

Capítulo 3

Conjunto de dados

Os ignorantes afirmam, os sábios duvidam, os sensatos refletem.

Aristóteles

O conteúdo deste capítulo foi publicado em forma de artigo na conferência “XIII Simpósio Brasileiro de Tecnologia da Informação – SBTI” com o título “NEODATASET - Um conjunto de dados com User Stories e Story Points” [79] e posteriormente selecionado pela e publicado na “RMP (Revista dos Mestrados Profissionais)” (disponível no apêndice B)

Resumo

CONTEXTO: As equipes geralmente utilizam ferramentas de gerenciamento para acompanhar as User Stories pendentes, controlar o seu código-fonte, registrar suas estimativas de esforço e os responsáveis pela abertura e fechamento dos chamados. Essas ferramentas contêm dados que podem ser utilizados em diversas pesquisas de engenharia de software. É desafiador encontrar dados para pesquisas, pois as empresas privadas são relutantes em compartilhar seus dados.

OBJETIVO: O objetivo deste capítulo é apresentar um conjunto de dados contendo dados brutos de 34 Projetos de Software Ágil de código aberto, minerados do GitLab [40], totalizando 163.897 Story Points e 40.014 Tarefas.

CONCLUSÃO: Foram disponibilizados esses dados publicamente nos formatos CSV

e JSON para facilitar o seu uso pela comunidade científica interessada. Acredita-se que esse conjunto de dados pode ser utilizado em várias linhas de pesquisa de engenharia de software, incluindo estimativa de esforço.

3.1 Introdução

Apesar da necessidade de dados consolidados em pesquisas, as empresas privadas são relutantes em divulgá-los. Além disso, nem sempre eles estão em um formato de fácil acesso para os pesquisadores. O processo de consolidação e preparação dos dados, para transformá-los em informações e *insights*, ainda exige considerável esforço, pois geralmente os dados são armazenados em bancos de dados relacionais.

Para contribuir com pesquisas de engenharia de software, e inspirados pela contribuição de outros conjuntos de dados [57, 124], foi disponibilizado um novo conjunto de dados (chamado NEODATASET). Esse conjunto de dados abrange dados de 34 projetos, com 163.897 tarefas (ou *issues*) retirados de repositórios do GitLab [40], totalizando 40.014 Story Points. Ele é disponibilizado no GitHub para que toda a comunidade interessada possa contribuir, semelhante ao que acontece com outros conjuntos de dados [131].

As seções seguintes apresentam uma descrição do conjunto de dados, como ele foi extraído, suas características e estrutura, a sua originalidade e relevância e as considerações finais.

3.2 Proposta

3.2.1 Extração

Este conjunto de dados foi extraído durante os meses de janeiro de 2023 e abril de 2023. O processo de mineração teve como alvo os principais projetos de código aberto do GitLab [40]. Os projetos selecionados empregam metodologias ágeis de desenvolvimento de software e tiveram registrado o tamanho em Story Points de suas tarefas.

Para minerar informações do GitLab [40] foi criada uma ferramenta de extração implementada em Python que se conecta ao GitLab via *API*, o código dessa aplicação está

disponível no GitHub.

Foram coletadas somente User Stories com o atributo *State* (situação): *Closed* e que tenham o atributo *weight* (esse campo, no GitLab [40], é utilizado para registrar o esforço em Story Points) preenchido. Mais informações sobre os projetos incluídos no conjunto de dados também estão disponíveis diretamente no *GitLab*.

3.2.2 Armazenamento

O conjunto de dados foi armazenado no formato JSON e CSV, dada a simplicidade de lidar com ambos os formatos, e ambos estão disponíveis no GitHub.

3.2.3 Característica

O conjunto de dados NEODATASET contém um total de 163.897 tarefas oriundas de 34 projetos, totalizando 40.014 *Story Points*. Os projetos possuem diferentes características, abrangem diferentes linguagens de programação, diferentes domínios de negócio e diferentes localizações geográficas da equipe. A Tabela 3.1 apresenta informações resumidas do conjunto de dados, tais como o nome do Projeto, o Total de Tarefas e o total de Story Points. A coluna ID da Tabela apresenta o ID do projeto no site do Gitlab [40]. O Total de Tarefas é o total de *User Stories* daquele projeto que está com a situação de *Closed*. O Total SP de um Projeto é o somatório dos *Story Points* de todas as Tarefas informadas pelos usuários daquele projeto.

