Datasheet

Monika Leszniewska, Józef Piechaczek, Nazar Pokutycki oraz Krzysztof Szafraniak

Politechnika Wrocławska

1 Pliki danych

- step_1_1_stack_users_in.csv
- step_1_2_stack_users_out.csv
- step_1_3_stack_users_out_fs.csv
- step_2_github_data.json
- step_2_github_data_fs.json
- step_3_processed_ground_truth.csv
- step_3_processed_ground_truth_fs.csv

2 Motywacja

Przedstawione zbiory danych zostały zgromadzone podczas realizacji publikacji na temat rozpoznawania technicznych ról użytkowników. Wykorzystano je w celu stworzenia modeli uczenia maszynowego. Finalne zbiory danych, których nazwa rozpoczyna się od prefiksu *step_3* zawierają zestaw cech i *Ground truth* na temat ról użytkowników. Pozostałe dane są reprezentują różne etapy gromadzenia i przetwarzania danych. Źródłem danych zawartych w pliku *step_1_1_stack_users_in.csv* jest platforma StackExchange. Pozostałe pliki zostały wygenerowane przez zespół badawczy w którego skład wchodzą: Monika Leszniewska, Józef Piechaczek, Nazar Pokutycki oraz Krzysztof Szafraniak. Są one wynikami działania wykonywanych algorytmów, przy czym należy zaznaczyć, że pliki *step_2_github_data.json* i *step_2_github_data_fs.json* są zależne od informacji zwracanych przez API platformy GitHub.

3 Opis danych

Plik *step_1_1_stack_users_in.csv* jest podzbiorem informacji zawartych w zrzucie danych portalu StackOverflow, zawierający informację na temat wszystkich użytkowników portalu. Każdy wiersz zawiera identyfikator. wskaźnik reputacji, datę założenia konta, nazwę użytkownika, ostatnią datę logowania, link do prywatnej strony, lokalizację, treść sekcji AboutMe, liczbę wyświetleń profilu, liczby oddanych pozytywnych i negatywnych głosów, link do zdjęcia, hash adresu email oraz identyfikator profuilu Stack Exchange Network. Pierwszy wiersz zbioru danych zawiera nagłówek, pozwalający zidentyfikować odpowiednie wartości. Zawarte w zbiorze informacje mogą być niekompletne ze względy na ich

nieobecność w profilu użytkownika. Kolumna zawierająca treść sekcji AboutMe nie jest w żaden sposób cenzurowana. Zawartych informacji nie da się bezpośrednio powiązać z konkretnymi osobami, jednak posiadając inne zbiory danych potencjalnie możliwe jest rozszerzanie informacji o użytkownikach na podstawie linku do profilu prywatnego lub wartości skrótu adresu email.

Plik *step_1_2_stack_users_out.csv* zawiera przetworzone informacje z opisanego powyżej zbioru. W jego wyniku powstał zbiór danych opisujący role użytkowników. Każdy wiersz zawiera identyfikator użytkownika StackOverflow, nazwę konta GitHub oraz tablicę ról, które zostały mu przypisane na podstawie danych zawartych w pliku *step_1_1_stack_users_in.csv*. Tablica jest niepustym podzbiorem zestawu rozpoznawanych ról, czyli: *Frontend*, *Backend*, *DevOps*, *Mobile* oraz *DataScience*. W plikiu znajuje się 1871 wierszy opisujących role poszczególnych użytkowników. Wśród opisywanych zbiorów danych znajduje się bliźniaczy plik *step_1_3_stack_users_out_fs.csv*. Tym co je rozróżnia jest rozszerzenie rozpoznawanych ról o wartość *FullStack*. Skutkiem tego jest również różna liczba wierszy danych która w tym przypadku wynosi 2770. W szczególności zbiór nazw użytkowników w pierwszym pliku jest podzbiorem tych z drugiego.

Pliki *step_2_github_data.json* i *step_2_github_data_fs.json* są zbiorami informacji pozyskanych z portalu GitHub dla użytkowników, których nazwy znajdują się w plikach kroków 1_2 oraz 1_3. Dane są podzielone na 4 grupy. Jedną z nich jest lista wejściowych nazw użytkowników, których nie udało się dopasować do żadnego profilu na platformie GitHub. Inną jest lista użytkowników którzy posiadają mniej niż 5 repozytoriów, trzecią - ci, dla których nie udało się zebrać danych w wyniku nieoczekiwanych błędów. Ostatnią grupą są użytkownicy dla których pomyślnie zebrano dane. Informację które pozyskano obejmują krótki opis profilu, a dla każdego z posiadanych repozytoriów: jego nazwę, opis, listę zależności projektowych, tematy, główny język programowania, całkowitą liczbę rewizji oraz liczbę tych wykonanych przez właściciela repozytorium. Poszczególne grupy dla pliku bez roli *"FullStack"* liczą odpowiednio: 74, 225, 0, 1567, natomiast dla drugiego pliku: 112, 314, 0, 2337. Ich sumy różnią się od długości wejściowych list nazw użytkowników ze względu na pojawiające się duplikaty. Zawarte dane nie pozwalają na bezpośrednie zidentyfikowanie tożsamości użytkownika.

