Addressa Dataset Description

Lemei Zhang, Peng Liu

February 2018

1 Introduction

The collection of this dataset is a part of the project for Recommendation Technology (RT). It aims to develop the next-generation news recommender system with the combination of content analysis, consumer intelligence and other advanced recommendation algorithms. The dataset is offered by Adresseavisen, which is a local newspaper company in Trondheim, Norway. This dataset will be helpful in achieving a better understanding of the news articles in conjunction with their readers. This document describes the structure of this dataset and the explanation of the crucial properties that could help you to get start with.

1.1 Technical Support

This dataset is collected with the help of Jørgen H. Frøland, Cxense's technical support Team and people from Adresseavisen. The news and customer related dataset for Adresseavisen are stored in Cxense platform, and thus we can obtain raw data from requesting Cxense APIs.

1.2 State of Collection

We have released 2 versions of datasets: full version and compact/light version. In each version of datasets, we will collect as long as 1 week(from 1 January to 7 January 2017) and 10 weeks (from 1 January to 31 March 2017) datasets. Full version of datasets contain full attributes in 3 folders *rawdata*, *artdata* and *contentdata*, storing raw data, article related data and article content information respectively. Light version of datasets are saved in one folder and contain only basic attributes.

The compact version of dataset can be found online, but the full version of dataset can be achieved by request owing to some articles behind paywall. The size the the one_week.tar.gz and three_month.tar.gz are 1.4G and 16G repectively.

2 Collection Interface

In this chapter, we briefly explain how to collect data from Cxense APIs [1].

2.1 Pull Request

Cxense offers us different APIs to get various data we want. The wiki of Cxense gives us the description and explanation of the function of their APIs. Here, we mainly make use of three kind of APIs to receive the data we need, including */traffic/data*, */profile/content/fetch* and */document/search*.

3 Data Structure

This chapter is a brief overview over the dataset structure from Cxense APIs.

3.1 Structure and Types of Data

The data is collected and stored in JSON, which is also used for all processing procedure.

3.2 Item Attributes

The following table lists the meaning and type of all attributes in full version of datasets. Data attributes in *rawdata* folder.

[1] https://wiki.cxense.com/display/cust/Home

Attributes of Raw Data		
Attribute	type	Description
start	int	The start time of the period of interest specified according to Traffic time specification. The default value is one hour ago.
stop	int	The stop time of the period of interest specified according to Traffic time specification. The default value is right now.
event	list	Requesting events, separated by map
eventId	int	The identifier used to differentiate distinct events from the same user.
activeTime	int	The active time on a page in seconds, if known.
os	String	Operating system that the user used when log in.
referrerHostClass	String	
site	string	The site identifier.
referrerUrl	string	The URL of the referrer page.
city	string	The city name inferred from the IP address.
sessionBounce	boolean	Indicates whether the event is considered as the only event in session.
referrerSearchEngine	string	The name of the referrer search engine.
deviceType	string	The type of the device.
sessionStop	boolean	Indicates whether the event is considered as the last event in session.
sessionStart	boolean	Indicates whether the event is considered as the first event in session.
exitLinkUrl	string	The name of the site hosting the exit link page.
customParameters	list	The list of custom parameters.
userId	string	The cross-site user identifier which can be used to differentiate devices/browsers, or identify different subscription users by the user id.
externalUserIds	string	The requested and known external user identities associated with the user.
userParameters	list	The list of user profile parameters.
url	string	The URL of the visited page.
country	string	The country code inferred from the user's IP address.
region	string	The region code inferred from the IP address.
time	int	The time of event, measured in Unix time.
browser	string	The name of the browser.

The data in *artdata* folder contains following attributes

Attributes of Article Data		
Attribute	type	Description
profile	list	Array of profile objects containing the generated content profile for the URL. Profile objects are defined below.
item	string	Item extracted or generated from the page content. Usually a string or keyword extracted from the page.
groups	list	Array of profile group objects associated with the item.
canonicalUrl	string	Canonical URL as calculated based on incoming events and the fetched page content.
title	string	Extracted title from the last time the page was fetched.
url	string	The URL of the page, corresponding to the URL of the request. (search key, may change afterwards)
httpStatus	int	HTTP status response from the last time the page was fetched.
lastFetched	string	ISO 8601 timestamp representing the last time the page was fetched.
siteId	string	The site identifier for the URL. This is either from the content of the page itself, or based on the site identifier of index pages on the same domain.
id	string	Unique identifier for the content profile. This will be the same for different URLs that are considered equivalent according to Cxense's normalization algorithm. E.g., http://www.example.com/, http://www.example.com and http://example.com will all have the same id value.
thumbnails	list	Array of thumbnail objects representing an image of the page from when the page was last fetched. Thumbnail objects are defined below.

