НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТІМЕНІ ІГОРЯ СІКОРСЬКОГО»
ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ
КАФЕДРА МАТЕМАТИЧНИХ МЕТОДІВ СИСТЕМНОГО АНАЛІЗУ

Дипломна робота на тему:

Підвищення ефективності надання послуг в сфері медичного туризму через кластеризацію веб-користувачів

Виконала:
Лєднікова Анна, студентка групи КА-33

Науковий керівник:
к.т.н., доцент Тимошенко Ю.О.

Вступ

Об’єкт дослідження:
поведінка користувачів, їх взаємодія з сайтом

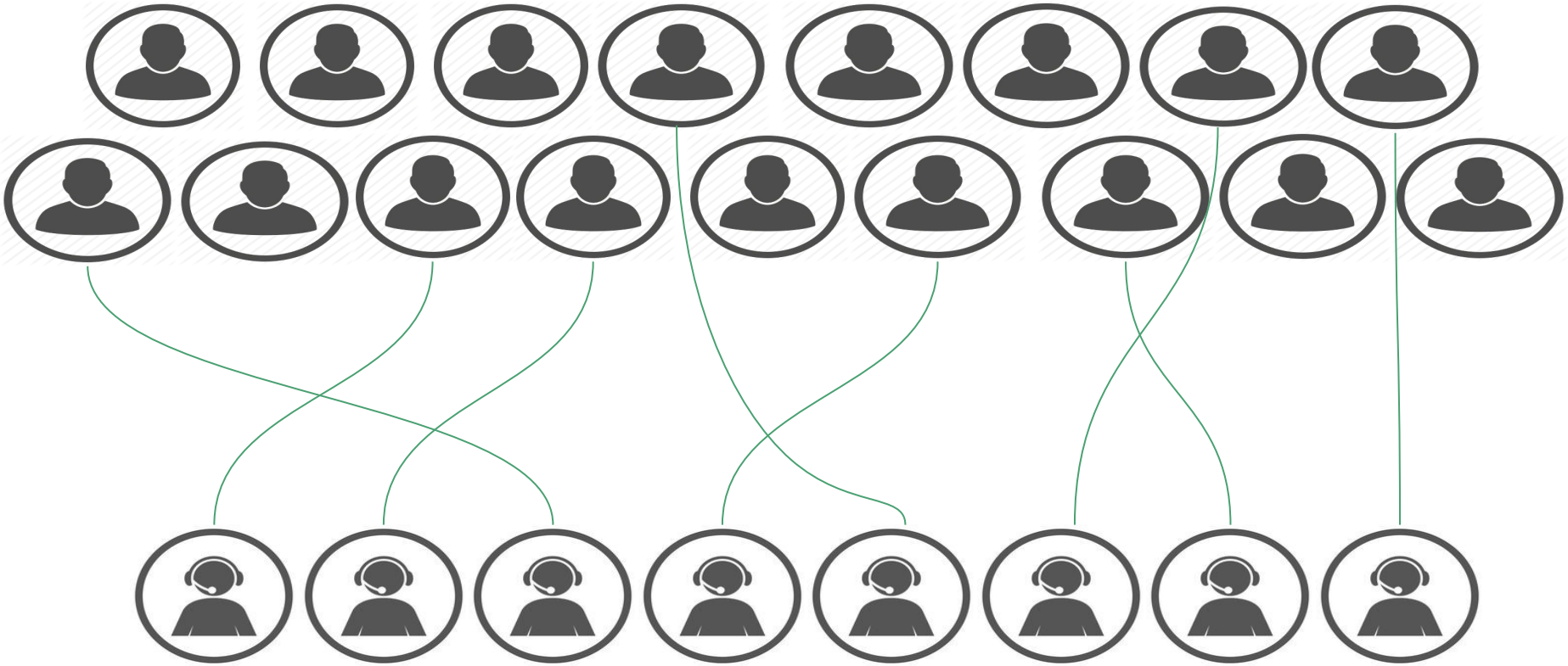
Предмет дослідження:
журнал подій веб-користувачів

Мета дослідження:
отримання описів сегментів

Актуальність роботи

- 3000+ звернень на лікування кожного місяця
- 1 пацієнт в середньому ведеться 1,5 місяця
- Менеджер обирає пацієнта випадково з ~10 заявок ~30 разів/день.
- Клієнт побажав отримати типажі користувачів
- Для них можна розробити стратегії ведення клієнта
- Це основа для автоматичного розподілу клієнтів між менеджерами
- Ця економія часу дасть змогу допомагати більшій кількості клієнтів