3.2.4 Estrutura

A Tabela 3.2 apresenta uma descrição dos principais atributos da entidade Tarefas; outros campos foram omitidos.

Um exemplo dos dados do conjunto de dados está na Tabela 3.3. Nessa tabela são apresentados somente os campos título, descrição e Story Point, outros campos foram omitidos.

A Figura 3.1 mostra um Diagrama de Classes que representa visualmente a estrutura JSON do conjunto de dados. A entidade principal é Tarefa (no diagrama, *Issue*), que contém as principais informações. O conjunto de dados tem mais de 70 atributos.

Tabela 3.1: Estatísticas descritivas dos dados do NEODATASET.

ID	Projeto	Total Tarefas	Story Points							
			Total SP	Média	STD	Mín	25%	50%	75%	Max
10152778	Minds	521	1594	3,1	4,2	0	1	2	3	80
10171280	Minds Mobile	2796	7213	2,6	2,7	0	1	2	3	50
12894267	MLReef	285	1419	5,0	5,0	0	2	4	6	40
3828396	GitLab Chart	15	31	2,1	1,2	1	1	2	2	5
278964	GitLab	19548	41270	2,1	1,8	0	1	2	3	160
6206924	Tildes	42	91	2,2	1,0	1	1,25	2	3	4
12584701	StackGres	171	1106	6,5	7,4	1	2	4	8	48
7764	Gitlab.com	355	970	2,7	7,2	0	1	2	3	128
12450835	Duplicity	424	5798	13,7	22,1	4	6	6	12	260
28847821	Buyer Experience	1004	2665	2,7	2,0	0	1	2	4	14
10174980	Veloren	178	502	2,8	2,8	1	1	2	3	15
250833	Gitlab Runner	13	35	2,7	2,2	1	1	2	3	9
23285197	Subway	284	507	1,8	1,5	0	1	1	2	13
1304532	Gitlab GL Infra Reliability	1724	4520	2,6	2,5	0	1	2	3	21
14052249	Mythictable	167	509	3,0	2,9	1	2	2	3	20
28419588	Lazarus	144	17800	123,6	48,7	100	100	100	100	300
4456656	Pajamas Design System	344	684	2,0	1,6	0	1	1	3	13
3836952	Tezos	103	2771	26,9	38,6	0	2,5	5	32	100
7603319	Meltano	237	1107	4,7	5,3	0	1	4	8	40
7776928	Triage ops	216	442	2,0	1,2	1	1	2	2	10
2670515	Customers gitlab com	1574	3137	2,0	1,2	0	1	2	2	15
21149814	Opengeoweb	1942	5668	2,9	1,5	1	2	3	3	20
15502567	Kicad	3284	34692	10,6	14,2	4	6	6	10	268
1714548	Petals Vockpit	154	400	2,6	1,5	1	1	2	3	8
7128869	Nlx	327	1355	4,1	3,6	0	2	3	5	21
10171263	Minds Backend Engine	982	3742	3,8	3,7	0	2	3	5	32
14976868	Database Lab Engine	113	721	6,4	7,6	1	2	4	8	42
2009901	Gitaly	171	401	2,3	1,5	1	1	2	3	13
28644964	FPC Source	102	14200	139,2	63,2	100	100	100	200	300
7071551	Gitlab UI	310	600	1,9	2,0	0	1	2	2	32
734943	Gitlab Pages	121	265	2,2	2,6	1	1	1	3	20
10171270	Minds Frontend	1845	5695	3,1	3,2	0	1	2	4	40
5261717	Gitlab vscode extension	106	185	1,7	0,7	1	1	2	2	4
19921167	Glaxnimate	420	1802	4,3	4,0	1	2	3	5	40
Total		40.022	163.897							