A thumbnail object contains following attributes:

Attributes of Thumbnails		
Attribute	type	Description
width	int	The width of the thumbnail in pixels.
url	string	Either a data URL representing the thumbnail itself, or an URL pointing to an image service that will return the thumbnail.
type	string	Identifies the type or nature of the thumbnail. E.g., if it's a screenshot or something else. Currently valid values are screenshot or dominant
height	int	The height of the thumbnail in pixels.

A groups object contains listed attributes:

Attributes of Groups		
Attribute	type	Description
count	int	Number of appearance in the page.
group	string	Represents the category or type of information.
weight	float	Indicates the relative prominence of the item/group combination.

The data in *contentdata* folder contains following attribute

Attributes of Article Content		
Attribute	type	Description
score	float	The score for the document, if score sort was used
siteId	string	The siteId where the document was found
id	string	The document ID
fields	list	The document fields. It contains lots of (field, type, value) combinations.
field	string	The field's name. E.g., body (article body), keywords (article keywords) and category.
type	string	The type of the captured field.
value	string	The field's value.

The following lists the attributes in light version of dataset.

Attributes of Light Version Dataset		
Attribute	type	Description
eventId	int	The identifier used to differentiate distinct events from the same user.
activeTime	int	The active time on a page in seconds, if known.
os	string	Operating system that the user used when log in.
referrerUrl	sting	The URL of the referrer page.
deviceType	string	The type of the device.
sessionStart	boolean	Indicates whether the event is considered as the first event in session.
sessionStop	boolean	Indicates whether the event is considered as the last event in session.
userId	string	The cross-site user identifier which can be used to differentiate devices/browsers, or identify different subscription users by the user id.
category	sting	The category of the news article.
city	string	The city name inferred from the IP address.
country	string	The country code inferred from the user's IP address.
region	string	The region code inferred from the IP address.
time	int	The time of event, measured in Unix time.
canonicalUrl	sting	Canonical URL as calculated based on incoming events and the fetched page content.
documentId	string	The document id. This will be the same for different URLs that are considered equivalent according to Cxense's normalization algorithm. E.g., http://www.example.com/, http://www.example.com and http://example.com will all have the same id value.
title	string	The title of the article.
keywords	list	The keywords of the article.
namedEntities	list	The named entities of the article, including their types, counts and weights.
author	string	The author of the article.
publishTime	string	The publish time of the article.
profile	Array of object	A set of items which are extracted or generated from the page content. Usually a string or keywords or Named Entities from the page. The possible types and weight of items are given.
item	sting	Item extracted or generated from the page content. Usually a string or keyword extracted from the page.

3.3 Example Item

The following listing shows an example of the event entry saved in *rawdata* folder in JSON format.

```
{
  "eventId": 1082287123,
  "userId": "cx:i8i85z793m9j4yy0:cv8ghy3v45j8",
  "url": "http://adressa.no",
  "time": 1487572383,
  "activeTime": 23,
  "country": "no",
  "city": "verdal",
  "region": "nord-trondelag",
  "referrerHostClass": "direct",
  "site": "9222270286501375973",
  "sessionStart": false,
  "sessionStop": false,
  "sessionBounce": false,
  "deviceType": "Desktop",
  "os": "Windows",
  "browser": "Chrome",
  "intents": [],
  "externalUserIds": [],
  "customParameters": [],
  "userParameters": []
}
```


Figure 1: An example of the event entry.

