

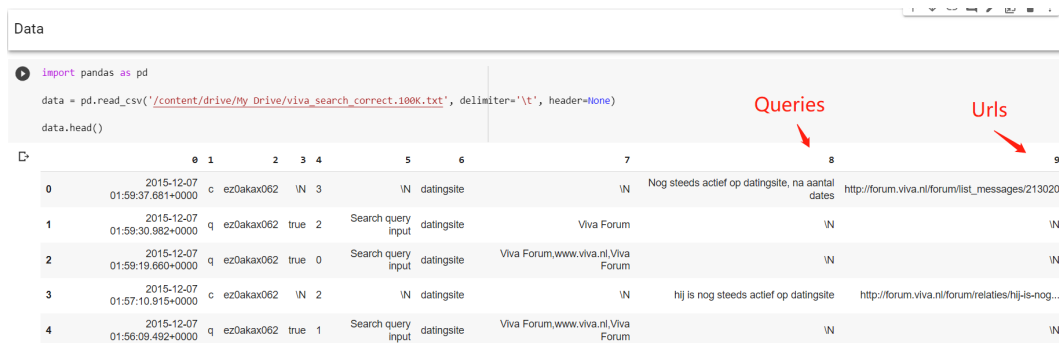
# Information retrieve

## Individual assignment 10

Wei Chen

### 1 Exercise 1

The data is roughly as figure 1 shown.



```
import pandas as pd
data = pd.read_csv('/content/drive/My Drive/viva_search_correct.100K.txt', delimiter='\t', header=None)
data.head()
```

	0	1	2	3	4	5	6	7	8	9
0	2015-12-07 01:59:37.681+0000	c	ez0akax062	\N	3	\N	datingsite	\N	Nog steeds actief op datingsite, na aantal dates	http://forum.viva.nl/forum/list_messages/213020
1	2015-12-07 01:59:30.982+0000	q	ez0akax062	true	2	Search query input	datingsite	Viva Forum	\N	\N
2	2015-12-07 01:59:19.660+0000	q	ez0akax062	true	0	Search query input	datingsite	Viva Forum,www.viva.nl,Viva Forum	\N	\N
3	2015-12-07 01:57:10.915+0000	c	ez0akax062	\N	2	\N	datingsite	\N	hij is nog steeds actief op datingsite	http://forum.viva.nl/forum/relaties/hij-is-nog...
4	2015-12-07 01:56:09.492+0000	q	ez0akax062	true	1	Search query input	datingsite	Viva Forum,www.viva.nl,Viva Forum	\N	\N

Figure 1: Some data

The number of unique queries is 16897 (See figure 2).

## ▼ The number of unique queries

```
✓ [28] queries, urls = data[8], data[9]
0s
    print('Number of unique queries: ', len(set(queries)))
```

➤ Number of unique queries: 16897

Figure 2: Number of unique queries

The top-10 most frequent queries are shown as below (See figure 3):

## ▼ The top-10 most frequent queries

```

✓ [30] queries.value_counts()[1:11]
0s

Kind in je Uppie - Deel 6                                37
Hersensbloeding bij vriend                                36
On Topic - Hij wil zo vaak anaal                          34
Kind misbruikt, hoe nu verder?                            33
beschrijf je laatste neuk                                 32
Hier schrijf ik graag verder van mij af.                  29
Ik wil gewoon eens genomen worden                         29
MMV, kan het iedereen aanraden                            25
Bantopic                                                    24
Blog Zimra: Zo geef je een goede blowjob                 24
Name: 8, dtype: int64

```

Figure 3: Top-10 most frequent queries

Lastly, the top-10 most clicked URLs are demonstrated as figure 4.

## ▼ The top-10 most clicked URLs

```

✓ [33] urls.value_counts()[1:11]
0s

http://forum.viva.nl/?utm_medium=cpc&utm_source=startpagina&utm_campaign=20140702_viva_startpagina&utm_content=tekstlink&utm_term=forumleesmeer 47
http://forum.viva.nl/forum/relaties/hersensbloeding-bij-vriend/list_messages/269993 36
http://forum.viva.nl/forum/seks/on-topic-hij-wil-zo-vaak-anaal/list_messages/259609 34
http://forum.viva.nl/forum/zwanger/kind-in-je-uppie-deel-6/list_messages/247697 31
http://forum.viva.nl/forum/list_messages/244326 29
http://forum.viva.nl/forum/psychie/hier-schrijf-ik-graag-verder-van-mij-af/list_messages/249131 29
http://forum.viva.nl/forum/kinderen/kind-misbruikt-hoe-nu-verder/list_messages/248992 27
http://forum.viva.nl/forum/overig/bantopic/list_messages/235701 26
http://forum.viva.nl/forum/seks/beschrijf-je-laatste-neuk/list_messages/71343 26
http://www.opwindend.net/ 25
Name: 9, dtype: int64

```

Figure 4: Top-10 most clicked URLs