Dwa pozostałe pliki zawierają informację gotowe do wykorzystania w procesie uczenia maszynowego. Ponownie rozróżnia je uwzględnienie roli *FullStack* przez co ilość wierszy jest też różna i w *step_3_processed_ground_truth.csv* wynosi 1567, a dla *step_3_processed_ground_truth_fs.csv*: 2337 - zgodnie z licznością ostatniej grupy omówionej w poprzednim paragrafie. Kolumny dzielą się na dwie części. Pierwsza z nich to zestaw cech. Są nimi różnego rodzaju dane pozyskane na podstawie różnych źródeł (poszczególnych danych z pliku *step_2_github_data.json*). Źródła te zostały powiązane z cechami dekorując ich nazwy odpowiednimi skrótami umieszczonymi w nawiasach. Drugą część w zależności od pliku stanowi 5 lub 6 kolumn, z której każda reprezentuje jedną z ról technicznych użytkownika. Wartość "0óznacza, że na podstawie danych z

StackOverflow danemu użytkownikowi nie przypisano konkretnej roli (bazując na informacjach pochodzących z kroków 1_2 oraz 1_3).

4 Wstępne przetwarzanie danych

W pierwszym kroku pozyskano informację o użytkownikach, którzy zawarli na swoim profilu link do GitHuba. W tym celu użyto narzędzia Stack Exchange Data Explorer[**?**]. Pozwala ono na uzyskanie informacji zebranych na poszczególnych podstronach serwisu Stack Exchange za pomocą zapytań SQL.

```
select *
from Users
where WebsiteUrl like '%github%'
```

Powyższe zapytanie zwróciło 30990 użytkowników, a otrzymane wyniki zostały zapisane w pliku *step_1_1_stack_users_in.csv*. W kolejnym kroku odrzucono użytkowników z pustą sekcją *AboutMe*. Operacja ta zostawiła 17970 użytkowników w zbiorze. W następnej operacji użyto wyrażeń regularnych w celu uzyskania nazwy konta na GitHubie. Poniżej przedstawiono użyte wyrażenia.

```
/https?:\/\/www\.([^\/\.]+)\.github\.io.*/i
/https?:\/\/([^\/\.]+)\.github\.io.*/i
/https?:\/\/github\.com\/([^\/\.]+)/i
/https?:\/\/www\.github\.com\/([^\/\.]+)/i
```

Wykonanie tej operacji pozwoli w kolejnych krokach powiązać zebrane *Ground truth* z odpowiednimi kontami na platformie GitHub. Ostatnią wykonaną operacją było odnalezienie ról użytkowników oraz odrzucenie osób bez żadnej roli. W tym celu użyto wyrażeń regularnych przedstawionych w tablicy 1.

Tablica 1. Wyrażenia regularne dla uzyskiwania ról

Rola	Wyrażenie
Frontend	`{ /.*front.0,1end.*/i}`
Backend	`{/.*back.0,1end.*/i}`
DevOps	`{/.*dev.0,1ops.*/i}`
DataScience	`{/.*data.0,1scientist.*/i}`
Mobile	`{/.*mobile.*/i}`
FullStack	`{.*full.0,1stack.*}`

Wyniki tych operacji zostały zapisane do plików *step_1_2_stack_users_out.csv* oraz *step_1_3_stack_users_out_fs.csv*. Przy czym pierwszy z nich nie uwzględnia roli FullStack.

Bacując na zebranej w poprzednim kroku prawdzie - dokładniej na uzyskanej w nim liście nazw użytkowników - wykonano dla nich procedurę pobrania

informacji o kontach i posiadanych repozytoriach na portalu GitHub. Wśród nich są nazwy, opisy i tematy w których użytkownicy często zamieszczają słowa kluczowe charakteryzujące zawartość repozytorium. Wydobywane są również informacje o językach programowania oraz udziale użytkownika w rozwoju danego projektu. Do tego celu wykorzystywana jest historia domyślnej gałęzi repozytorium. Przetwarzane są dane dotyczące całkowitej liczby rewizji (ang. commit) w projekcie, a także liczby tych wykonanych przez jego właściciela. Ważne informacje - ze względu na rozpoznawanie ról - może dostarczać lista zależności projektowych, które często różnią się w zależności od typu oprogramowania i wykorzystywanego języka. Poza danymi bazującymi na repozytoriach pobierany jest krótki opis profilu użytkownika, mogący zawierać hasła określające obszar jego działania. Pozyskanie tych informacji odbyło się przy użyciu skryptu wykorzystującego interfejs GraphQ, który wraz z zapytaniem znajdują się w źródłach dołączonych do badań.

