Project Dataset Description (COSC 320 2022/2023)

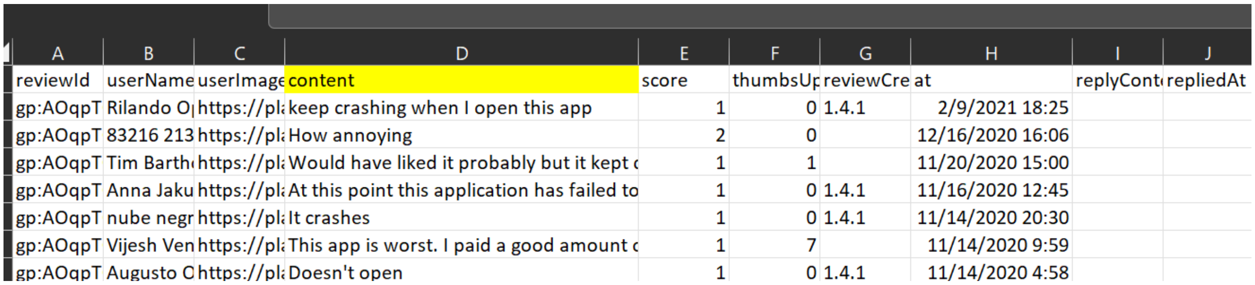
You can find the link to the dataset I mentioned for Topic 1 of the projects below.

Project-Dataset

Note that you need to access it through your UBC email address.

For those working on Topic 2, you can use the same dataset and write a script to copy some parts of the records and then use it as the copied records (or combination of them) as the plagiarism dataset. Then both of these datasets can be used for your plagiarism detection algorithms.

The dataset includes separate csv files, each about one app. You need to combine all of the csv files into one file and then run your algorithms. Each csv file looks like below.



A	B	C	D	E	F	G	H	I	J
reviewId	userName	userImage	content	score	thumbsUp	reviewCre	at	replyCont	repliedAt
gp:AOqpT	Rilando O	https://pl	keep crashing when I open this app	1	0	1.4.1	2/9/2021 18:25		
gp:AOqpT	83216 213	https://pl	How annoying	2	0		12/16/2020 16:06		
gp:AOqpT	Tim Barth	https://pl	Would have liked it probably but it kept	1	1		11/20/2020 15:00		
gp:AOqpT	Anna Jaku	https://pl	At this point this application has failed to	1	0	1.4.1	11/16/2020 12:45		
gp:AOqpT	nube negr	https://pl	It crashes	1	0	1.4.1	11/14/2020 20:30		
gp:AOqpT	Vijesh Ven	https://pl	This app is worst. I paid a good amount	1	7		11/14/2020 9:59		
gp:AOqpT	Augusto C	https://pl	Doesn't open	1	0	1.4.1	11/14/2020 4:58		

You need to work with the column named “content” as this is the app reviews written by the users.

Important note about licensing:

You are not allowed to share this dataset with anyone out of this class or publish it online. Any other uses other than the term projects for COSC 320 (2022-2023) is not allowed. This is part of the dataset my students have collected and are working on as their graduate thesis.