Як це відбувається зараз



Як це має бути

Постановка задачі

В наявності:

- сайт компанії, де користувачі залишають запити
- журнал дій 800 000 користувачів сайта за 1,5 року

Треба:

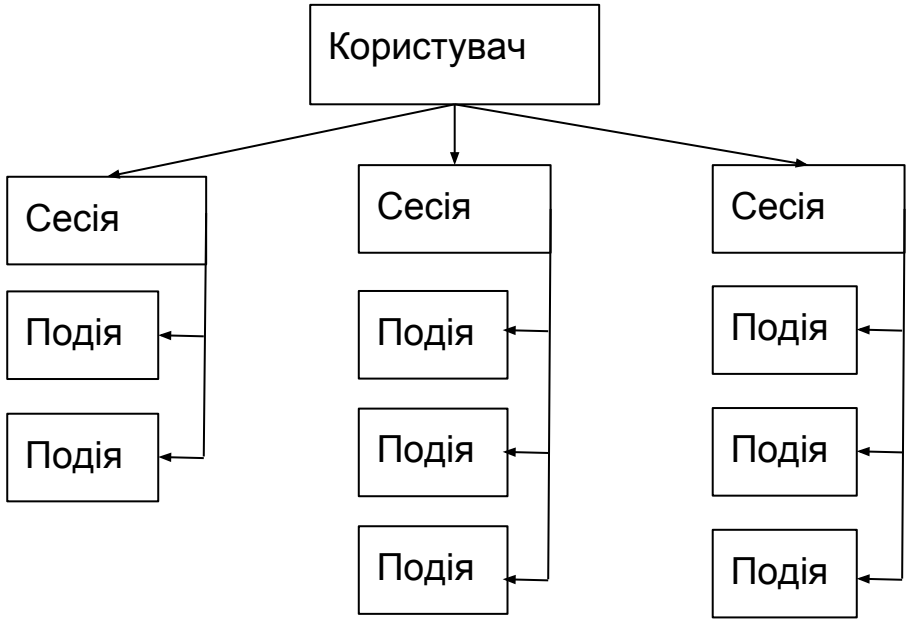
- розробити необхідний інструментарій для перетворення цього журналу в датасет для сегментації
- натренувати моделі та обрати кращу за метриками якості
- скласти кількісний і якісний опис основних сегментів

Сегментація

Приклад сегменту:

- Країна:
  Казахстан
- Браузер:
  Mozila FireFox
- Операційна система:
  Windows
- Кількість переглядів статей:
  >10

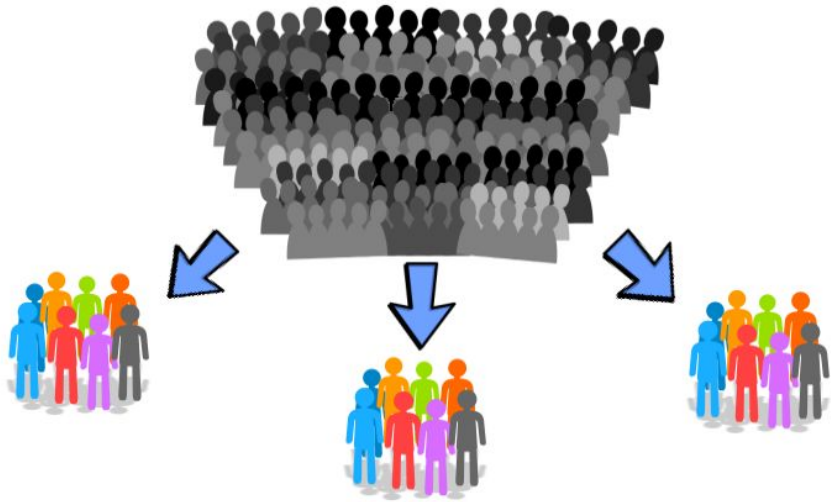
Приклад сесії:



Інструменти сегментації

«Підготовлені дані» передаються **алгоритму інтелектуального аналізу даних**, який видає висновок що інтерпретується для отримання **нових і потенційно корисних знань**.

