Info 7374 Final Project

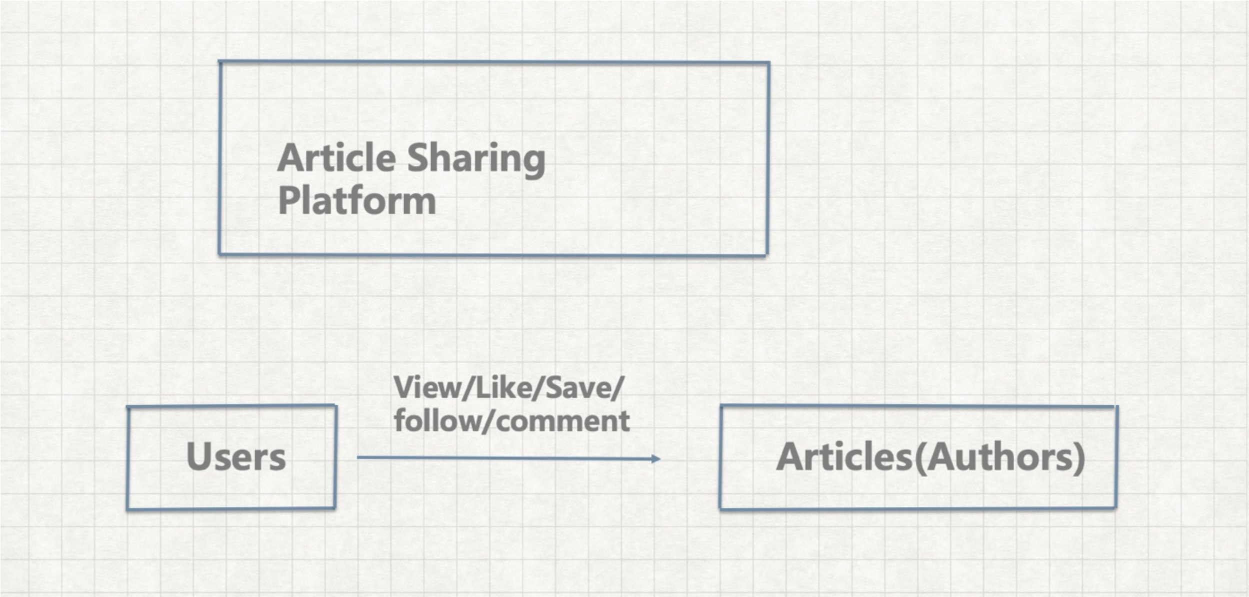
Project Member:

Yufan Yang

Vividh Talesara

Project Description:

● Overview



This project is aimed to build a social media content-based recommendation system to enhance the user experience, which includes recommending or predicting some new items (articles) they will like based on their preference. This project includes the following functions: article classification, article topic keyword extraction, article similarity analysis, article recommendation, user recommendation, user preference analysis, article influence ranking, article influence prediction. This project explores the details and meaning of the marketing recommendation system through different angles and methods.

● Project requirements and goals

Use concepts covered in the class (See textbook and various class links) involving algorithmic marketing to perform data analysis based on large datasets.

Goals:

Build a social media content-based recommendation system to recommend, in other words, predict some new items (articles) users will like based on their preference, recommend new users to authors and so on.

Perform text analysis and keywords extraction based on related datasets contain the key interaction information with users and contents such as: user_id, user_action (view/like/comment/share), content_id, content_text.

● **Problems to be addressed**

Existing recommendation systems have imperfections, such as not considering different user behaviors representing different levels of interest. The new recommendation system will try to Improve the accuracy of recommendations and remove annoying recommendations. It's important to use preference levels which means the record we have for users who interact with the social media posts (in our case including view, like, share, etc) can have a decent effect on the recommendation.

Explanation:

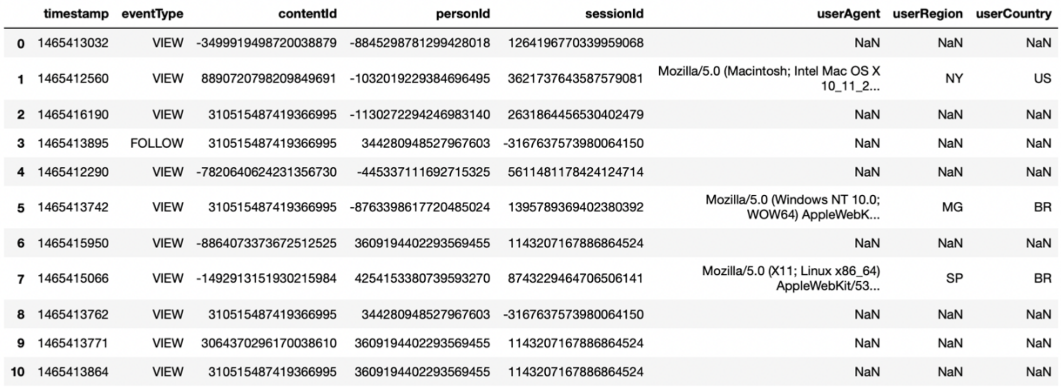
High Accuracy of recommendations means users received new items(contents) with similar interest belongs to contents that the user once liked, commented on, or shared.

Annoying recommendations means users received new items(contents) are only something they viewed before but not really had an interest in.

Data:

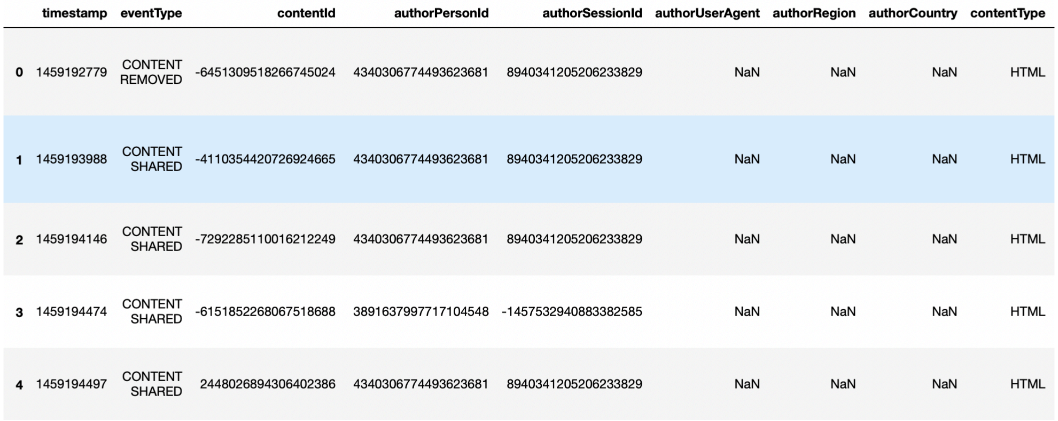
We mainly used the dataset from kaggle： https://www.kaggle.com/gspmoreira/articles-sharing-reading- from-cit-deskdrop. The two datasets are about the information of articles sharing and reading from CI&T DeskDrop. Deskdrop is an internal communications platform developed by CI&T, focused in companies using Google G Suite. Among other features, this platform allows companies employees to share relevant articles with their peers, and collaborate around them.

User Interaction dataset:

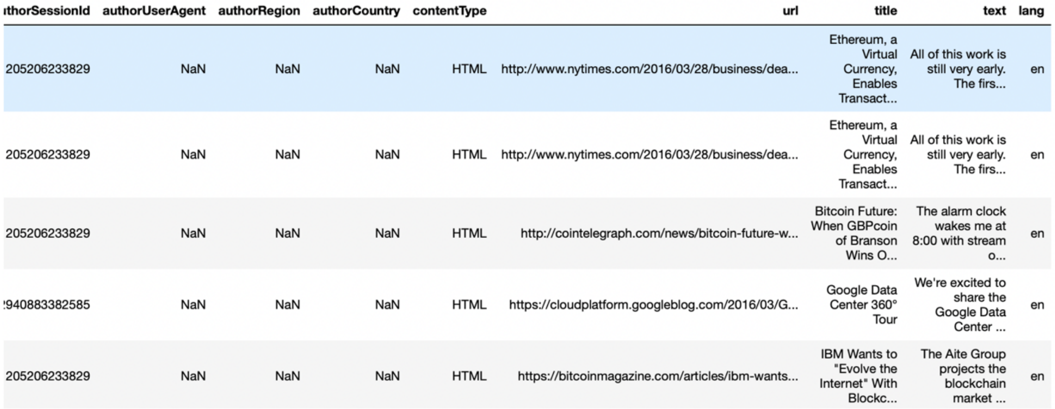


	timestamp	eventType	contentId	personId	sessionId	userAgent	userRegion	userCountry
0	1465413032	VIEW	-3499919498720038879	-8845298781299428018	1264196770339959068	NaN	NaN	NaN
1	1465412560	VIEW	8890720798209849691	-1032019229384696495	3621737643587579081	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_2...	NY	US
2	1465416190	VIEW	310515487419366995	-1130272294246983140	2631864456530402479	NaN	NaN	NaN
3	1465413895	FOLLOW	310515487419366995	344280948527967603	-3167637573980064150	NaN	NaN	NaN
4	1465412290	VIEW	-7820640624231356730	-445337111692715325	5611481178424124714	NaN	NaN	NaN
5	1465413742	VIEW	310515487419366995	-8763398617720485024	1395789369402380392	Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebK...	MG	BR
6	1465415950	VIEW	-8864073373672512525	3609194402293569455	1143207167886864524	NaN	NaN	NaN
7	1465415066	VIEW	-1492913151930215984	4254153380739593270	8743229464706506141	Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/53...	SP	BR
8	1465413762	VIEW	310515487419366995	344280948527967603	-3167637573980064150	NaN	NaN	NaN
9	1465413771	VIEW	3064370296170038610	3609194402293569455	1143207167886864524	NaN	NaN	NaN
10	1465413864	VIEW	310515487419366995	3609194402293569455	1143207167886864524	NaN	NaN	NaN

Articles dataset:



	timestamp	eventType	contentId	authorPersonId	authorSessionId	authorUserAgent	authorRegion	authorCountry	contentType
0	1459192779	CONTENT REMOVED	-6451309518266745024	4340306774493623681	8940341205206233829	NaN	NaN	NaN	HTML
1	1459193988	CONTENT SHARED	-4110354420726924665	4340306774493623681	8940341205206233829	NaN	NaN	NaN	HTML
2	1459194146	CONTENT SHARED	-7292285110016212249	4340306774493623681	8940341205206233829	NaN	NaN	NaN	HTML
3	1459194474	CONTENT SHARED	-6151852268067518688	3891637997717104548	-1457532940883382585	NaN	NaN	NaN	HTML
4	1459194497	CONTENT SHARED	2448026894306402386	4340306774493623681	8940341205206233829	NaN	NaN	NaN	HTML



ıthorSessionId	authorUserAgent	authorRegion	authorCountry	contentType	url	title	text	lang
205206233829	NaN	NaN	NaN	HTML	http://www.nytimes.com/2016/03/28/business/dea...	Ethereum, a Virtual Currency, Enables Transact...	All of this work is still very early. The firs...	en
205206233829	NaN	NaN	NaN	HTML	http://www.nytimes.com/2016/03/28/business/dea...	Ethereum, a Virtual Currency, Enables Transact...	All of this work is still very early. The firs...	en
205206233829	NaN	NaN	NaN	HTML	http://cointelegraph.com/news/bitcoin-future-w...	Bitcoin Future: When GBPcoin of Branson Wins O...	The alarm clock wakes me at 8:00 with stream o...	en
940883382585	NaN	NaN	NaN	HTML	https://cloudplatform.googleblog.com/2016/03/G...	Google Data Center 360° Tour	We're excited to share the Google Data Center ...	en
205206233829	NaN	NaN	NaN	HTML	https://bitcoinmagazine.com/articles/ibm-wants...	IBM Wants to "Evolve the Internet" With Blockc...	The Aite Group projects the blockchain market ...	en

Process Outline:

1. Data Preprocessing

2. Exploratory Data Analysis

3. Text feature/keywords extraction and perform article auto-classification

4. Build recommendation system

5. Design of a pipeline and system to implement this approach and discussion on the system's capabilities

6. Deploy the Model on Azure/AWS or Google Cloud Computing Platform

7. Build a web application to demonstrate the prediction and recommendation results.

Deliverables:

1.Generate a recommendation item list for a specific user; Generate a recommendation users list for a specific item;

2. Achieve article auto-classification and keywords extraction.

3. Further Steps: new article potential influence prediction and new contents suggestions for authors

Use Cases:

1. Select one topic of articles and get a list of top n popular articles on the platform

2. Get recommendation of articles based on the reading record and preference level.

3. Get recommendation of potential readers based on the reading record and preference level.

4. Check how popular an author's articles are on the platform.

5. An author will get keywords extraction and article auto-classification to tel

Milestones:

No	Tasks	Timeframe
1	Environment set up & Data Preparation	Day 1-2
2	Data Exploration & Set preference Level	Day 3-4
3	Text classification & Keywords Extraction	Day 5-6
4	Determine and implement algorithm to build the recommendation model and complete prediction	Day 7-8
5	Evaluation & Presentation	Day 9-10