News Feed

Summarization And Recommendation

Team Members :-

- Karim Atef Henry
- Mohamed Ahmed Hassan
- Fady Nasser Fawzy
- Hoyam Nabil

Orange Innovation Egypt Project

Supervised by :-

- Eng. Ahmed Abdelatty
- Eng. Ghada Soliman

Agenda

- Define Project
- Dataset Definition
- Methodology
- News Feed Modules
  - News Feed Summarization
  - News Feed Recommendation
- Deployment
- Conclusion
- Future Work

Define Project

Define Project

- **Business Challenge:**
  With the vastly growing news feeds, it became harder for users to stay up to date with the latest news in different categories.

- **Project Aim:**
  Introduce to the users a summarized version of the news feeds' article they might be interested in.

Define Project

- **Modules of development:**
  - Divided into **2 concurrent modules**

Summarization Module	Recommendation Module
• Summarize key information of a given **article**. • Reduce **time** needed to read the article.	• **Recommend** to users a set of summarized articles matching their interests. • Introduce the news users the **top trending articles**.

Dataset Definition

Dataset Definition

- **Language :-**
  - English
- **Datasets :-**
  - Approach A: CNN - DailyMail News.
  - Approach B: BBC News.
  - Approach C: MIND Microsoft News.

CNN-DailyMail News

- An **English language** dataset containing over **300k** unique news articles written by journalists at CNN and the Daily Mail.

- The current dataset version supports both **Extractive and Abstractive** summarization.

CNN-DailyMail News

- **It Consists of :-**
  - ID
  - Articles
  - Highlights

- **Dataset Split No. of Articles**

Train	287,113
Validation	13,368
Test	11,490

CNN Dataset

Article:

The bishop of the Fargo Catholic Diocese in North Dakota has exposed potentially hundreds of church members in Fargo, Grand Forks and Jamestown to the hepatitis A virus in late September and early October. The state Health Department has issued an advisory of exposure for anyone who attended five churches and took communion. Bishop John Folda (pictured) of the Fargo Catholic Diocese in North Dakota has exposed potentially hundreds of church members in Fargo, Grand Forks and Jamestown to the hepatitis A. State Immunization Program Manager Molly Howell says the risk is low, but officials feel it's important to alert people to the possible exposure. The diocese announced on Monday that Bishop John Folda is taking time off after being diagnosed with hepatitis A. The diocese says he contracted the infection through contaminated food while attending a conference for newly ordained bishops in Italy last month. Symptoms of hepatitis A include fever, tiredness, loss of appetite, nausea and abdominal discomfort. Fargo Catholic Diocese in North Dakota (pictured) is where the bishop is located

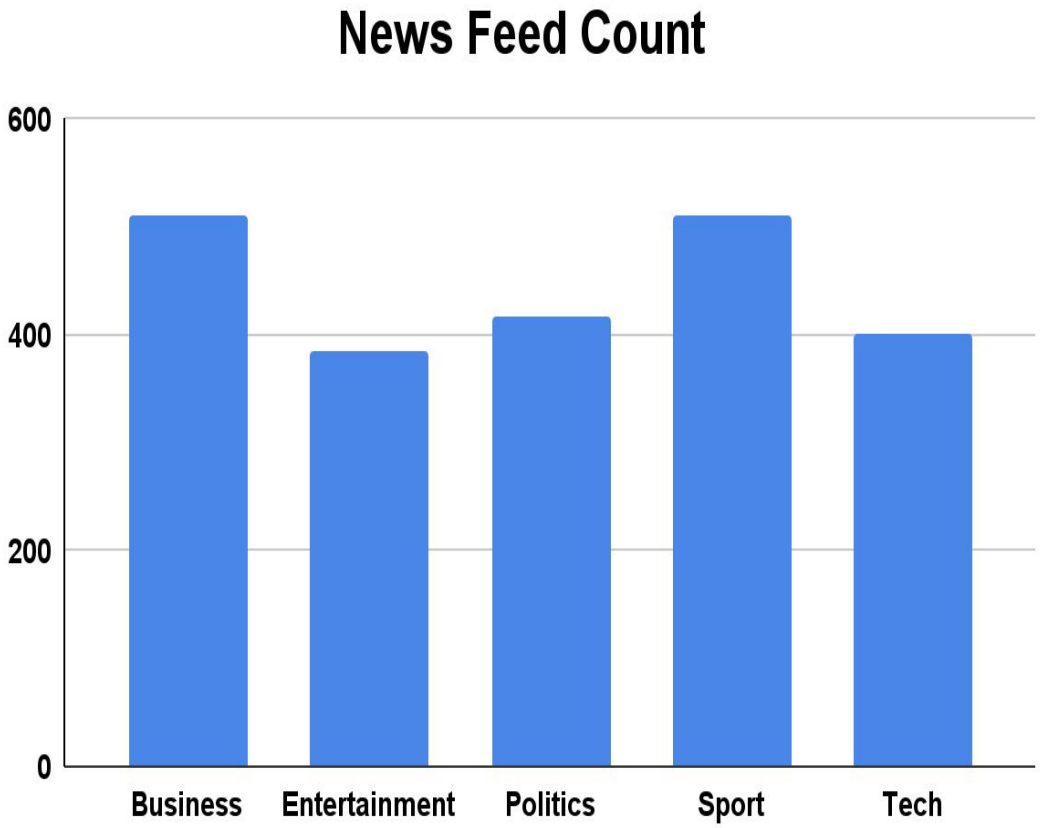
Highlight:

Bishop John Folda, of North Dakota , is taking time off after being diagnosed .\nHe contracted the infection through contaminated food in Italy .\nChurch members in Fargo, Grand Forks and Jamestown could have been exposed .

BBC News Summary

- **2225** documents from the BBC news website corresponding to stories in **five** topical areas from 2004-2005.

- **<u>News Categories Distribution :-</u>**
  - Business 510
  - Entertainment 386
  - Politics 417
  - Sport 511
  - Tech 401



BBC News Summary

- **Divided into categories**, each has related articles and their **extractive** summarization.

Summary:

TimeWarner said fourth quarter sales rose 2% to $11.1bn from $10.9bn. For the full-year, TimeWarner posted a profit of $3.36bn, up 27% from its 2003 performance, while revenues grew 6.4% to $42.09bn. Quarterly profits at US media giant TimeWarner jumped 76% to $1.13bn (£600m) for the three months to December, from $639m year-earlier. However, the company said AOL's underlying profit before exceptional items rose 8% on the back of stronger internet advertising revenues. Its profits were buoyed by one-off gains which offset a profit dip at Warner Bros, and less users for AOL. For 2005, TimeWarner is projecting operating earnings growth of around 5%, and also expects higher revenue and wider profit margins. It lost 464,000 subscribers in the fourth quarter profits were lower than in the preceding three quarters. Time Warner's fourth quarter profits were slightly better than analysts' expectations.

BBC Dataset Sample

Quarterly profits at US media giant TimeWarner jumped 76% to $1.13bn (£600m) for the three months to December, from $639m year-earlier. The firm, which is now one of the biggest investors in Google, benefited from sales of high-speed internet connections and higher advert sales. TimeWarner said fourth quarter sales rose 2% to $11.1bn from $10.9bn. Its profits were buoyed by one-off gains which offset a profit dip at Warner Bros, and less users for AOL.

Time Warner said on Friday that it now owns 8% of search-engine Google. But its own internet business, AOL, had has mixed fortunes. It lost 464,000 subscribers in the fourth quarter profits were lower than in the preceding three quarters. However, the company said AOL's underlying profit before exceptional items rose 8% on the back of stronger internet advertising revenues. It hopes to increase subscribers by offering the online service free to TimeWarner internet customers and will try to sign up AOL's existing customers for high-speed broadband. TimeWarner also has to restate 2000 and 2003 results following a probe by the US Securities Exchange Commission (SEC), which is close to concluding.

Time Warner's fourth quarter profits were slightly better than analysts' expectations. But its film division saw profits slump 27% to $284m, helped by box-office flops Alexander and Catwoman, a sharp contrast to year-earlier, when the third and final film in the Lord of the Rings trilogy boosted results. For the full-year, TimeWarner posted a profit of $3.36bn, up 27% from its 2003 performance, while revenues grew 6.4% to $42.09bn. "Our financial performance was strong, meeting or exceeding all of our full-year objectives and greatly enhancing our flexibility," chairman and chief executive Richard Parsons said. For 2005, TimeWarner is projecting operating earnings growth of around 5%, and also expects higher revenue and wider profit margins.

MIND Microsoft News

- **The Microsoft News Dataset (MIND)** was collected from anonymized behavior logs of Microsoft News website.
- The data randomly sampled **1 million** users.
- This dataset is a small version of MIND **(MIND-small)**, by randomly sampling **50,000 users** and their **behavior logs**.
- Only **training and validation** sets are contained in the (MIND-small) dataset.