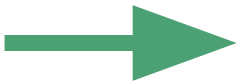
Кластеризація пов'язана з угрупованням об'єктів, які схожі один на одного і відрізняються від об'єктів, що належать до інших кластерів.



Опис використаних даних

Подія

{u'anonymousId':
u'79c96d10-9814-4aba-a04d-7893c36300b2',
u'ip': u'93.75.22.246',
u'path':
u'/clinics/country=germany/direction=nejrohirurgiya/',
u'referrer': u'https://www.google.com.ua/',
u'timestamp': u'2016-02-15T13:09:32.982+0000',
u'title': ‘Нейрохирургия в Германии - цены и отзывы
о лечении за границей’,
u'type': 1,
u'version': 1}



Приклад шляху

[('clinic/medicinskij-centr-anadolu/',
'2016-02-15T13:10:47.284+0000'),
('/clinics/country=germany/direction=nejroh
irurgiya/',
'2016-02-15T13:09:32.982+0000'),
('https://www.google.com.ua/',
'2016-02-15T13:09:32.982+0000')]

Математична модель

Кожна подія a складається з певних атрибутів p_j, впорядкованх за значущістю: $a = (p_0, p_1, \ldots, p_{N_p})$ де N_p - кількість атрибутів.

Кожен атрибут приймає значення з відповідної множини кінцевого розміру P_j. Значущі атрибути: p_0 - номер користувача; p_1 - час, коли подія відбулася; p_2 - тип події;

Множину подій журналу зобразимо у вигляді множини: $A = \{ a_i \in \Re^{N_p} \mid i = \overline{1, N_a} \}$ де A - множина всіх подій, a_i - окремо взята подія, N_a - кількість подій.

Нехай множина A впорядкована за атрибутом часу: $\forall j < k, j, k = \overline{1, N_a} : (a_j < a_k) \Leftrightarrow (p_1^j < p_1^k)$

Далі потрібно розбити множину подій за користувачами, чиї номера зазначені в атрибуті p_0 кожної події.

$$\forall\, i = \overline{1, N_a} : p_0^i \in P_0 = \{ p_{0_k} \mid k = \overline{1, N_{p_0}} \}$$

Тоді множина подій користувача: $U_k = \{ a_i \in \Re^{N_p} \mid \forall i\ p_0^i = p_{0_k},\ i = \overline{1, N_{U_k}} \},\ k = \overline{1, N_{p_0}}$

Поділ на сесії

1. Нехай $T = 30$ хв. , де T - максимальний інтервал між подіям, лічильник сесій $m \in \aleph$. Починаємо з першого користувача, тобто $k = 1$.
2. Визначаємо для користувача k першу сесію S_1^k , тобто та $m = 1$. Т.к. події впорядковані за часом, то

$$S_1^k = \{ \ a_i \in \Re^{N_p} \mid i = \overline{1, N_{S_1^k}} \},$$

де $N_{S_1^k}$ - номер останньої події першої сесії, може приймати наступні значення:

   a. $N_{S_1^k}$ таке що: $(p_1^{N_{S_1^k}+1} - p_1^{N_{S_1^k}}) > T$

   b. $N_{S_1^k} = N_{U_k}$, якщо $\forall l = \overline{1, N_{U_k} - 1}, : (p_1^{l+1} - p_1^l) < T$

3. У випадку 2.а продовжуємо. В загальному випадку сесія m визначається так:

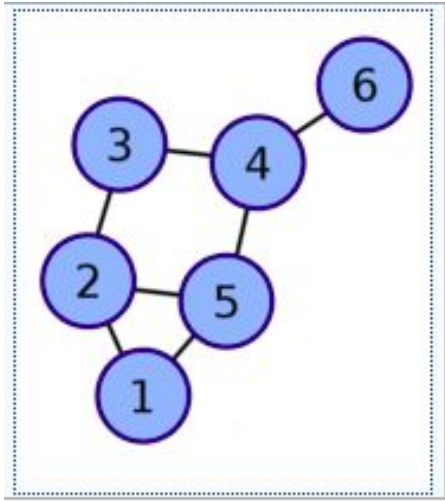
$$S_m^k = \{ \ a_i \in \Re^{N_p} \mid i = \overline{N_{S_{m-1}^k}, N_{S_m^k}} \},$$

де $N_{S_{m-1}^k}$ та $N_{S_m^k}$ - номера останніх події сесії під номером $m-1$ та m відповідно.