The following listing shows an example of one article fetch entry saved in *artdata* folder in JSON format.


```
{
  "canonicalUrl": "http://www.adressa.no/nyheter/okonomi/2017/02/06/Br%c3%b8ndbos-Nav-kritikk-skaper-kraftig-debatt-14179352.ece",
  "title": "Brøndbos Nav-kritikk skaper kraftig debatt",
  "url": "http://adressa.no/nyheter/okonomi/2017/02/06/br%c3%b8ndbos-nav-kritikk-skaper-kraftig-debatt-14179352.ece",
  "httpStatus": 200,
  "lastFetched": "2017-02-07T18:14:16Z",
  "siteId": "9222270286501375973",
  "id": "74578862b1c9202de8ae89369542b52c1b2e2f25",
  "profile": [
    {
      "item": "bjarne brøndbos",
      "groups": [
        {
          "count": 1,
          "group": "person",
          "weight": 1
        }
      ]
    },
    {
      "item": "nav",
      "groups": [
        {
          "count": 1,
          "group": "entity",
          "weight": 0.734375
        }
      ]
    },
    {
      "item": "egne erfaringer",
      "groups": [
        {
          "count": 1,
          "group": "concept",
          "weight": 0.1328125
```


Figure 2: An example of article fetch entry.

The following listing shows an example of one content entry saved in *contentdata* folder in JSON format.

```
{
  "score": 1.3201813,
  "siteId": "9222270286501375973",
  "id": "9757814edc2d346dfcf6f54e349f404c4e9775cf",
  "fields": [
    {
      "field": "modifiedtime",
      "type": "time",
      "value": "2017-02-20T09:45:47.000Z"
    },
    {
      "field": "author",
      "value": "mia kristin midtbø"
    },
    {
      "field": "category0",
      "value": "nyheter"
    },
    {
      "field": "body",
      "value": [
        "Saken oppdateres.",
        "En varebil og en personbil har frontkollidert på fylkesvei 714 ved Mjønes i Snillfjord.",
        "Fem personer er involvert, og alle kom seg ut av bilene.",
        "- Fire personer blir sendt til sykehuset Orkdal. Ingen av dem skal være alvorlig skadd, sa operasjonsleder Bjørn Handegard ved Trøndelag politidistrikt kort tid etter ulykken.",
        "Klokka 10.30 melder politiet at de fire skadde fraktes videre til St. Olavs Hospital for undersøkelser.",
        "Personbilen skal ha fått sleng på det slapsete føret, og kolliderte med den andre bilen, som kom kjørende fra Orkdal. Bilen som var truffet, var den andre i et følge på tre biler, opplyser politiet.",
        "Framkommeligheten er redusert på strekningen på fylkesvei 714, melder Statens vegvesen.",
        "Det er mulig å passere, men oppsto noe kø. Trafikken ble dirigert forbi ulykkesstedet.",
        "Bilberging skal bestilles, opplyser Handegard."
      ]
    },
```


Figure 3: An example of content entry.

4 Examples on the Use of the Dataset

1. How to get the location of the user(country, city and region)?

Answer: We can use attributes "country", "city" and "region" to get location information (see red square bellow).

Figure 4: An example of getting location info.

2. How to differentiate subscriber users from ordinary users?

Answer: Since only subscriber users have the access to "pluss" articles, we can filter subscriber users if url attributes or canonicalUrl attributes contains "pluss". As we shows in Fig. 5:

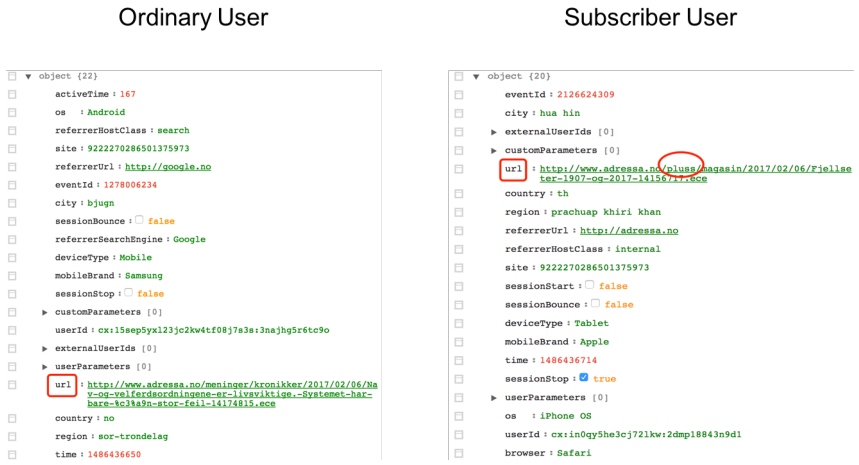

Figure 5: An example of extracting subscriber user.