MIND Microsoft News

- **<u>The Dataset Contains:</u>**
  - **Behaviors.tsv:**

    The click histories and impression logs of users.
  - **News.tsv:**

    The information of news articles.

MIND Behaviours Dataset

- The **behaviors.tsv** file contains the **impression** logs and users' news click **histories**.
- **It has 5 columns divided by the tab symbol:**
  - Impression ID
  - User ID
  - Time
  - History (Ordered by time)
  - Impressions: **(1 for click and 0 for non-click)**
    - List of news displayed with user's click **behaviors** on them.
    - The orders of news in a impressions have been shuffled.

Behaviours Dataset Example

Impression ID	1
User ID	U13740
Time	11/11/2019 9:05:58 AM
History	N55189 N42782 N34694 N45794 N18445 N63302 N10414 N19347 N31801
Impressions	N55189-1 N35729-0

MIND New Dataset

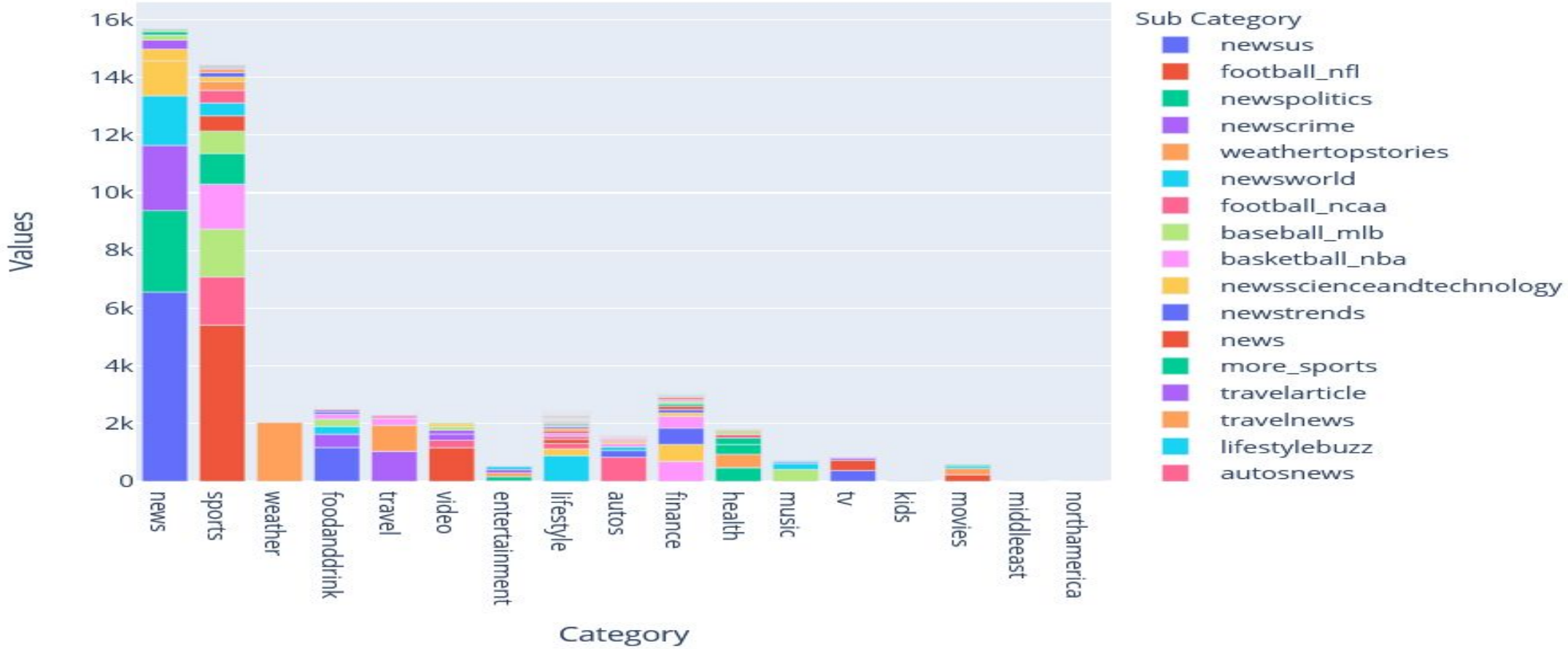
- The **news.tsv** contains the detailed information of news articles involved in the behaviors.tsv file.

- **<u>It has 6 columns Separated by the tab symbol:</u>**
  a. News ID
  b. Category
  c. SubCategory
  d. Title
  e. Abstract
  f. URL

News Dataset Example

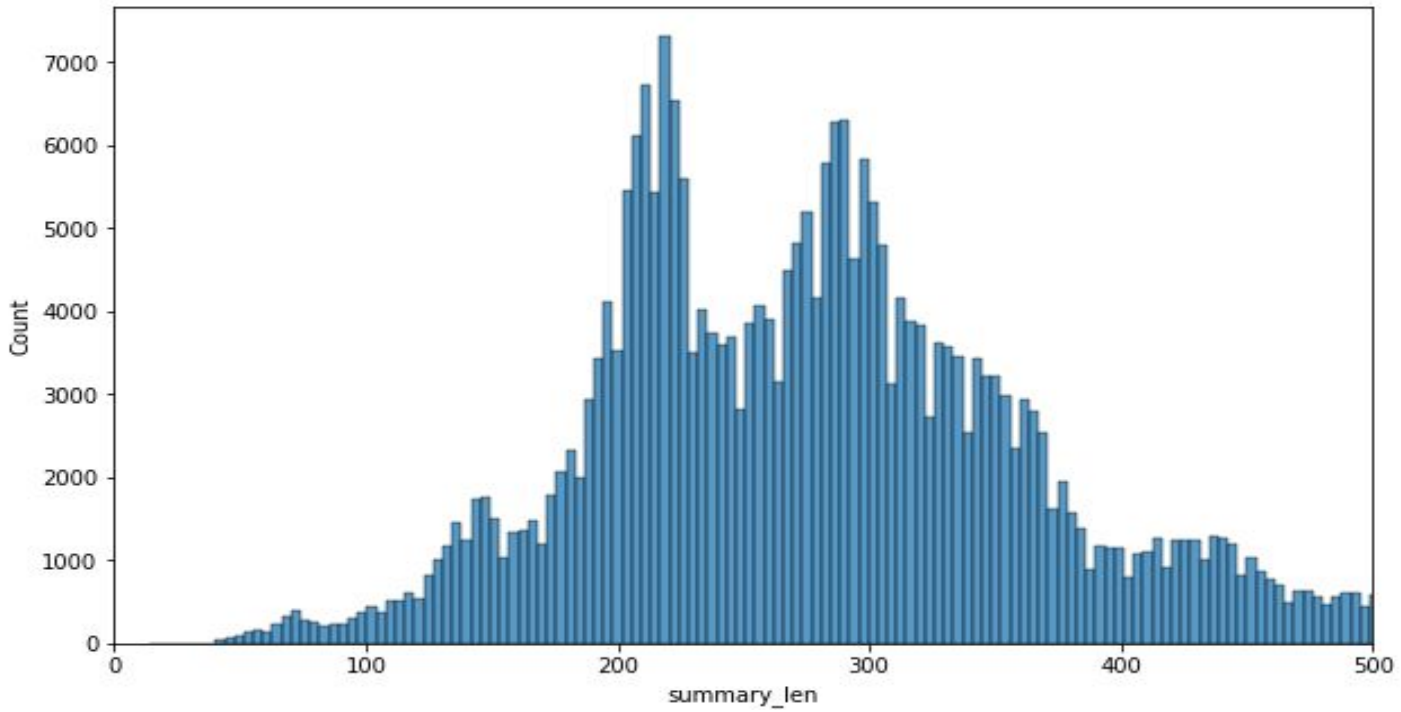
News ID	N24416
Category	news
SubCategory	newsscienceandtechnology
Title	Google to fix 'bug' that uploads free full-quality iPhone pics to Photos
Abstract	Google is about to patch a quirk in Photos that effectively gives iPhone users a free ride. The company told Android Police in a statement that it's planning to fix a Google Photos "bug" that stores iOS photos in their original quality without counting toward Google Drive usage ....
URL	https://assets.msn.com/labs/mind/AAJ4ORa.html

MIND News Dataset Categories



MIND Summary Lengths

- **Articles** Average Length is in the range of **4000** words.
- **Summaries** Average Length is in the range of **300** words.




```
(df.article.str.len()).describe()
```

```
count    287113.000000
mean       4033.660865
std        1954.339234
min          48.000000
25%        2583.000000
50%        3682.000000
75%        5117.000000
max       15925.000000
```

```
(df.highlights.str.len()).describe()
```

```
count    287113.000000
mean        294.770390
std         120.197405
min          14.000000
25%         218.000000
50%         280.000000
75%         342.000000
max        7388.000000
```