4. У разі 2.б збільшуємо k . Якщо $k = N_{p_0}$, то алгоритм закінчено, інакше повторюємо 2 та 3 крок.

Перетворення логів у шляхи за допомогою графів

- **Вузол** - це подія
- **Ребра** - переходи (відношення між подіями), заповнюється відповідним групі числом.
- Прикладі груп відношень:
  - перехід від перегляду сторінки до виклику форми
  - перехід від виклику форми до її відправки
- Події сортуються за часом
- Перші події - корені, останні - листи



$$\begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Утворення шляхів

Множина типів подій:

$$P_2 = \{p_{2_1}, p_{2_2}, p_{2_3}\},$$

де p_{2_k} - індикатор типу події: “перехід на сторінку”, “дія на сторінці”, “вхід до системи”.

Множина переходів між подіями:

$$R = P_2 \times P_2$$

Поставимо у відповідність кожному елементу множини R елемент натурального ряду і отримаємо множину індексів I.

Матрицю сесії:

$$M_S = \left\| m_{ij} \right\|_{N_S}^{N_S},$$

де $m_{ij} \in I$ - відношення між подіями e_i та e_j,

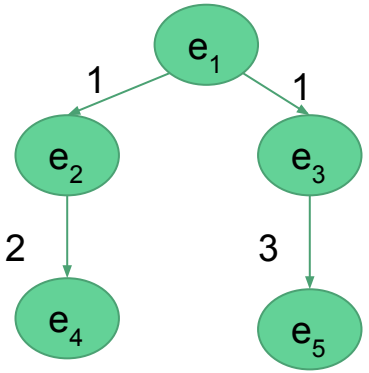
а N_S - кількість подій у сесії.

Далі виділяємо шляхи та строїмо дерево.

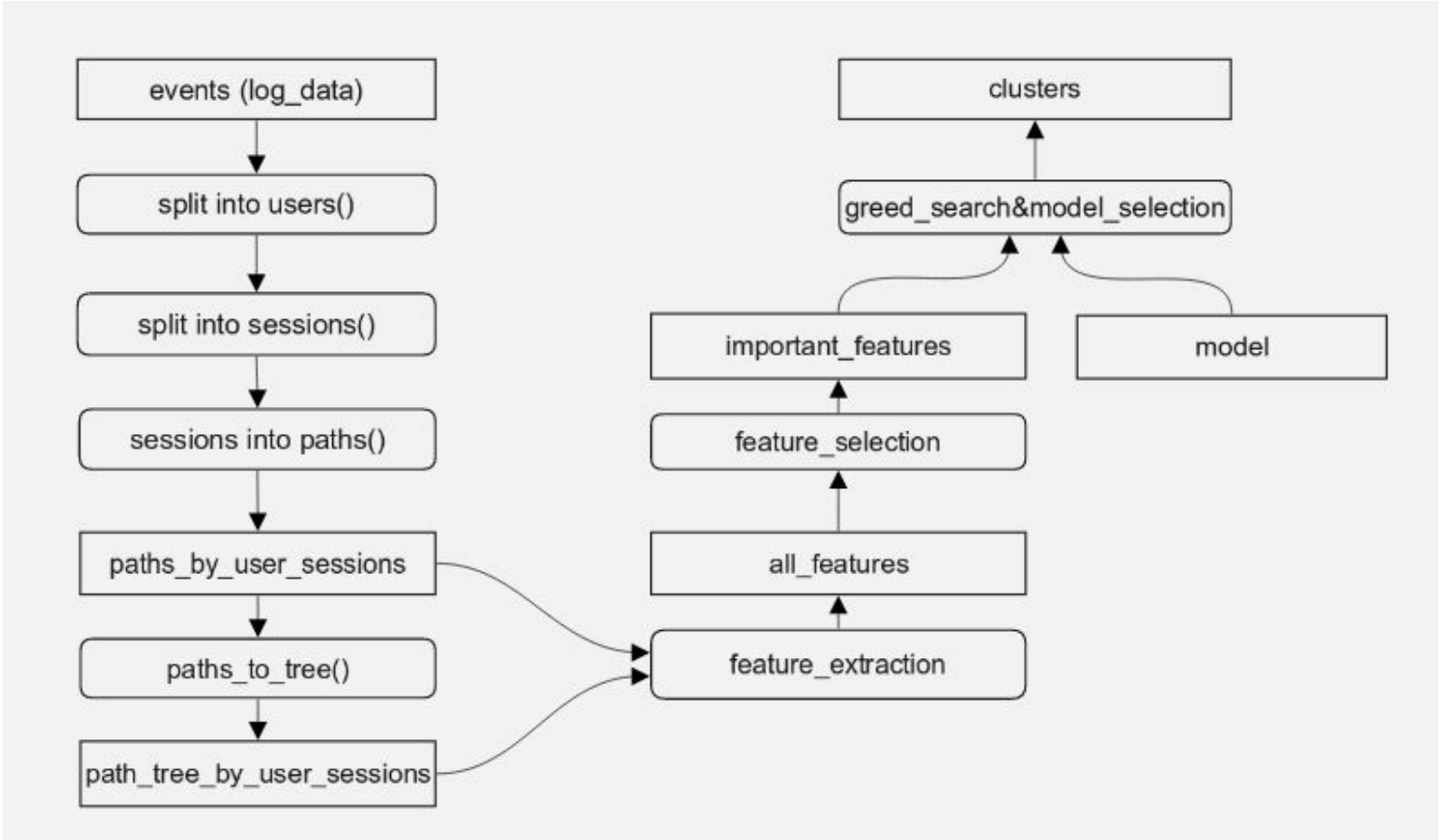
$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \end{pmatrix}$$