3. How to string article views (events) into a session for a particular user (start, stop, time, userId, etc.)?

Answer: First, we need to find "sessionStart" being true for this particular user, which means the session starts. Then, we follow the user's next record and make sure if the session stops or not according to "sessionStop" attribute. If "sessionStop" is true, which means the session has stoped and we could know that there are 2 events within this session. If "sessionStop" is false, we need to keep this record and find the next record until the record is found with "sessionStop" being true.

Fig. 6 shows us the steps of how to get the number of events within a session. Fig. 7 are the events within a session for a particular user. The events are ordered chronologically. Timestamp are marked with red, "sessionStart" and "sessionStop" with true value are marked with bold red, and userId is marked with color green.

Figure 6: The steps of calculating the number of events within a session for a specific user.

- {"activeTime": 3, "os": "Android", "referrerHostClass": "direct", "site": "9222270286501375973", "eventId": 287321134, "city": "trondheim", "sessionBounce": false, "deviceType": "Mobile", "mobileBrand": "Samsung", "sessionStop": false, "exitLinkUrl": "http://adressa.no/nyheter/okonomi/2017/02/06/br%c3%b8ndbos-nav-kritikk-skaper-kraftig-debatt-14179352.ece", "customParameters": [], "userId": "cx:img2qifppitfzxe0:1kpyj1v51zzt", "externalUserIds": [], "userParameters": [], "url": "http://adressa.no", "country": "no", "region": "sor-trondelag", "time": 1486436654, **"sessionStart": true**, "browser": "Chrome"}

- {"activeTime": 14, "os": "Android", "referrerHostClass": "direct", "site": "9222270286501375973", "eventId": 1192197545, "city": "trondheim", "sessionBounce": false, "deviceType": "Mobile", "mobileBrand": "Samsung", "sessionStop": false, "exitLinkUrl": "http://www.adressa.no/nyheter/trondheim/2017/02/06/Ungdommer-innr%c3%b8mmer-%c3%a5-ha-knust-busskur-i-Trondheim-14176468.ece", "customParameters": [], "userId": "cx:img2qifppitfzxe0:1kpyj1v51zzt", "externalUserIds": [], "userParameters": [], "url": "http://adressa.no", "country": "no", "region": "sor-trondelag", "time": 1486436755, "sessionStart": false, "browser": "Chrome"}

- {"activeTime": 92, "os": "Android", "referrerHostClass": "internal", "site": "9222270286501375973", "referrerUrl": "http://adressa.no", "eventId": 471985159, "city": "trondheim", "sessionBounce": false, "deviceType": "Mobile", "mobileBrand": "Samsung", "sessionStop": false, "customParameters": [], "userId": "cx:img2qifppitfzxe0:1kpyj1v51zzt", "externalUserIds": [], "userParameters": [], "url": "http://adressa.no/nyheter/okonomi/2017/02/06/br%c3%b8ndbos-nav-kritikk-skaper-kraftig-debatt-14179352.ece", "country": "no", "region": "sor-trondelag", "time": 1486436661, "sessionStart": false, "browser": "Chrome"}

- {"activeTime": 6, "os": "Android", "referrerHostClass": "internal", "site": "9222270286501375973", "referrerUrl": "http://adressa.no", "eventId": 461051268, "city": "trondheim", "sessionBounce": false, "deviceType": "Mobile", "mobileBrand": "Samsung", "sessionStop": false, "customParameters": [], "userId": "cx:img2qifppitfzxe0:1kpyj1v51zzt", "externalUserIds": [], "userParameters": [], "url": "http://www.adressa.no/nyheter/trondheim/2017/02/06/Ungdommer-innr%c3%b8mmer-%c3%a5-ha-knust-busskur-i-Trondheim-14176468.ece", "country": "no", "region": "sor-trondelag", "time": 1486436779, "sessionStart": false, "browser": "Chrome"}