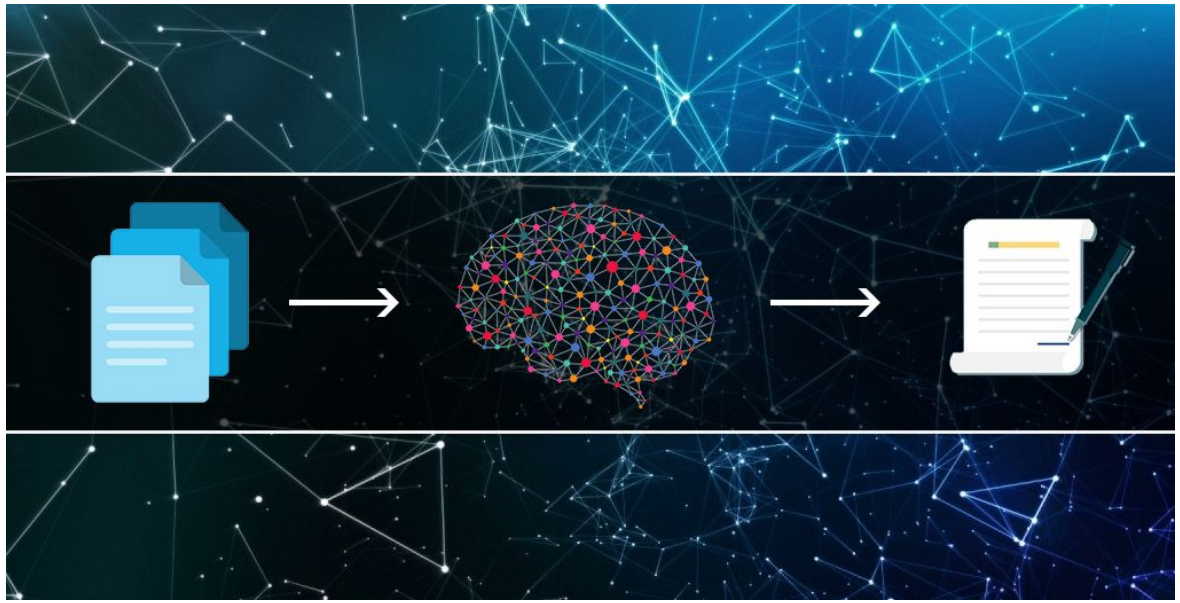
Comparison

Comparison	CNN News	BBC News	MIND News
Dataset Count	300K	2225	160k
Categorized	No	Yes	Yes
Summarized	Yes	Yes	Only Abstract
User's Impressions	No	No	More than 15 million impression logs generated by 1 million users
Pros	Large Dataset with Provided Summary	Summary is Provided	Contains behaviours of users which will help in recommendation
Cons	No Categories or Behaviour dataset	Only 5 Categories and No Behaviour dataset	Need Web Scraping to obtain the articles

Methodology

Methodology

1. Dataset Selection
2. Data Preprocessing
3. Text Summarization
4. Recommendation System
5. Evaluation and Tuning
6. Trying advanced techniques



Technologies

- **Extractive Summarization Techniques** such as:
  - TextRank
  - LexRank
- **Abstractive Summarization Techniques** such as:
  - Deep Learning using Encoder and Decoder based models (BART - Transformer Based)
- **Recommendation Techniques** such as:
  - Rank Based
  - Matrix Factorization

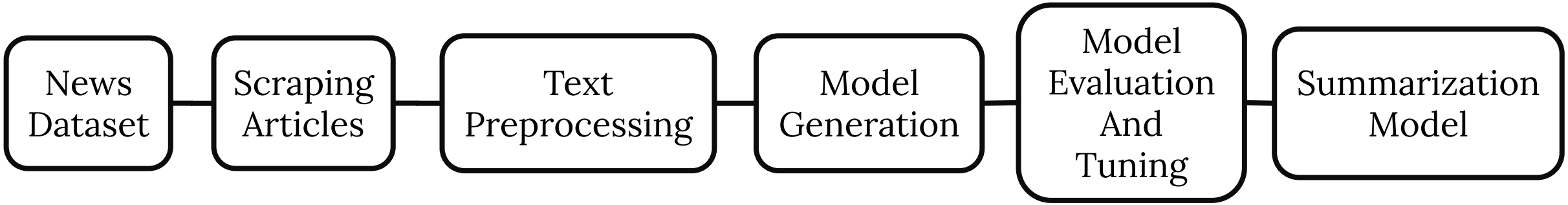
News Feed Modules

News Feed Modules

- We have worked on **Two modules concurrently**:
  A) News Feed **Summarization** Module
  B) News Feed **Recommendation** Module

News Feed Summarization

News Feed Summarization



Articles Scraping

- **<u>Using the URLS in provided in MIND News dataset news:</u>**
  - We scraped the article for each URL using **Trafilatura** package.
  - Then used **Newspaper3k** package for **Video Category**, because **Trafilatura** It can't recognize the difference between text and videos which **Newspaper3k** package could do.

Articles Scraping Packages

Comparison	Trafilatura	Newspaper3k
Pros	• Very good with Text only • Time Saving	• Mainly for News articles • Text and image extraction from html • Keyword, Summary and Author extraction from text • Works in 10+ languages
Cons	• It can't recognize the difference between text, videos and images	• Time consuming

Text Preprocessing

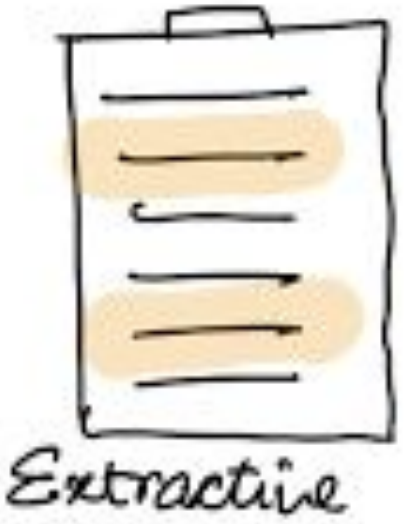
- **We applied the following Text preprocessing :**
  1. Convert text to lowercase
  2. Contraction mapping
  3. Handling commas (‘s)
  4. Remove HTML tags
  5. Remove any text inside the parenthesis “( )”
  6. Remove punctuations and special characters
  7. Handling stopwords
  8. Text stemming

Model Generation

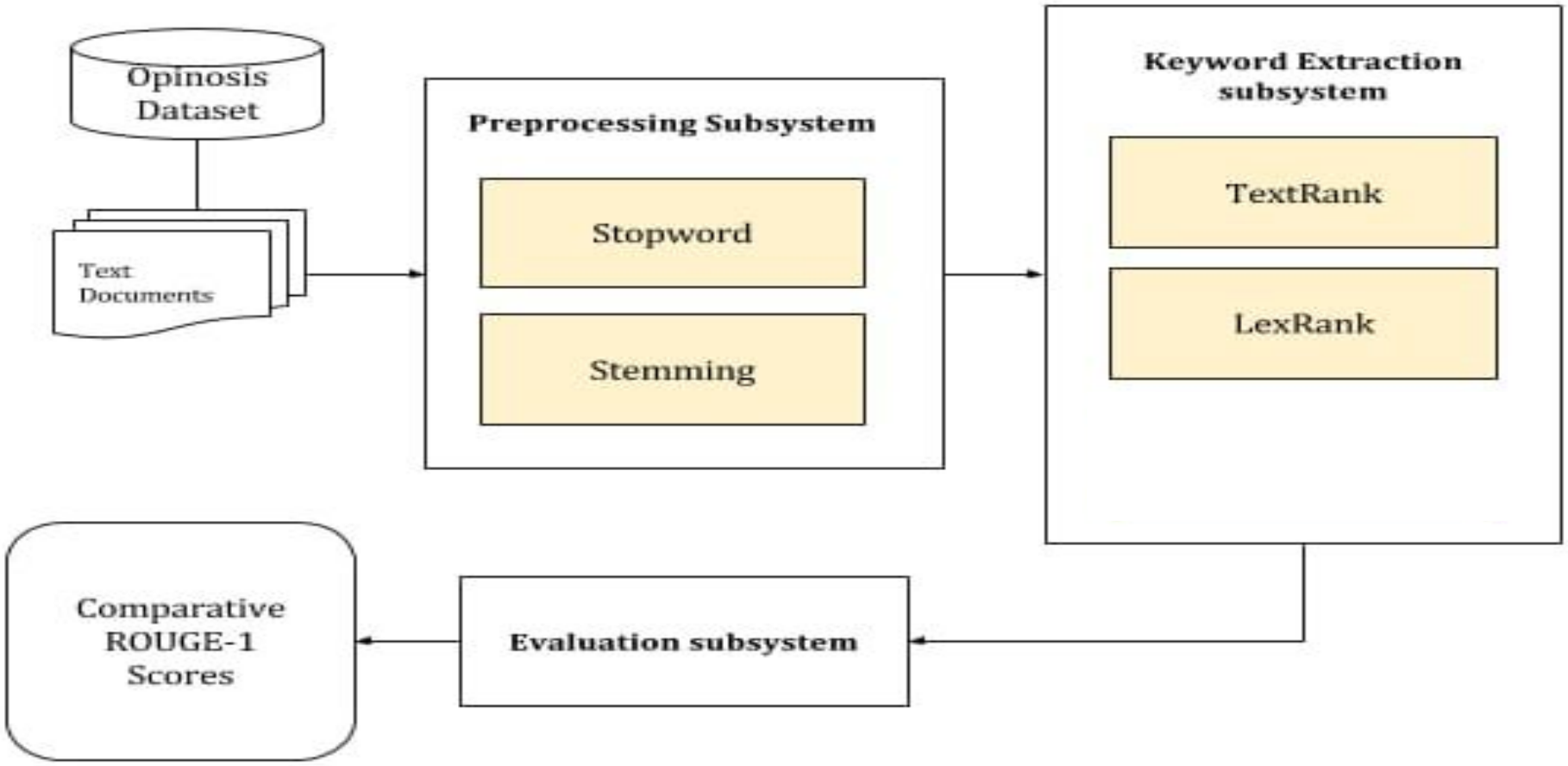
- We tackled both **Extractive and Abstractive** text summarization techniques :
  - **Extractive Text Summarization:**
    - We used a **rank based** extractive sentence selection algorithm to ensure a pure sentence abstraction.
  - **Abstractive Text Summarization:**
    - We used several novel sentence abstraction techniques which jointly perform sentence compression, fusion, and paraphrasing at the sentence level.