W wyniku działania programu, przetwarzana lista nazw użytkowników, została podzielona na kilka grup. Pierwszą z nich były duplikaty, czyli powtarzające się w danych wejściowych wartości. Kolejnym zbiorem były ciągi tekstowe, które nie stanowiły nazwy użytkownika dla żadnego z indywidualnych kont na portalu Github. Do tej grupy wliczały się również nazwy istniejących kont organizacji. Warunkiem kwalifikującym użytkownika do dalszego etapu analizy, jest posiadanie co najmniej pięciu repozytoriów. Użytkownicy niespełniający wymienionego kryterium zostali przydzieleni do odrębnej grupy. Warto zaznaczyć, że repozytoria bazujące na innym projekcie (ang. forked) zostały pominięte i nie zostały wykorzystane do określenia liczby repozytoriów oraz nie są przetwarzane w dalszej analizie. Pozostałą grupę stanowią użytkownicy dla których proces przebiegł pomyślnie oraz spełnili wymaganie ilości repozytoriów. Do tej grupy zakwalifikowano również tych, dla których z przyczyn błędów nie udało się pobrać danych o części z ich repozytoriów. Repozytorium dla którego wystąpił błąd zostało całkowicie odrzucone w procesie i nie wliczano go do liczby projektów kwalifikujących do dalszego etapu. Te rezultaty zostały uwiecznione w pliku *step_2_github_data.json*.

Ostatni krok przetwarzania ma na celu scalenie zebranych w pierwszym oraz drugim kroku informacji oraz ich przygotowanie do wykorzystania w procesie nauczania maszynowego. Aby uzyskać listę cech, dane z GitHub podzielono na kategorie:

- *Informacje o użytkowniku:* Dla każdego z użytkowników zestawiono informacje dotyczącego jego nazwy oraz opisu profilu. Dane te zestawiono w ramce, po czym usunięto słowa znajdujące się na stop-liście dla słownika języka angielskiego oraz zastosowano algorytm *bag-of-words*.

- *Informacje o repozytoriach:* W tej kategorii zestawiono w ramce nazwę użytkownika, nazwę, opis oraz tagi repozytorium. Listę tagów połączono w jeden ciąg znaków, oddzielając spacją poszczególne tagi. Dla każdej z wymienionych informacji, z wyłączeniem nazwy użytkownika, wykonano przetwarzanie analogiczne do poprzedniej grupy - usunięto słowa ze stop-listy oraz użyto

algorytmu *bag-of-words*.

– *Informacje o językach:* Zebrane informacje o głównych językach umieszczono w zbiorze, a następnie dla każdego z nich określono łączną liczbę commitów utworzonych przez danego użytkownika, całkowitą liczbę commitów oraz stosunek pomiędzy nimi. Ze względu na dużą liczbę uzyskanych cech odrzucono języki o wysokim współczynniku korelacji Spearmana.

– *Informacje o bibliotekach:* W początkowej ramce zestawiono nazwę użytkownika wraz z nazwami bibliotek, używanych we wszystkich jego repozytoriach. Aby zredukować długość listy ograniczono się do 1000 najczęściej używanych bibliotek. W kolejnym kroku odrzucono repozytoria o wysokim współczynniku korelacji Spearmana.

Cechy uzyskane we wszystkich kategoriach oraz informacje o rolach umieszczono w jednej ramce, indeksowanej nazwami użytkowników. Ramkę tą następnie wyeksportowano do plików CSV (*step_3_processed_ground_truth.csv* *step_3_processed_ground_truth_fs.csv*), będących wejściowymi dla algorytmów nauczania maszynowego.

5 Wykorzystanie danych

Omawiane zbiory danych zostały wykorzystane w badaniach dotyczących detekcji technicznych ról użytkowników bazując na informacjach pochodzących z kont i repozytoriów GitHub. Podjęto próbę odpowiedzenia na pytania "Jak dokładne są klasyfikatory uczenia maszynowego w identyfikowaniu technicznych ról użytkowników?", "Jakie cechy są najbardziej odpowiednie do rozróżnienia ról technicznych?", "Czy role techniczne wpływają na siebie nawzajem podczas klasyfikacji?" oraz "Jak skutecznie możemy zidentyfikować programistów Full-Stack?". Szczegółowe informacje dotyczący tych badań znajdują się w dokumencie je opisującym.

6 Źródła

Zgromadzenie powyższych danych wymagało wykorzystana dwóch zewnętrznych źródeł. Pierwszym z nich jest platforma StackExchange umożliwiająca dostęp do danych o użytkownikach portal StackOverflow. Dane są ogólno dostępne oraz nie wymagają specjalnych uprawnień. Drugim źródłem jest interfejs w wersji 4. udostępniany przez platformę GitHub. Aby z niego skorzystać w wymaganym zakresie konieczne jest posiadanie bezpłatnego konta oraz wygenerowanie odpowiedniego klucza. Wspomniany klucz pozwala pobierać dane z limitem 5000 punktów na godzinę. Z tego powodu gromadzenie danych z tej platformy jest procesem podzielonym na części.

7 Wsparcie

Zebrane dane stanowią kompletny zestaw danych potrzebny do reprodukcji prowadzonych badań. Stanowią jednorazową migawkę z życia dynamicznie zmieniających się danych na portalach GitHub oraz StackOverflow. Obecnie nie ma planów, aby opisane zbiory były aktualizowane, jednak źródła powiązanych z nimi badań pozwalają na samodzielne ich zgromadzenie.