- {"eventId": 1922402924, "city": "trondheim", "externalUserIds": [], "customParameters": [], "url": "http://adressa.no", "country": "no", "region": "sor-trondelag", **"sessionStop": true**, "referrerHostClass": "direct", "site": "9222270286501375973", "sessionStart": false, "sessionBounce": false, "deviceType": "Mobile", "mobileBrand": "Samsung", "time": 1486436786, "userParameters": [], "os": "Android", "userId": "cx:img2qifppitfzxe0:1kpyj1v51zzt", "browser": "Chrome"}

Figure 7: An example of all events within a session for a specific user.

4. How to retrieve article text from event in dataset?

Answer:First, we get the url from *rawdata*, then we get article title from *artdata* according to url, finally, we can extract article text by filtering "body" property according to title from *contentdata*, as illustrated in Fig. 8

Figure 8: An example of extracting article text.

5. How to estimate user's rating of article?

Answer: The "activeTime" attribute measuring how much time the user spent on one article, can be a good way to estimate user's rating to this article.

5 Access to Dataset

The light version of dataset can be found in two different sizes (1.4 GB and 16 GB - compressed size) on Smartmedia webpage `http://reclab.idi.ntnu.no/dataset/` and saved in files according to days. The full content of news articles and the full version of dataset can be accessed on request.

6 Statistics of Dataset

Here we give a preliminary statistics of our dataset using one week data info.

Data Statistics	
Items	Values
Number of articles	11207
Number of entries	2286835
Number of total users	561733
Number of subscriber users	126723
Ratio of subscriber users to total users	0.2256

Number of Articles Per Category	
Category	Number of Articles
nyheter	7516
bil	20
abonnement	6
meninger	383
pluss	1093
vaeret	8
student	1
100sport	659
bedriftsannonser	5
migration catalog	16
tjenester	74
bolig	139
forbruker	563
streaming	2
tema	69
omadresseavisen	5
sport	202
kultur	446

Average Number of Entities Per Articles Per Category	
Category	Number of Entities
nyheter	26.63
bil	31.55
abonnement	15.17
meninger	28.88
pluss	36.13
vaeret	28.88
student	26
100sport	31.77
bedriftsannonser	17.8
migration catalog	25.94
tjenester	9.58
bolig	19.63
forbruker	26.12
streaming	11.5
tema	44.74
omadresseavisen	14
sport	25.25
kultur	31.17

Data Density Per Day	
Day	Density (%)
Day 1	0.081
Day 2	0.065
Day 3	0.059
Day 4	0.054
Day 5	0.044
Day 6	0.054
Day 7	0.060

Number of Views VS. Articles (Top 15)	
Article Id	Number of Views
fbccdbd868387c01497a9b5d13319bc93cce07db	93961
cf83d342459ce871e2a8562a91b7dca946e3201a	70225
aa6a5862cb2ae9fb8996f35a692192559b9083e1	54409
764fc962174c526c1d24932ca6951ca37f93d6a2	52101

c065a4298b0dc55060e4dc6cf62b9467216d42d0	51649
300849a183ef001c215bb8714e4a342f287cfae6	47196
14e07e0abfcdc157d09a6c8b054c276450b73f34	43911
b791d1e6c3355e64e0c33b5e8580ea3fb1fc9ae1	41745
68d1503c73ad169dcfff48214fd0274c4d612e63	36672
0867dbb33bb90970ae48592057be34246a0124ac	36526
87a215563ffbcc513b5387fd90f82d876eba11c4	33000
4e2fd6d44bc38e54519e93ae1e0dae5f0d4dd2af	32400
a81a8d98980b5fc373f395a28264eb31fd1e16f9	32217
5154d39394e4c1b9a26b757e200631bfbef8a6e4	29675
004edd2345395b2b798b941e8b92ed9c550a749a	29657

Here we give a preliminary statistics of our dataset using 3 months data info.