Extractive Text Summary

- It is the **traditional** method developed first.
- The main objective is to identify the **significant** sentences of the text and **add** them to the summary.
- You need to note that the summary obtained contains **exact sentences** from the original text.
- **<u>We used two extractive methods:</u>**
  - TextRank
  - LexRank

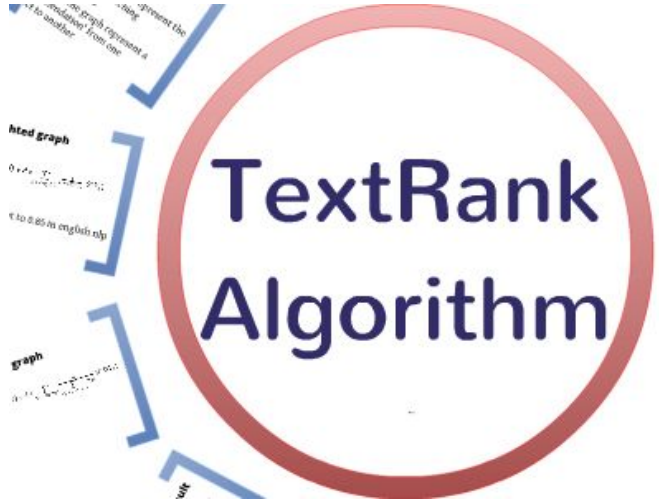


Extractive Summary Pipeline

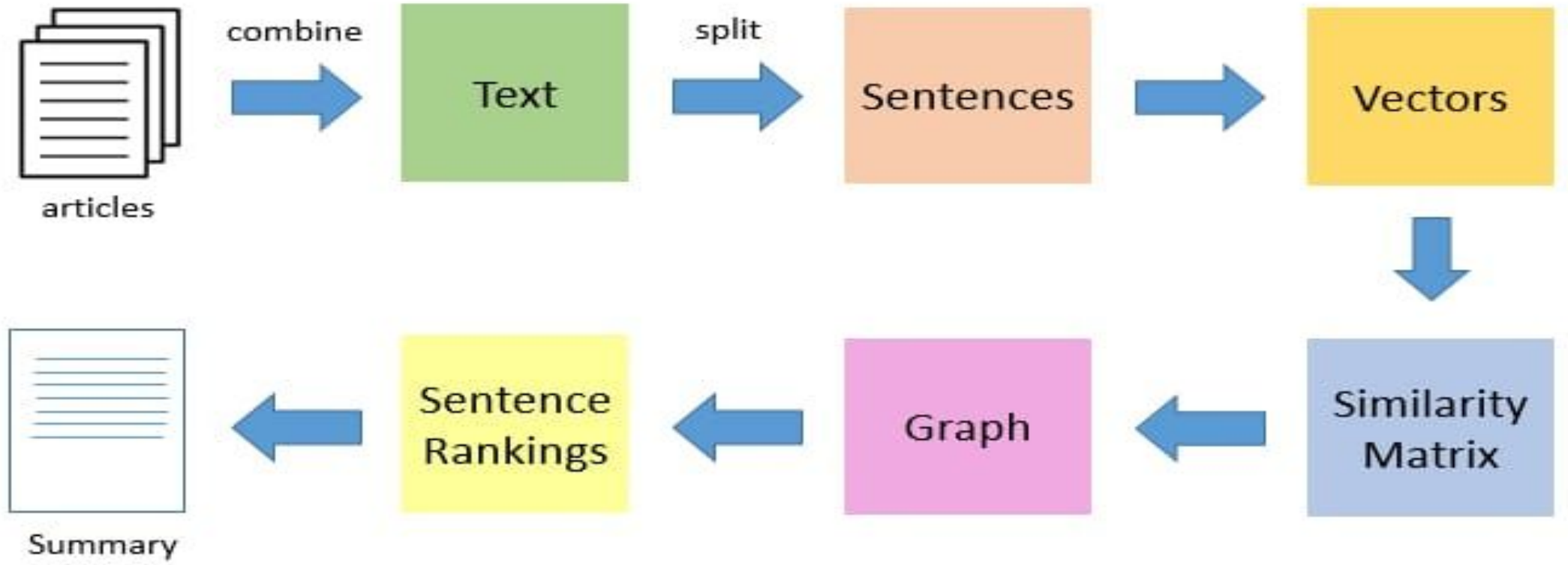


TextRank Extractive Summary

- It is based on the concept that words which occur **more frequently** are significant.
- Hence, the sentences containing highly **frequent words** are important.
- TextRank algorithm assigns **scores** to each sentence in the text.
- The **top-ranked sentences** make it to the summary.



TextRank Extractive Summary



TextRank Model Architecture

TextRank Extractive Summary

Article " It's been Orlando's hottest October ever so far, but cooler temperatures on the way "¶

there won't be a chill down to your bones this halloween in orlando , unless you count the sweat dripping from your armpits . halloween temperatures are supposed to come near or tie the record for the hottest halloween in orlando , but the month of october has already beaten the record for the hottest october ever recorded in the city beautiful , according to the national weather service . the record to beat was an average of 802 degrees with two days remaining 2019's october is on track to record 809 degrees , said nws meteorologist derrick weitlich . yeah with just two days left , there's no way it won't break the record . this october has been above normal , weitlich said . daytona and vero beach are also expected to hit record breaking months . so why is it so hot was central florida cursed by a coven of witches the answer is less magical and more meteorological as a ridge , or an area of blocking high pressure , has been sitting over florida preventing frontal passages of cooler air from entering , weitlich said . that's been the trend this month , he said . so while the central us . gets colder temperatures , we're getting warmer than normal . the hottest halloween on record came in 1992 when pumpkins wilted in the heat of 90 degrees . this halloween is expected to hit a temperature of 89 degrees , but the possibility for hotter temperatures does exist , weitlich said . this week's forecast isn't all heat though . cooler temperatures are expected to start this weekend where the high is forecast to be in the upper 70s , and the low could be in the low 70s by saturday morning . saturday night should see lows in the mid 60s , weitlich said . but temperatures are expected to return to the 80s by monday . we'll get , at least , a taste of fall , weitlich said .'

.....

Summary (TextRank)¶

'halloween temperatures are supposed to come near or tie the record for the hottest halloween in orlando , but the month of october has already beaten the record for the hottest october ever recorded in the city beautiful , according to the national weather service .cooler temperatures are expected to start this weekend where the high is forecast to be in the upper 70s , and the low could be in the low 70s by saturday morning .' .....

Our Abstract Ref¶

'there will not be a chill down to your bones this halloween in orlando , unless you count the sweat dripping from your armpits .'

