor: Rens van de Schoot

Details:

Request type
access to compute

Title(s)
prof. dr.

First name:
Rens

Last name:
van de Schoot

E-mail address:
a.g.j.vandeschoot@uu.nl

ORCID iD
https://orcid.org/0000-0001-7736-2091

Phone:
[redacted]

Function:
hoogleraar

Institute:
Universities - Universiteit Utrecht

(If applicable) Department:
Methods and Statistics

For this application, I am the Signing Authority
Yes

Signing Authority Staff Position within the organization
Hoogleraar

Title of the project
Automated Systematic Reviewing with active learning

Scientific Area
Informatics

Scientific project description
Scholars are confronted with ever-larger amounts of textual data. All this data present new and unique opportunities to scholars, while simultaneously confronting them with unprecedented challenges. How to select relevant text effectively and efficiently from an almost unlimited amount of data? Conducting a systematic review on this data is often a very time consuming and tedious task. Reviewers have to manually scan thousands of abstracts of scientific articles and assess their relevance to the research question at hand. For an experienced reviewer, it takes between 30 seconds to a couple of minutes to classify a single abstract, which easily results in hundreds of hours spent merely on abstract screening. In this day and age when the field of artificial intelligence is thriving at an unprecedented pace, one would imagine that this large amount of manual work can be reduced or even completely replaced by some smart machine learning software. The AI pipeline that we propose to develop will include the newest insights from the field of deep learning combined with active learning. The ultimate goal of using our software, titled ASReview, is to reduce the number of abstracts and potentially full-texts which have to be screened by the reviewer, thereby expediting the reviewing process.
To test the performance of our AI-pipeline we need to run simulations on different datasets with varying parameters such as statistical and 3 machine learning models (Naïve Bayes, Support Vector Machine, Random Forest), seven query strategies (e.g., random, most probably, max-pool), 2 feature extraction strategies (TFIDF, Doc2Vec), and 3 levels of amount of training data (5 inclusions + 5 exclusions; 10 + 10; 10 + 5). This leads to a total amount of at least 210 different combinations of parameter configurations.
Such simulations are impossible to run on our local devices and we would like to make use of cluster computing.

Technical project requirements
A typical simulation of 10 datasets with 10 runs (with 4 cores per run) will take approximate 1 day and uses 8Gb of memory. It uses 1Gb size of input dataset and generates and 200Gb size of output data. The application makes use of Python, Tensorflow, and related packages to execute the runs. The total amount of runs to complete is 210*10*10=21,000. The application runs are dependent to each other.

Can we make this information public on the SURFsara website?
Yes

Best-fit service(s) and respective resources.
Cartesius

Cartesius: how much SBU? (Max. 500.000)
500

Cartesius: how much project space in Terabyte? (Max. 50 TB)
1

Cartesius: Do you need access to GPU's?
No

Service Preference
Basic support

Support
We would like to make use of Cartesius because of the highly parallel setup of the simulation (embarrassingly parallel). The simulation study can benefit from multiple cores and machines. We have experienced engineers in the team who will work with us on this project.

Start of the project (at SURFsara):
3/Feb/20

End of the project (at SURFsara):
31/Dec/20

Created
16-01-2020 10:06

Current status of your case is: **Submitted**

Kind Regards

SURF Compute Access Team | Science Park 140 | 1098 XG Amsterdam | T +31 (0) 20 800 1300 | https://surf.nl |

We are ISO 27001 certified and meet the high requirements for information security.

View request · Turn off this request's notifications