Dokumentacja Storyboardu

Bartłomiej Lechowicz

Dataset: **10000 Most Common Passwords**
Źródło: https://www.kaggle.com/shivamb/10000-most-common-passwords
Licencja: CC0: Public Domain

Dataset wchodzi w skład SecLists, który jest zbiorem narzędzi do testowania bezpieczeństwa zabezpieczeń komputerowych i zawiera 10000 najczęściej używanych haseł.

Wybór datasetu:
Postanowiłem wybrać ten dataset z kilku powodów. Pierwszym z nich jest fakt, że na co dzień jesteśmy otoczeni komputerami, poświęcamy bardzo dużo czasu na interakcje w środowisku cyfrowym. Otaczają nas hasła, loginy i maile, to nasze karty dostępu o których rzadko myślimy. Jednym z pierwszych etapów tworzenia kont, z których tak często korzystamy jest utworzenie hasła. Niektóre z serwisów wymagają stworzenia hasła zawierającego określoną liczbę i rodzaj znaków, jednak nic nie stoi na przeszkodzie żeby użytkownik wykorzystywał to samo hasło i maila jako dane dostępowe do wielu witryn. W wyniku coraz popularniejszych ataków hakerskich wiele haseł trafia w ręce cyberprzestępców, którzy próbują uzyskać dostęp do kont (w tym kont bankowych) ofiar.

Wydaje mi się, że temat bezpieczeństwa danych i kont nie jest jeszcze szeroko zakorzeniony wśród społeczeństwa, a prywatne informacje, zdjęcia i oszczędności ludzi są ciągle zagrożone. Drugim powodem była moja chęć wdrożenia się w użytkowanie biblioteki Pandas w środowisku Jupyter Notebook dla Pythona 3. Dokonałem tam analizy danych w datasecie, która wymagała przetwarzania języka naturalnego. Ostatnim powodem, który przekonał mnie do wyboru tego datasetu był fakt, że na 13. miejscu listy haseł zobaczyłem ciąg znaków, który odblokowuje wszystkie komputery w laboratoriach, w którym mieliśmy zajęcia, a problem stosowania za słabych haseł i ich wycieków wcale nie ustaje, co pokazują portale takie jak niebezpiecznik.pl.

Celem analizy była ocena siły haseł, z uwzględnieniem ich długości oraz znalezienie standardowych schematów i częstości występowania konkretnych schematów, które stosują ludzie przy tworzeniu haseł. Analiza danych i storyboard powstały na jej skutek są czysto opisowe - skupiłem się na znalezieniu i opisaniu już istniejących cech zbioru danych.

Link do repozytorium

Pliki:

1. Plik źródłowy głównego datasetu: ‘common_passwords.csv’ liczący 10000 rekordów i 9 kolumn:
    1. password - hasło
    2. length - długość hasła
    3. num_chars - ilość liter w haśle
    4. num_digits - ilość cyfr w haśle
    5. num_upper - ilość wielkich liter w haśle
    6. num_lower - ilość małych liter w haśle

7. num_special - ilość znaków specjalnych w haśle
8. num_vowels - ilość samogłosek w haśle
9. num_syllables - ilość sylab w haśle

Pliki dodatkowe:

1. ‘bad_words.csv’ - lista 1617 słów sklasyfikowanych jako przekleństwa lub wyrażenia obraźliwe w ramach projektu Toxic Comment Classification Challenge. Licencja CC0: Public Domain.
2. 'baby names-clean.csv' - lista 6781 amerykańskich imion pochodząca z danych Social Security Administration US. Licencja CC0: Public Domain.

Przygotowanie danych:
Plik źródłowy ‘common_passwords.csv’ oraz pliki dodatkowe zostały zaimportowany do środowiska Anaconda Jupyter Notebook. Korzystając z metody str.match pochodzącej z biblioteki Pandas języka Python sprawdziłem, czy słowa znajdujące się na liście ‘bad_words.csv’ wchodzą w skład haseł, w celu sprawdzenia jak często przekleństwa, obraźliwe i wulgarne wyrażenia są wykorzystywane jako całe hasło lub stanowią jego część.

Podobnie postąpiłem z plikiem ‘babynames-clean.csv’ - chciałem sprawdzić, jak często ludzie ustawiają imię jako hasło lub jego część. W efekcie tych działań powstały dwie nowe kolumny w dataframe utworzonym z oryginalnego pliku:

1. is_name - wartość True jeśli hasło zawiera imię, False jeśli hasło nie zawiera imienia
2. is_ bad - wartość True jeśli hasło zawiera wulgarne wyrażenie, wartość False jeśli nie hasło nie zawiera

Dodatkowo wyszukałem hasła, które są ciągiem cyfr stanowiącym datę w formacie DDMMYYYY oraz stworzyłem kolumnę ‘is_date’, której komórka ma wartość True jeśli ciąg znaków jest datą w ww. formacie a wartość False, jeśli ciąg znaków nie pasuje kryteriów daty.
Dataframe rozszerzony o trzy dodatkowe kolumny zapisałem do pliku wynikowego ‘common_passwords_modified.csv’.
Kod, którym dokonałem wyżej opisanych operacji jest zapisany w pliku ‘passwords.ipynb’.

Storyboard składa się z 7 slajdów.

1. Slajd tytułowy, który zawiera krótki opis z wprowadzeniem do omawianego zagadnienia oraz linki do źródłowych datasetów
2. Strona 1 - interaktywny wykres słupkowy pokazujący ilość haseł ze względu na liczbę znaków, liczniki średniej długości hasła, średniej ilości liter wchodzących w skład hasła oraz średniej ilości cyfr wchodzących w skład hasła
3. Strona 2 - interaktywne drzewo dekompozycji ukazujące ilość haseł zawierających określone rodzaje znaków (znaki specjalne, wielkie litery, małe litery, cyfry)
4. Strona 3 - wykres kołowy przedstawiający ilość haseł spełniających kryteria daty (8 cyfr, format DDMMRRRR), podgląd rekordów spełniających kryteria oraz licznik rekordów spełniających kryteria
5. Strona 4 - drzewo dekompozycji, którego celem jest pokazanie ile procent spośród datasetu stanowią hasła składające się wyłącznie z małych liter, podgląd rekordów spełniających kryteria oraz wykres kołowy pokazujący ile haseł posiada wartość ‘True’ w kolumnie ‘is_name’

6. Strona 5 - podsumowanie analizy, wykres kołowy z procentowym udziałem haseł posiadających wartość ‘True’ w kolumnie ‘is_bad’, przycisk z przekierowaniem do slajdu ‘Duplikat sekcji Strona 5’
7. [ukryty] Duplikat sekcji Strona 5 - ta sama zawartość co Strona 5 + podgląd haseł spełniających kryteria ‘is_bad’

Bibliografia

1. *10000 Most Common Passwords*. (b.d.). Pobrano 11 styczeń 2022, z https://kaggle.com/shivamb/10000-most-common-passwords
2. *Bad Bad Words*. (b.d.). Pobrano 11 styczeń 2022, z https://kaggle.com/nicapotato/bad-bad-words
3. *First Names—Dataset by alexandra | data.world*. (b.d.). Pobrano 11 styczeń 2022, z https://data.world/alexandra/baby-names