Sample of TextRank Result on MIND Dataset

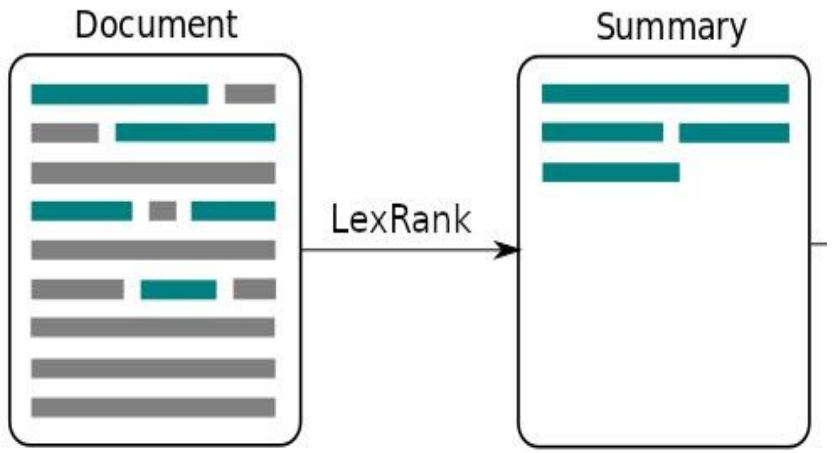
TextRank Extractive Summary

```
The Whole Article :
-------------------
 the eu and us have agreed to begin talks on ending subsidies given to aircraft makers , eu trade commissioner peter
mandelson has announced . , both sides hope to reach a negotiated deal over state aid received by european aircraft
maker airbus and its us rival boeing , mr mandelson said . airbus and boeing accuse each other of benefiting from il
legal subsidies . mr mandelson said the eu and us hoped to avoid having to resolve the dispute at the world trade or
ganisation wto . , with this agreement the eu and us have confirmed their willingness to resolve the dispute which h
as arisen between them , mr mandelson said . i hope our negotiations in the next three months will lead to an agreem
ent ending subsidies to development and production of large civil aircraft . last year , the us terminated an agreem
ent with the eu , reached in 1992 , which limits the subsidies countries can hand over to civil aircraft makers . th
e us filed a complaint against brussels with the wto over state aid to airbus , prompting a retaliatory eu complaint
over us support for boeing . however , both sides agreed to suspend their requests for wto arbitration at the beginn
ing of december , to allow bilateral talks to continue . eads and bae systems , the european defence and aerospace f
irms which own airbus , welcomed mr mandelson's announcement . it has always been preferable that any differences be
tween the us and europe on this matter be overcome through constructive discussion rather than through legal recours
e , the companies said in a joint statement . , separately , the world's largest package delivery company , ups , sa
id it had placed an order for 10 airbus a380 superjumbo freight carrying jets , with an option to buy 10 more of the
triple decker aircraft . the us company said it needed to expand its air freight capacity following strong internati
onal growth , and would begin receiving deliveries of the a380s from 2009 . however , ups said it was cutting a prev
ious order for smaller airbus a300s from 90 planes to 53 . so far , airbus has delivered 40 a300s to ups . airbus ov
ertook boeing as the world's largest manufacturer of commercial airliners in 2003 .
--------------------------------------------------------------------------------------------------------------------
Summary (TextRank) :
--------------------
 the eu and us have agreed to begin talks on ending subsidies given to aircraft makers , eu trade commissioner peter
mandelson has announced . , both sides hope to reach a negotiated deal over state aid received by european aircraft
maker airbus and its us rival boeing , mr mandelson said . , separately , the world's largest package delivery compa
ny , ups , said it had placed an order for 10 airbus a380 superjumbo freight carrying jets , with an option to buy 1
0 more of the triple decker aircraft .
--------------------------------------------------------------------------------------------------------------------
Our Summary Ref :
-----------------
 both sides hope to reach a negotiated deal over state aid received by european aircraft maker airbus and its us riv
al boeing , mr mandelson said . with this agreement the eu and us have confirmed their willingness to resolve the di
spute which has arisen between them , mr mandelson saidthe us filed a complaint against brussels with the wto over s
tate aid to airbus , prompting a retaliatory eu complaint over us support for boeingthe eu and us have agreed to beg
in talks on ending subsidies given to aircraft makers , eu trade commissioner peter mandelson has announcedmr mandel
son said the eu and us hoped to avoid having to resolve the dispute at the world trade organisation wto . separately
, the world's largest package delivery company , ups , said it had placed an order for 10 airbus a380 superjumbo fre
ight carrying jets , with an option to buy 10 more of the triple decker aircrafthowever , ups said it was cutting a
previous order for smaller airbus a300s from 90 planes to 53 .
```

Sample of TextRank Result on BBC Dataset

LexRank Extractive Summary

- LexRank is based on the concept that sentence which is **similar** to many other sentences of the text has a **high probability** of being important.
- The approach of LexRank is that a particular sentence is recommended by **other similar sentences** and hence is ranked higher.
- **Higher the rank**, higher the priority of being included in the summarized text.



LexRank Extractive Summary

Article " It's been Orlando's hottest October ever so far, but cooler temperatures on the way "¶

there won't be a chill down to your bones this halloween in orlando , unless you count the sweat dripping from your armpits . halloween temperatures are supposed to come near or tie the record for the hottest halloween in orlando , but the month of october has already beaten the record for the hottest october ever recorded in the city beautiful , according to the national weather service . the record to beat was an average of 802 degrees with two days remaining 2019's october is on track to record 809 degrees , said nws meteorologist derrick weitlich . yeah with just two days left , there's no way it won't break the record . this october has been above normal , weitlich said . daytona and vero beach are also expected to hit record breaking months . so why is it so hot was central florida cursed by a coven of witches the answer is less magical and more meteorological as a ridge , or an area of blocking high pressure , has been sitting over florida preventing frontal passages of cooler air from entering , weitlich said . that's been the trend this month , he said . so while the central us . gets colder temperatures , we're getting warmer than normal . the hottest halloween on record came in 1992 when pumpkins wilted in the heat of 90 degrees . this halloween is expected to hit a temperature of 89 degrees , but the possibility for hotter temperatures does exist , weitlich said . this week's forecast isn't all heat though . cooler temperatures are expected to start this weekend where the high is forecast to be in the upper 70s , and the low could be in the low 70s by saturday morning . saturday night should see lows in the mid 60s , weitlich said . but temperatures are expected to return to the 80s by monday . we'll get , at least , a taste of fall , weitlich said .'

.....

Summary (LexRank)¶

'this october has been above normal , weitlich said .cooler temperatures are expected to start this weekend where the high is forecast to be in the upper 70s , and the low could be in the low 70s by saturday morning .' .....

Our Abstract Ref¶

'there will not be a chill down to your bones this halloween in orlando , unless you count the sweat dripping from your armpits .'

Sample of LexRank Result on MIND

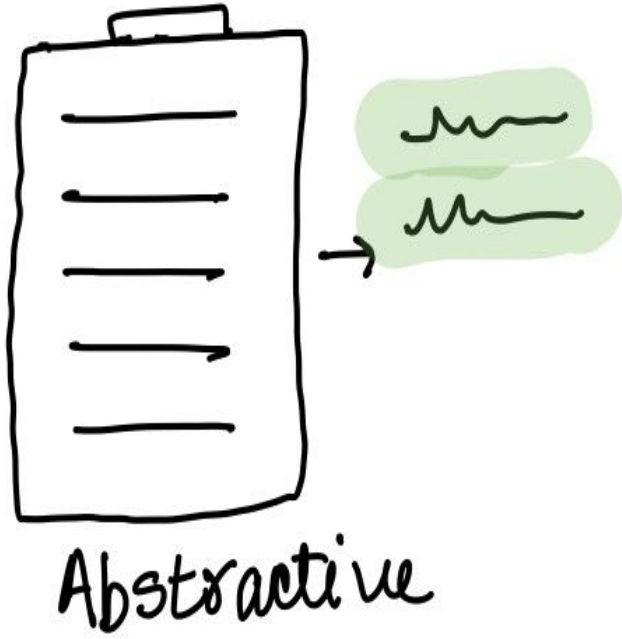
LexRank Extractive Summary

```
The Whole Article :
-------------------
 the eu and us have agreed to begin talks on ending subsidies given to aircraft makers , eu trade commissioner peter
mandelson has announced . , both sides hope to reach a negotiated deal over state aid received by european aircraft
maker airbus and its us rival boeing , mr mandelson said . airbus and boeing accuse each other of benefiting from il
legal subsidies . mr mandelson said the eu and us hoped to avoid having to resolve the dispute at the world trade or
ganisation wto . , with this agreement the eu and us have confirmed their willingness to resolve the dispute which h
as arisen between them , mr mandelson said . i hope our negotiations in the next three months will lead to an agreem
ent ending subsidies to development and production of large civil aircraft . last year , the us terminated an agreem
ent with the eu , reached in 1992 , which limits the subsidies countries can hand over to civil aircraft makers . th
e us filed a complaint against brussels with the wto over state aid to airbus , prompting a retaliatory eu complaint
over us support for boeing . however , both sides agreed to suspend their requests for wto arbitration at the beginn
ing of december , to allow bilateral talks to continue . eads and bae systems , the european defence and aerospace f
irms which own airbus , welcomed mr mandelson's announcement . it has always been preferable that any differences be
tween the us and europe on this matter be overcome through constructive discussion rather than through legal recours
e , the companies said in a joint statement . , separately , the world's largest package delivery company , ups , sa
id it had placed an order for 10 airbus a380 superjumbo freight carrying jets , with an option to buy 10 more of the
triple decker aircraft . the us company said it needed to expand its air freight capacity following strong internati
onal growth , and would begin receiving deliveries of the a380s from 2009 . however , ups said it was cutting a prev
ious order for smaller airbus a300s from 90 planes to 53 . so far , airbus has delivered 40 a380s to ups . airbus ov
ertook boeing as the world's largest manufacturer of commercial airliners in 2003 .
--------------------------------------------------------------------------------------------------------------------
Summary (LexRank) :
-------------------
 the eu and us have agreed to begin talks on ending subsidies given to aircraft makers , eu trade commissioner peter
mandelson has announced . , both sides hope to reach a negotiated deal over state aid received by european aircraft
maker airbus and its us rival boeing , mr mandelson said . , separately , the world's largest package delivery compa
ny , ups , said it had placed an order for 10 airbus a380 superjumbo freight carrying jets , with an option to buy 1
0 more of the triple decker aircraft .
--------------------------------------------------------------------------------------------------------------------
Our Summary Ref :
-----------------
 both sides hope to reach a negotiated deal over state aid received by european aircraft maker airbus and its us riv
al boeing , mr mandelson said . with this agreement the eu and us have confirmed their willingness to resolve the di
spute which has arisen between them , mr mandelson saidthe us filed a complaint against brussels with the wto over s
tate aid to airbus , prompting a retaliatory eu complaint over us support for boeingthe eu and us have agreed to beg
in talks on ending subsidies given to aircraft makers , eu trade commissioner peter mandelson has announcedmr mandel
son said the eu and us hoped to avoid having to resolve the dispute at the world trade organisation wto . separately
, the world's largest package delivery company , ups , said it had placed an order for 10 airbus a380 superjumbo fre
ight carrying jets , with an option to buy 10 more of the triple decker aircrafthowever , ups said it was cutting a
previous order for smaller airbus a300s from 90 planes to 53 .
```