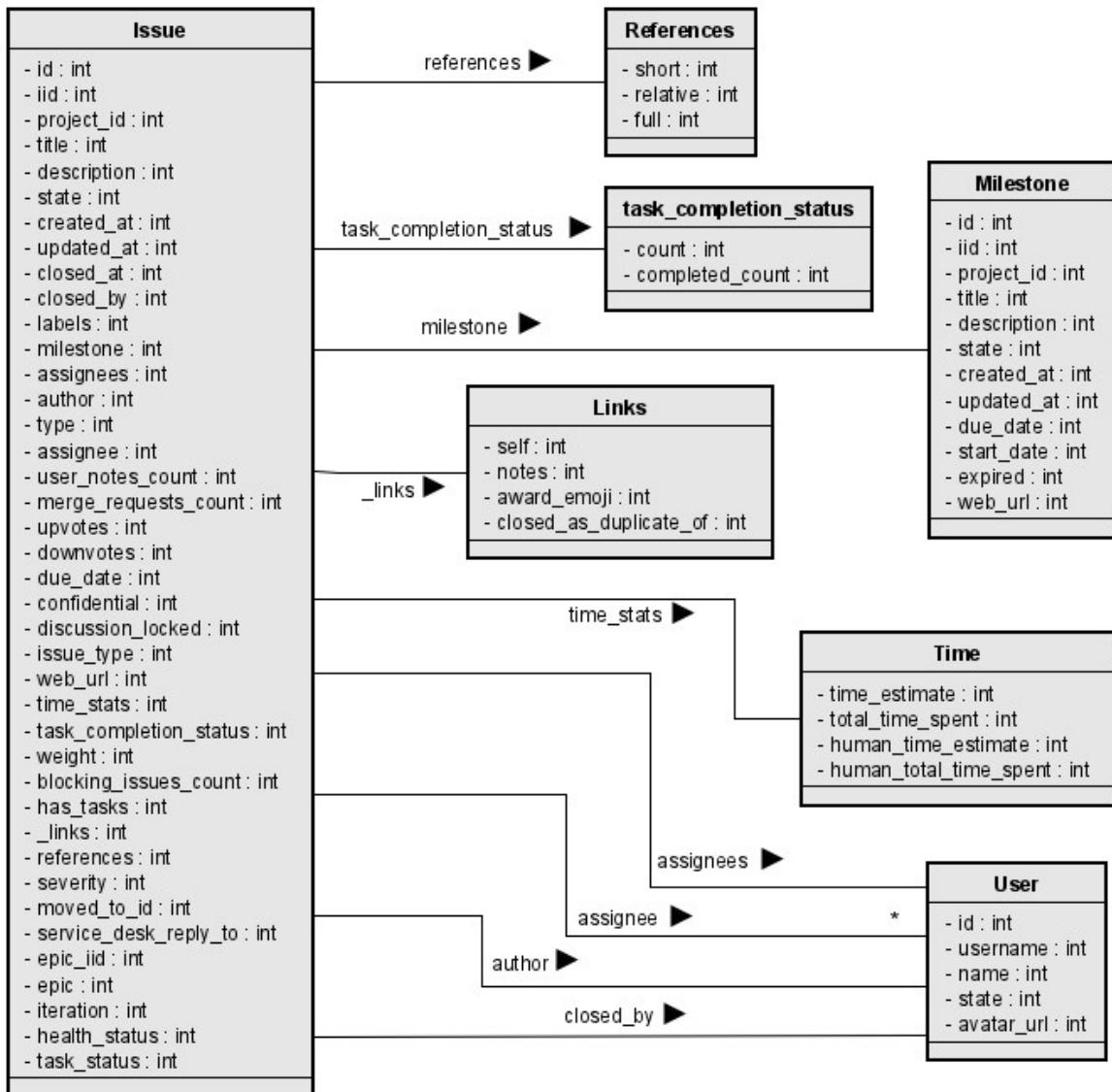
Tabela 3.2: Descrição dos principais atributos.

Nome	Descrição
Título	Título da User Story
Descrição	Descrição da User Story
SP (ou Weight)	Story Points estimados

Tabela 3.3: Exemplo dos dados.