$e_1e_2e_4$

$e_1e_3e_5$



Загальна блок-схема процесу



Обрані інструменти

Моделі:

- Kmeans
- Bisect Kmeans

Метрики:

- Згуртованість кластерів

$$SSE = \sum_{i} \sum_{x \in C_i} (x - c_i)^2$$

- Поділ кластерів

$$Total\ SSB = \sum_{i} |C_i|(c - c_i)^2$$

- Корисність кластеризації

$$\sum_{i=1}^{K} max(conv_i - conv\ , 0) * \frac{|C\,| - |C_i|}{|C\,|}$$

C_i - кластер i,
c_i - центр кластеру i,
c - загальне середнє,
|Ci|- це розмір кластера i,
$conv_i$ - конверсія кластеру i,
|C|- кількість усіх елементів.

python™

jupyter

mongoDB®

Spark

Обрання найкращої моделі кластеризації

Модель	SSE	SSB	Користність кластеризації
Kmeans, k=8	321 278,37	581 067,65	0,016
Kmeans, k=12	**298 513,47**	**667 236,51**	**0,028**
Bisect Kmeans, k=5	390 116,84	475 737,06	0,008
Bisect Kmeans, k=9	342 124,65	523 729,25	0,017

Аналіз результатів - 1

В результаті отримано:

- **4** кластера з більшої конверсією, ніж середня (1.78%).
  Ці кластера об’єднують **30%** всіх користувачів.
  Максимальна конверсія (3.25%) в **1.8 більша** за середню.
  Позначені *червоним* кольором.
- **1** кластер з нульовою конверсією, що складає **0.7%**.
  Позначений *синім* кольором.
- *Жовтим* позначені 7 кластерів, що мають схожу на середню конверсію або меншу.

Аналіз результатів - 2

- Умовно сегменти можна поділити на поведінкові та інформаційні, наприклад, кластер 10 та 5 відповідно.

Кластер 10 - 145309 людей - 2,36% поїхало на лікування

Зазвичай заходять на своє захворювання або його напрямку та одразу залишають заявку. Частіше: цукровий діабет, сколіоз, хвороба Паркінсона.

Кластер 5 - 13573 людей - 3,2% поїхало на лікування

Шукають в Туреччині: ринопластика, пересадка кісткового мозку/печінки, рхінопласти, хірургічне зменшення ваги, остеопороз, гідроцефалія, неврит лицьового нерва, ревматоїдний артрит.