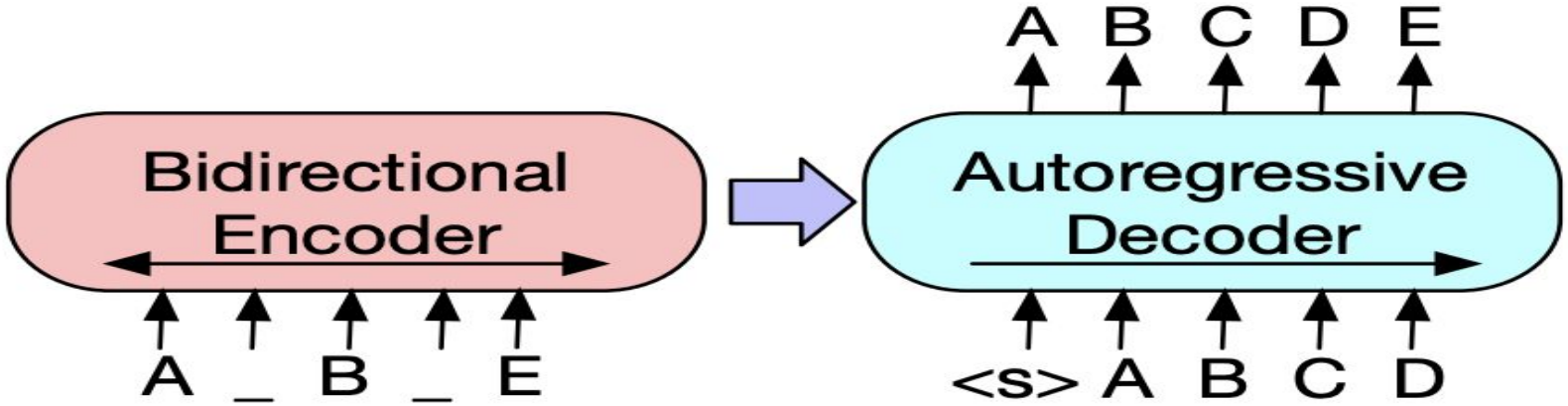
Sample of LexRank Result on BBC Dataset

Abstractive Text Summary

- It is the **new state of art method**, which **generates new sentences** that could best represent the whole text.
- This is **better than** extractive methods where sentences are **just selected** from original text for the summary.
- **<u>Some common abstractive techniques are:-</u>**
  - BART
  - OpenAI
  - GPT
  - T5



Bart Abstractive Summary



- **BART** stands for **Bidirectional and Auto-Regressive Transformer**.
- **BART** is a Transformer that combines the **Bidirectional** Encoder (i.e. BERT like) with an **Autoregressive** decoder (i.e. GPT like) into one Seq2Seq model.

Bart Abstractive Summary

Article " It's been Orlando's hottest October ever so far, but cooler temperatures on the way "¶

there won't be a chill down to your bones this halloween in orlando , unless you count the sweat dripping from your armpits . halloween temperatures are supposed to come near or tie the record for the hottest halloween in orlando , but the month of october has already beaten the record for the hottest october ever recorded in the city beautiful , according to the national weather service . the record to beat was an average of 802 degrees with two days remaining 2019's october is on track to record 809 degrees , said nws meteorologist derrick weitlich . yeah with just two days left , there's no way it won't break the record . this october has been above normal , weitlich said . daytona and vero beach are also expected to hit record breaking months . so why is it so hot was central florida cursed by a coven of witches the answer is less magical and more meteorological as a ridge , or an area of blocking high pressure , has been sitting over florida preventing frontal passages of cooler air from entering , weitlich said . that's been the trend this month , he said . so while the central us . gets colder temperatures , we're getting warmer than normal . the hottest halloween on record came in 1992 when pumpkins wilted in the heat of 90 degrees . this halloween is expected to hit a temperature of 89 degrees , but the possibility for hotter temperatures does exist , weitlich said . this week's forecast isn't all heat though . cooler temperatures are expected to start this weekend where the high is forecast to be in the upper 70s , and the low could be in the low 70s by saturday morning . saturday night should see lows in the mid 60s , weitlich said . but temperatures are expected to return to the 80s by monday . we'll get , at least , a taste of fall , weitlich said .'

.....

Summary (BART)¶

'The month of october has already beaten the record for the hottest october ever recorded in the city beautiful. The hottest halloween on record came in 1992 when pumpkins wilted in the heat of 90 degrees. cooler temperatures are expected to start this weekend where the high is forecast to be in the upper 70s.' .....

.....

Our Abstract Ref¶

'there will not be a chill down to your bones this halloween in orlando , unless you count the sweat dripping from your armpits .'