Título	Descrição	SP
Posts repeat in Groups feed	We have a few reports of posts repeating in groups. [This Gitlab issue](url...) makes reference to the issue in a test Group. Additionally, a user alerted me that [this Group](url...) exhibits the issue, though notably the posts repeat only after scrolling VERY deep into the timeline (~100+ posts). The pinned comment reappears in the feed, after which the feed appears to loop. Note timestamps on the posts in this sequence: !image(url...) !image(url...) !image(url...) ### Replication steps Effect can be seen here: (url...) To replicate a base-case: 1. Make a new group 2. Make a post titled 1, and a post titled 2 3. Refresh	5
Expose bot accuracy scores on front end for admins	In order to (a) help with making admin decisions, and (b) QA check the scores so that admins can see what Tasman is spitting out and provide feedback on the effectiveness of the current scoring, we'd like to make Tasman-derived trust scores visible on the front end of Minds website for users with Admin access. (url...) >Designs are in progress - Given Tasman trust scores for users are available, - and UserA is logged into Minds, - and UserA's account has Admin access, - when any user avatar + name component is displayed, - then a number is displayed alongside the avatar + name that exposes the Tasman trust score for that user. Proposed approach for displaying the correct colour based on a 0-100 value ->(url...) ### Notes * Uniquely an admin feature for the first iteration * In the future this will be visible to all users, including an explainer (via a modal) about the score, why scores are valuable and how to improve scores.	3
Refactor experiments to use events instead of contexts	## The Problem Currently all users are bucketed into experiments on app initialisation and not when they actually see the experiment. This was done so that pageviews could register the experiments as contexts when someone lands on the site. Issues arise when experiment reports rely on only including users who have seen the experiment to be included. For example, an experiment that forwards users to discovery instead of the newsfeed after registration should not just only include users that have just signed up, it should also only apply to users that do not have any referral logic (ie. take newlyloggedin/registered users back to previous page). ## Proposed change - [x] Create a new iglu schema 'growthbook_event' which includes the experiment_id that the user is in - [x] Remove backend registration and client on init logic - [x] Record a 'growthbook_event' via snowplow when an experiment is run - [x] To avoid unnecessary events, cache a reference to the events so that they only get sent every 24 hours. - [x] Now that we support psuedo_id, attach both the anonymous_id and user_id to growthbook events	8

Figura 3.1: Diagrama de classes do arquivo JSON representando sua estrutura.



3.3 Originalidade e relevância

O conjunto de dados aqui apresentado foi minerado da ferramenta de gerenciamento *GitLab* [40], e inclui projetos adicionais, que não foram utilizados por nenhum desses estudos anteriores. Já existem estudos anteriores que extraíram dados da ferramenta de gerenciamento Jira [8] para construir modelos preditivos [128, 18, 93, 109], mas projetos extraídos do *GitLab* [40] são mais raros.

Assim como fez [124], compartilho todos os dados coletados. Pois, o mais comum é compartilhar apenas os dados do conjunto de dados considerados no próprio estudo, como fez, por exemplo: [18], e não todos os dados coletados.

A contribuição esperada é que este conjunto de dados possa auxiliar pesquisas na área de ASD. Embora o conjunto de dados tenha sido projetado inicialmente para a pesquisa de estimativa de *Story Points* em ASD, ele também inclui informações relevantes para outras pesquisas de engenharia de software. Além de fornecer uma possibilidade para reproduzir achados de outros estudos.

3.4 Disponibilidade dos artefatos

- O código Python do extrator está disponível em <https://github.com/giseldo/neo-gitlab-extractor>;
- O conjunto de dados produzido (CSV e JSON) está disponível em <https://github.com/giseldo/neodataset>;
- O conjunto de dados disponível no HuggingFace pode ser acessado no endereço: <https://huggingface.co/datasets/giseldo/neodataset> e no Mendeley Data <https://data.mendeley.com/datasets/skk2wn9j86/1>.

3.5 Considerações finais

Nesse capítulo apresentou-se o NEODATASET, um conjunto de dados com projetos minerados do *GitLab* [40], para ser utilizado em pesquisas que exploram vários problemas

do desenvolvimento de software ágil. Como tal, ele será utilizado nos próximos capítulos.