Poniedziałek, 10:00

Eksploracja danych internetowych

Zadanie 3

Barbara Morawska 234096

Andrzej Sasinowski 234118

Wydział Fizyki Technicznej, Informatyki i Matematyki Stosowanej

Politechnika Łódzka

2020/2021

1 Cel zadania

Celem zadania było wykorzystanie danych o odwiedzonych przez użytkowników stronach internetowych w taki sposób, aby można było zarekomendować na tej podstawie strony nowemu użytkownikowi.

2 Przetwarzane dane

W zadaniu zostały wykorzystane dane z zadania pierwszego dotyczące użytkowników i stron internetowych na jakie wchodzili. Każdy rekord zawierał adres hosta oraz atrybuty odpowiadające odwiedzonym stronom przyjmujące wartość 0 – jeśli użytkownik nie wchodził na daną stronę oraz 1 – jeśli wchodził na stronę. Wczytywany plik *arff* został zaprezentowany na *Listingu 1*.

```
@RELATION user_attributes.arff


@ATTRIBUTE userID STRING

@ATTRIBUTE /shuttle/countdown/ {0, 1}

@ATTRIBUTE /shuttle/missions/sts-71/images/images.html {0, 1}

[…]

@ATTRIBUTE /shuttle/missions/sts-71/news/ {0, 1}

@ATTRIBUTE /shuttle/missions/sts-63/mission-sts-63.html {0, 1}


@DATA

199.72.81.55,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0

unicomp6.unicomp.net,1,0,0,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
```

Listing 1: Struktura przetwarzanych danych.

3 Opis rozwiązania

W celu wykonania zadania wykonane zostały następujące kroki:

1. Przygotowanie danych do klasteryzacji, które opierało się na usunięciu kolumny identyfikującej hosta i pozostawieniu jedynie informacji na temat wejść na strony internetowe przez danego użytkownika.
2. Klasteryzacja metodą *Simple K-means* na 2, 5 i 10 klastrów reprezentujących grupy użytkowników, którzy odwiedzili podobne strony.
3. Wygenerowanie losowego użytkownika.

4. Obliczenie podobieństwa Jaccarda (1) wektora wejść na strony nowego użytkownika i wektora wejść odpowiadającemu danemu klastrowi.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

gdzie A i B to wektory wejść na strony nowego użytkownika i reprezentującego dany klaster.

5. Wybranie klastra o najwyższym współczynniku podobieństwa względem wektora wejść nowego użytkownika.
6. Zarekomendowanie stron nowemu użytkownikowi (rekomendacja uwzględniająca tylko strony, na które nie wchodził nowy użytkownik).

4 Wyniki

Wygenerowany użytkownik:

Wygenerowany losowo użytkownik przyjął następujące atrybuty (*Listing 2*):

```
Jack.Strong,0,0,1,0,0,0,1,0,1,0,0,0,0,1,1,0,0,1,0,0,0,0,0,1,0,1,0,0
```

Listing 2: Atrybuty nowego użytkownika.

Klasteryzacja – 2 klastry

```
Cluster 0      Mean/Mode:  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Cluster 1      Mean/Mode:  1 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0


Clustered Instances

0      3362 ( 89%)

1       420 ( 11%)
```

Listing 3: Wyniki klasteryzacji stron odwiedzonych przez użytkowników dla 2 klastrów.

Podobieństwo Jaccarda wektora nowego użytkownika oraz wektorów reprezentujących dane klastry zostały przedstawione w *Tabeli 1*.

Nowy użytkownik	Klaster	Podobieństwo Jaccarda
	0	0.00
	1	0.10

Tabela 1: Podobieństwo Jaccarda wektora nowego użytkownika oraz wektorów reprezentujących dane klastry.

Najbardziej podobnym klastrem jest *klaster 1* (podobieństwo – 0.1).

Strony rekomendowane nowemu użytkownikowi:

```
/shuttle/countdown/

/ksc.html
```

Listing 4: Strony rekomendowane nowemu użytkownikowi.