Sample of BART Result on MIND

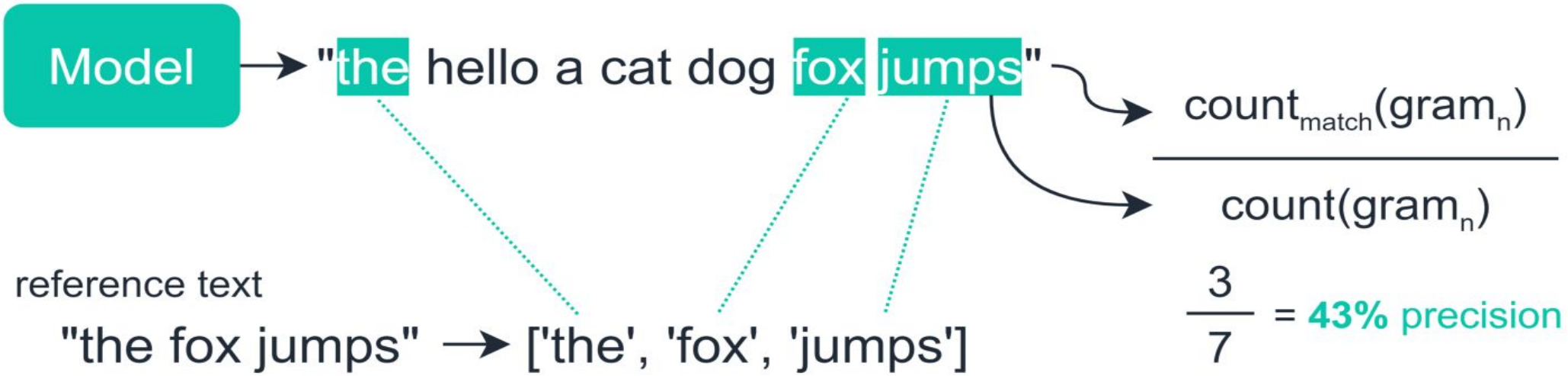
Bart Abstractive Summary

```
The Whole Article :
-------------------
 the eu and us have agreed to begin talks on ending subsidies given to aircraft makers , eu trade commissioner peter
mandelson has announced . , both sides hope to reach a negotiated deal over state aid received by european aircraft
maker airbus and its us rival boeing , mr mandelson said . airbus and boeing accuse each other of benefiting from il
legal subsidies . mr mandelson said the eu and us hoped to avoid having to resolve the dispute at the world trade or
ganisation wto . , with this agreement the eu and us have confirmed their willingness to resolve the dispute which h
as arisen between them , mr mandelson said . i hope our negotiations in the next three months will lead to an agreem
ent ending subsidies to development and production of large civil aircraft . last year , the us terminated an agreem
ent with the eu , reached in 1992 , which limits the subsidies countries can hand over to civil aircraft makers . th
e us filed a complaint against brussels with the wto over state aid to airbus , prompting a retaliatory eu complaint
over us support for boeing . however , both sides agreed to suspend their requests for wto arbitration at the beginn
ing of december , to allow bilateral talks to continue . eads and bae systems , the european defence and aerospace f
irms which own airbus , welcomed mr mandelson's announcement . it has always been preferable that any differences be
tween the us and europe on this matter be overcome through constructive discussion rather than through legal recours
e , the companies said in a joint statement . , separately , the world's largest package delivery company , ups , sa
id it had placed an order for 10 airbus a380 superjumbo freight carrying jets , with an option to buy 10 more of the
triple decker aircraft . the us company said it needed to expand its air freight capacity following strong internati
onal growth , and would begin receiving deliveries of the a380s from 2009 . however , ups said it was cutting a prev
ious order for smaller airbus a300s from 90 planes to 53 . so far , airbus has delivered 40 a300s to ups . airbus ov
ertook boeing as the world's largest manufacturer of commercial airliners in 2003 .
--------------------------------------------------------------------------------------------------------------------
BARTSummary :
-------------------
 The European Union and the United States have agreed to begin talks in the next three months to resolve a dispute o
ver state aid to civil aircraft makers. the eu and us hope to reach a deal by the end of the year, the EU trade comm
issioner has said. the EU and the us have announced the start of talks on ending subsidies for civil aircraft maker
airbus and its us rival, boeing, in the coming months. the two sides have said they will suspend their requests for
arbitration at the World Trade Organisation ( wto) at the beginning of december. the European Commission has announc
ed the talks.
--------------------------------------------------------------------------------------------------------------------
Our Summary Ref :
-------------------
 both sides hope to reach a negotiated deal over state aid received by european aircraft maker airbus and its us riv
al boeing , mr mandelson said . with this agreement the eu and us have confirmed their willingness to resolve the di
spute which has arisen between them , mr mandelson saidthe us filed a complaint against brussels with the wto over s
tate aid to airbus , prompting a retaliatory eu complaint over us support for boeingthe eu and us have agreed to beg
in talks on ending subsidies given to aircraft makers , eu trade commissioner peter mandelson has announcedmr mandel
son said the eu and us hoped to avoid having to resolve the dispute at the world trade organisation wto . separately
, the world's largest package delivery company , ups , said it had placed an order for 10 airbus a380 superjumbo fre
ight carrying jets , with an option to buy 10 more of the triple decker aircrafthowever , ups said it was cutting a
previous order for smaller airbus a300s from 90 planes to 53 .
```

Sample of BART Result on BBC Dataset

Model Evaluation

- We analyse the performance of our model using **ROUGE**.
- It is basically a software package used for **evaluating automatic summarization and machine translation** software in natural language processing.



Model Evaluation

- In general, **TextRank** model achieved better results on MIND.

	rouge-1(Text Rank)	rouge-2(Text Rank)	rouge-l(Text Rank)	rouge-1(Lex Rank)	rouge-2(Lex Rank)	rouge-l(Lex Rank)	rouge-1(BART)	rouge-2(BART)	rouge-l(BART)
r	0.339528	0.171428	0.314699	0.275836	0.134254	0.255765	0.269304	0.085884	0.225588
p	0.237216	0.129529	0.224566	0.247161	0.131568	0.233263	0.236842	0.073255	0.198175
f	0.255521	0.136177	0.240191	0.235558	0.120904	0.220827	0.228463	0.070343	0.190733

TextRank and LexRank Summaries Evaluation Based on Their Rouge Score on MIND Dataset

Model Evaluation

TextRank	LexRank	BART

	rouge-1	rouge-2	rouge-l
r	0.638095	0.532051	0.638095
p	0.985294	0.932584	0.985294
f	0.774566	0.677551	0.774566

	rouge-1	rouge-2	rouge-l
r	0.961905	0.935897	0.961905
p	0.528796	0.447853	0.528796
f	0.682432	0.605809	0.682432

	rouge-1	rouge-2	rouge-l
r	0.342857	0.179487	0.333333
p	0.529412	0.277228	0.514706
f	0.416185	0.217899	0.404624

TextRank, LexRank and BART Summaries Evaluation Based on Their Rouge Score on BBC Dataset

Model Evaluation

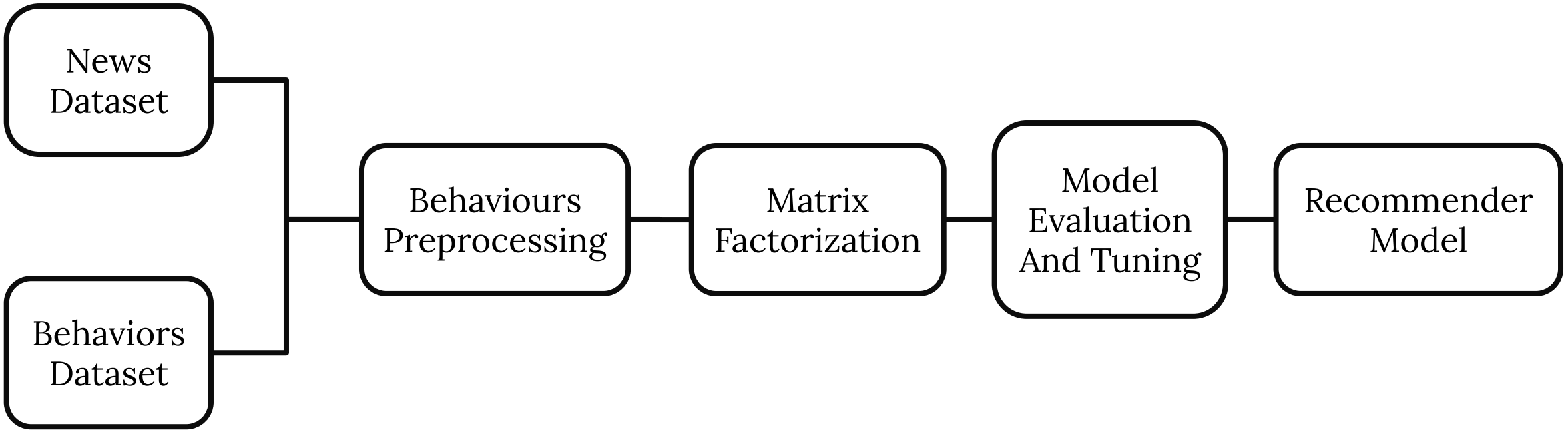
Task	Dataset	Model	Metric Name	Metric Value	Global Rank
Text Summarization	BBC XSum	MatchSum	ROUGE-1	24.86	# 1
			ROUGE-2	4.66	# 1
			ROUGE-L	18.41	# 1

BBC Dataset Benchmark Using MatchSum Model

Summarization Model

News Feed Recommendation

News Feed Recommendation



Behaviours Dataset Example

Impression ID	1
User ID	U13740
Time	11/11/2019 9:05:58 AM
History	N55189 N42782 N34694 N45794 N18445 N63302 N10414 N19347 N31801
Impressions	N55189-1 N35729-0

Behaviours PreProcessing

user_id	N61235	N54489	N42597	N7800	N61408	N26415	N20079	N58271	N6099	N57554	...	N25412	N23625	N14709	N42544	N64109	N49108	N51150
U100	1.0	1.0	1.0	2.0	1.0	1.0	1.0	1.0	1.0	1.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
U1000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
U10001	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
U10003	NaN	1.0	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
U10008	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
U11634	NaN	1.0	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
U11636	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
U11639	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
U11640	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
U11641	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN

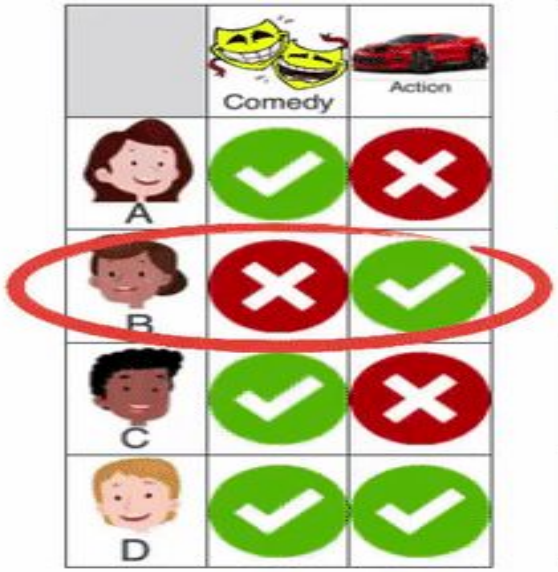
Behaviours PreProcessing

1) **Read** Behaviour and News Dataset.
2) **Splitted** the user's interactions,
   and create dictionary with those interactions.
3) **Collected** users' interactions into single record.
4) **Unpivoted** the updated behavior dataframe.
5) **Merged** the resulted dataframe with news dataset.