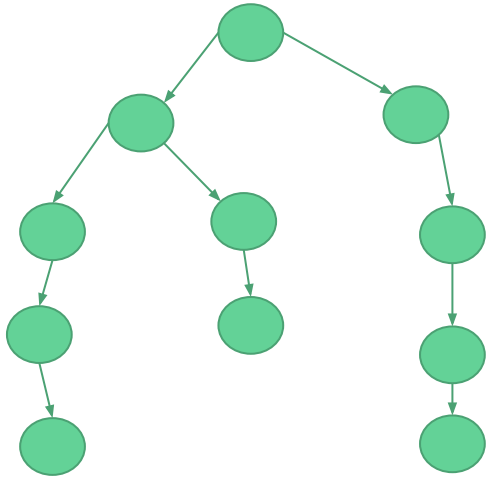
Аналіз результатів - 3

- Поведінкові відрізняються “емоційними” характеристиками отриманими завдяки особливостям графів

Кластер 2 - 30091 людей -
1,57% поїхали на лікування

Дуже ретельно та повільно переглядають сторінки майже всіх сутностей, але більше вивчають у глибину, ніж перебирають можливі варіанти. Заходять ~8 разів, відкривають 3 паралельні вкладки, але на кожній роблять ~2 кроки.

Типічна сесія такого користувача:



Висновки

- Розроблено інструмент для перетворення журналу подій у дата сет для кластеризації
- Були виділені основні сегменти користувачів, зроблено їх кількісний і якісний опис
- Отримана сегментація може допомогти з:
  - визначенням рівня усвідомлення та інформованості користувача, що дозволить раціонально використовувати час координаторів.
  - формуванням базових стратегії ведення заявок клієнтів певного сегмента.

Новизна:

- нелінійність сесій
- використання характеристик дерев для сесій користувача

Шляхи подальшого розвитку

- Автоматизація даного процесу
- Розробка програмного забезпечення для кінцевого користувача
- Виділення більшої кількості характеристик з шляхів та дерев сесій
- Відслідковувати більшого типу подій для більш детального опису користувача
- Розробка системи розподілу користувачів між менеджерами

Дякую за увагу!

Опис всіх кластерів кращої можеді

Кластер 0 - 25815 - 0,015244
Цікавляться певною клінікою у Німеччині. Мабуть, вже щось знають, бо не затримуються на сторінці більше 20 секунд.

Кластер 1 - 376881 - 0,008119
Випадкові: швидко зайшли і вийшли.

Кластер 2 - 30091 - 0,015662
Дуже ретельно та повільно переглядають сторінки майже всіх сутностей, але більше вивчають у глибину, ніж перебирають можливі варіанти. Заходять ~8 разів, відкривають 3 паралельні вкладки, але на кожній роблять ~2 кроки. Найчастіше шукають лікування склерозу або раку простати у Німеччині.

Кластер 3 - 48164 - 0,018256
Відрізняються дуже швидким переглядом сторінок: ~10 секунд. Ключове: напрямок захворювання та місто. Можливо, шукають лікування там, де мають знайомих. Кластер відрізняється лише країнами

Кластер 4 - 4184 - 0
Шукають в Естонії протезування зубів, лікування раку щитовидної залози, псоріазу, наркоманії, неврозу, депресії, лімфостазу, реабілітації після інсульту.

Кластер 5 - 13573 - 0,032468
Шукають в Туреччині: ринопластика, пересадка кісткового мозку/печінки, рхінопласти, хірургічне зменшення ваги, остеопороз, гідроцефалія, неврит лицьового нерва, ревматоїдний артрит.

Кластер 6 - 58766 - 0,01422
Більш всього шукають лікування вірусного гепатиту в Індії, Польщі, США, Австрії, тощо. Для них важливо обрати країну.

Кластер 7 - 10438 - 0,026706
Дуже схожий на 2 кластер, більше читають статті. Зазвичай шукають ЕКО в Туреччині. Шукають конкретно клініку.

Кластер 8 - 6448 - 0,017442
Шукають в Чехії протезування зубів, заміна колінного/тазо-стегнового суглоба, лікування хребта, артрозу, геморою, сколіозу, варикозу, лімфоми.

Кластер 9 - 24410 - 0,007508
Шукають докторів в Ізраїлі для лікування великої кількості хвороб, найбільш популярні: аутизм, рак грудей, меланома, псоріаз, ДЦП, рак легенів, епілепсія.

Кластер 10 - 145309 - 0,023608
Зазвичай заходять на своє захворювання або його напрямку та одразу залишають заявку. Частіше: цукровий діабет, сколіоз, хвороба Паркінсона.

Кластер 11 - 18500 - 0,011952
Їх цікавить в Кореї різні види пластичної хірургії туберкульоз, простатит, геморой, алергія, цироз печінки, міома матки, артроз.