Klasteryzacja – 5 klastrów

```
Cluster 0      Mean/Mode:  1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Cluster 1      Mean/Mode:  1 0 1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Cluster 2      Mean/Mode:  1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Cluster 3      Mean/Mode:  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Cluster 4      Mean/Mode:  0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0


Clustered Instances

0      1390 ( 37%)

1       142 (  4%)

2       602 ( 16%)

3      1339 ( 35%)

4       309 (  8%)
```

Listing 5: Wyniki klasteryzacji stron odwiedzonych przez użytkowników dla 5 klastrów.

Podobieństwo Jaccarda wektora nowego użytkownika oraz wektorów reprezentujących dane klastry zostały przedstawione w *Tabeli 2*.

<table>
<tr><td rowspan="6">Nowy użytkownik</td><td>Klaster</td><td>Podobieństwo Jaccarda</td></tr>
<tr><td>0</td><td>0.11</td></tr>
<tr><td>1</td><td>0.30</td></tr>
<tr><td>2</td><td>0.00</td></tr>
<tr><td>3</td><td>0.00</td></tr>
<tr><td>4</td><td>0.13</td></tr>
</table>

Tabela 2: Podobieństwo Jaccarda wektora nowego użytkownika oraz wektorów reprezentujących dane klastry.

Najbardziej podobnym klastrem jest *klaster 1* (podobieństwo – 0.3).

Strony rekomendowane nowemu użytkownikowi:

```
/shuttle/countdown/

/ksc.html
```

Listing 6: Strony rekomendowane nowemu użytkownikowi.

Klasteryzacja – 10 klastrów

```
Cluster 0      Mean/Mode:  1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Cluster 1      Mean/Mode:  1 0 1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Cluster 2      Mean/Mode:  1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Cluster 3      Mean/Mode:  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Cluster 4      Mean/Mode:  0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Cluster 5      Mean/Mode:  1 0 0 0 0 1 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0

Cluster 6      Mean/Mode:  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0

Cluster 7      Mean/Mode:  0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Cluster 8      Mean/Mode:  0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Cluster 9      Mean/Mode:  0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0


Clustered Instances


 0      1171 ( 31%)

 1       131 (  3%)

 2       547 ( 14%)

 3       861 ( 23%)

 4       230 (  6%)

 5        77 (  2%)

 6        41 (  1%)

 7        84 (  2%)

 8       279 (  7%)

 9       361 ( 10%)
```

Listing 7: Wyniki klasteryzacji stron odwiedzonych przez użytkowników dla 10 klastrów.

Podobieństwo Jaccarda wektora nowego użytkownika oraz wektorów reprezentujących dane klastry zostały przedstawione w *Tabeli 3*.

	Klaster	Podobieństwo Jaccarda
Nowy użytkownik	0	0.11
	1	0.30
	2	0.00
	3	0.00
	4	0.13
	5	0.00
	6	0.00
	7	0.11
	8	0.00
	9	0.00

Tabela 3: Podobieństwo Jaccarda wektora nowego użytkownika oraz wektorów reprezentujących dane klastry.

Najbardziej podobnym klastrem jest *klaster 1* (podobieństwo – 0.3).

Strony rekomendowane nowemu użytkownikowi:

```
/shuttle/countdown/

/ksc.html
```

Listing 8: Strony rekomendowane nowemu użytkownikowi.

5 Wnioski

Na podstawie wyników uzyskanych podczas zadania można wyciągnąć następujące wnioski:

- Liczba klastrów nie wpływa na wyniki rekomendacji.
- Za każdym razem użytkownik jest dopasowywany do tego samego klastra, który reprezentuje stosunkowo niewielką liczbę użytkowników. Sugeruje to, że wygenerowany użytkownik jest dosyć charakterystyczny i nietypowy dlatego trudno znaleźć dla niego odpowiednie rekomendacje.
- Ze względu na małą grupę, do której dopasowywany jest nowy użytkownik liczba proponowanych stron nie przekracza kilku pozycji.
- W każdym przypadku jednym z najbardziej licznych klastrów jest taki, który zawiera same zera, czyli składający się z użytkowników, którzy nie odwiedzili żadnej ze zdefiniowanych stron lub odwiedzali pojedyncze witryny. W tym przypadku podobieństwo Jaccarda zawsze będzie wynosiło 0.0.