Data Statistics	
Items	Values
Number of articles	48486
Number of entries	27223576
Number of total users	3083438
Number of subscriber users	380527
Ratio of subscriber users to total users	0.1234

Number of Articles Per Category	
Category	Number of Articles
nyheter	30722
bil	47
abonnement	14
meninger	1291
pluss	5342
vaeret	31
student	11
100sport	5855
bedriftsannonser	7
migration catalog	65
tjenester	218
bolig	521
forbruker	1533
streaming	14
tema	124
omadresseavisen	12
sport	754
kultur	1923
privatannonser	1
xyz 42 testseksjon	1

Average Number of Entities Per Articles Per Category	
Category	Number of Entities
nyheter	24.87
bil	26.77
abonnement	15.57
meninger	28.20
pluss	34.79
vaeret	29.58
student	25.27
100sport	31.27
bedriftsannonser	21.29
migration catalog	26.15
tjenester	19.33
bolig	19.23
forbruker	25.59
streaming	10.14
tema	40.59
omadresseavisen	15.5
sport	25.26
kultur	30.79
privatannonser	14
xyz 42 testseksjon	10

Number of Views VS. Articles (Top 15)	
Article Id	Number of Views
e9f7b8052b53839ddff821707ab00b559b48b9a6	220310
134aca17d8e7954f5bc95779e6c865c731700ada	164101
0c2c9247850cafc1143db88372116ace2b4085d3	124707
fbccdbd868387c01497a9b5d13319bc93cce07db	96124
b6f08e0fe6567ed39d3b244f2afeceef43e0dbd3	90442
6af41040924f1aa5cbc5427af8ecb63df0b36fc3	85495
74578862b1c9202de8ae89369542b52c1b2e2f25	77239
0023cf8c8637599ee493463a429f4af62817cf3a	75501
094e3ca8251f2a81626da7af88e25ef03ae7bd86	71346
cf83d342459ce871e2a8562a91b7dca946e3201a	71289
d333a6b9c64b858e0a1280dfebda505e409db1dd	69730
e9a8deeda6a04df6afb887619a3a1880250aed7a	68044
c47f63e7e6d046709c8222e5baebec3ba19e1ae8	67031
fa3f4c9983712f837925cbfe9bd096d09143a2ac	66916
844309b9439a3d4fa923e04acaa0e9f31684bfbd	66830

7 Comparison of Other Existing Dataset

There already have some datasets available on line [2].

Comparison With Existing Datasets					
Dataset	Users	Items	Ratings	Density(%)	Rating scale
Movielens 1M	6040	3883	1,000,209	4.26	[1-5]
Movielens 10M	69,878	10,681	10,000,054	1.33	[0.5-5]
Movielens 20M	138,493	27,278	20,000,263	0.52	[0.5-5]
Jester	124,113	150	5,865,235	31.50	[-10, 10]
Book-Crossing	92,107	271,379	1,031,175	0.0041	[1, 10], and implicit
Last.fm	1892	17632	92,834	0.28	Play Counts
Wikipedia	5,583,724	4,936,761	417,996,366	0.0015	Interactions
OpenStreetMap (Azerbaijan)	231	108,330	205,774	0.82	Interactions
Git (Django)	790	1757	13,165	0.95	Interactions
Cxense (one week)	561,733	11,207	2,286,835	0.036	Click Counts
Cxense (3 months)	3,083,438	48,486	27,223,576	0.0182	Click Counts

8 Frequently Asked Questions

1. **"wordCount" attribute appears in the paper.**
   Answer: "wordCount" isn't appear in the original dataset, which we count it ourselves so it won't be included in the dataset.
2. **Some events do not include "activeTime".**
   Answer: There is the threshold when computing the "activeTime", when it is too short then it won't be account in original dataset.
3. **Some events do not include article some attributes e.g "keywords"/"profile".**
   Answer: They do not appear in the original dataset because either the articles do not contain such attributes or they are not types of news articles.
4. **Some attributes cannot be found in several events.**
   Answer: Not all attributes are included in every event in both kinds of datasets. If the attributes cannot be tracked or extracted from the raw data, such kinds of attributes won't included. For instance, there is no profile if the user visited news homepage.

9 Cite

If you use this dataset, please cite this paper:

[2] http://www.kdnuggets.com/2016/02/nine-datasets-investigating-recommender-systems.html

Gulla, Jon Atle, et al. "The Adressa dataset for news recommendation." Proceedings of the International Conference on Web Intelligence. ACM, 2017.