Matrix Factorization

Matrix Factorization

	M1	M2	M3	M4	M5
Comedy	3	1	1	3	1
Action	1	2	4	1	3



	M1	M2	M3	M4	M5
A	3	1	1	3	1
B	1	2	4	1	3
C	3	1	1	3	1
D	4	3	5	4	4

Matrix Factorization

1) **Using Matrix Factorization Library:** (Link)

   a) **Splitted** dataset into training, testing and validation dataset.
   b) **Initialize** the matrix factorization model and fit the training data on it.
   c) **Updated** the model's parameters using Validation dataset.
   d) **Tested** the model on the testing dataset.
   e) **Calculated** the RMSE of the testing dataset with the predicted output.

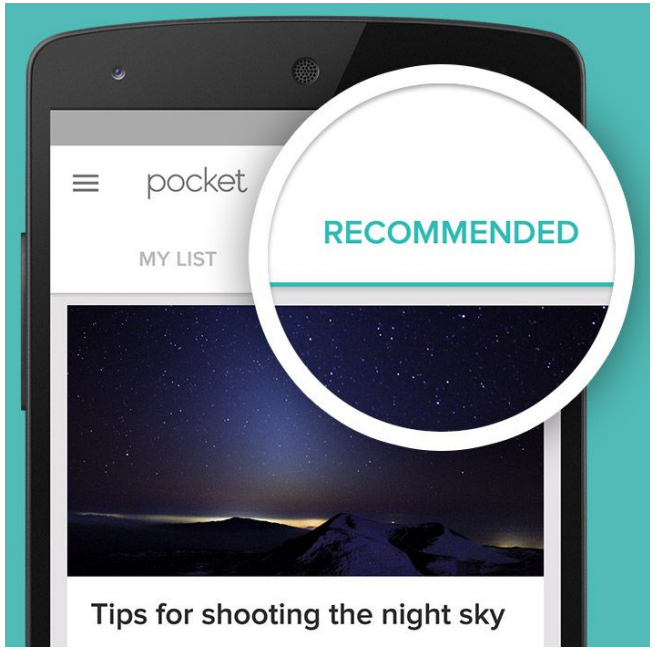
Matrix Factorization

2) **Using Surprise Library:** (Link)
   a) **Splitted** dataset into training and testing dataset using k folds.
   b) **Initialized** SVD Model.
   c) **Looped** for each training and testing, **fit** the model on the train and **get** the predictions.
   d) **Calculated** the Precision of the testing dataset with the predicted output.

Matrix Factorization

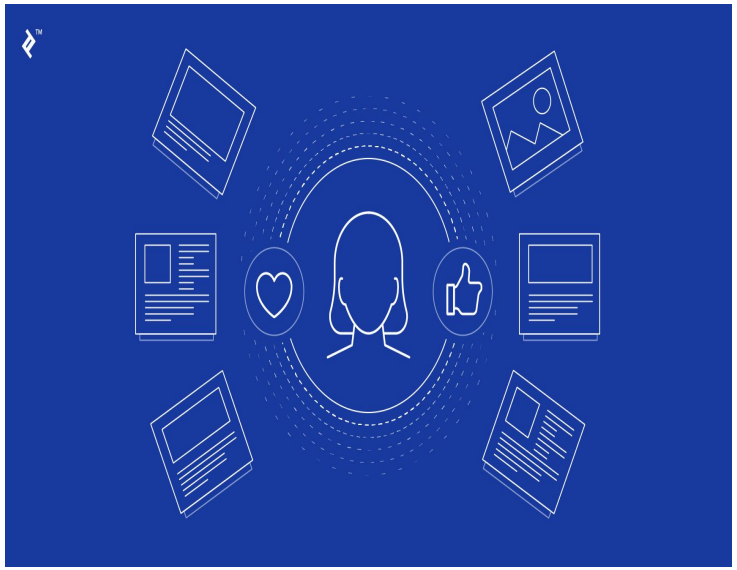
- **<u>We used 2 approaches:</u>**
  - Using Articles.
  - Using SubCategories.

- Both yield **nearly** same precision in surprise library so we proceeded with **articles** dataset.
- We also needed our results to provide **diversity** since many users have only viewed **few articles**.



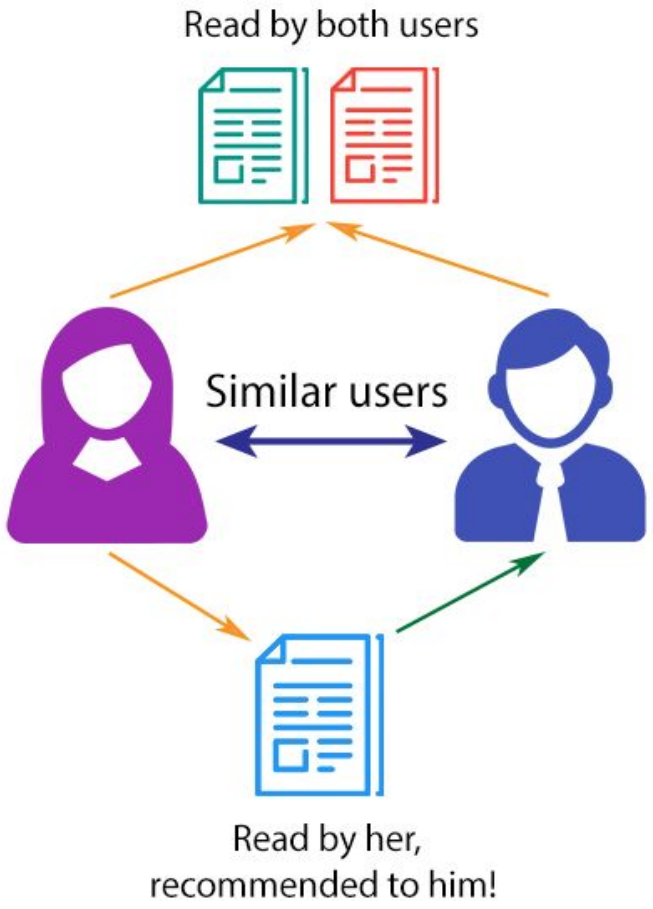
Cold Start Problem

- As for **new users**, we don't have previous interactions history.
- This Problem is known as "**Cold Start Problem**".
- They will be directed to a **welcome page** with **top trending articles** among all categories based on number of clicks. (**Rank Based**)



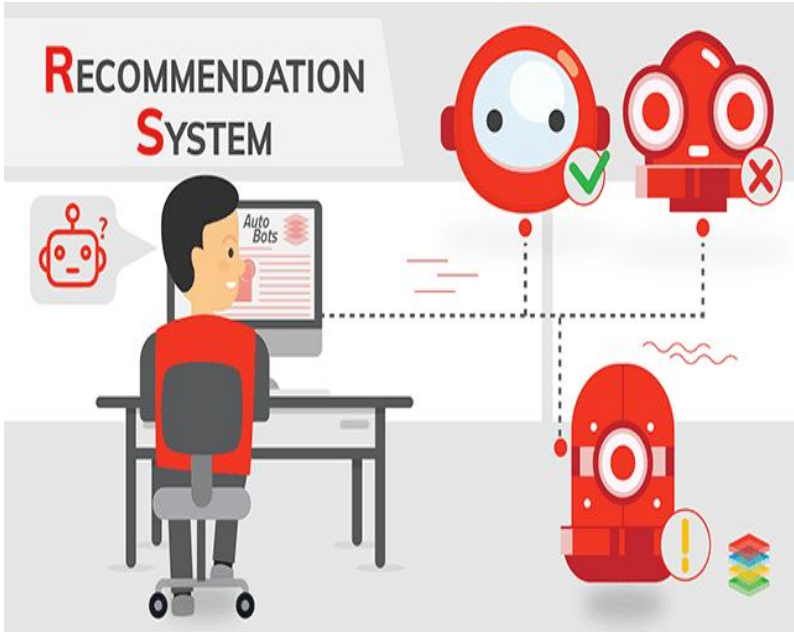
Model Evaluation

- **State-of-the-art precision values** in similar cases range from **10%** up to **30%** depending on dataset used.
- We used **precision@20** as our metric with threshold of **0.15** .
- Applying grid search to tune our model boosted precision up to **19%** .
- After parameter tuning, we ended up using **50 latent features** to capture interests of our users and different features for our articles.



Recommendation Model

- After training, we saved our model to be ready to generate top **50 predictions** for all users in offline mode as **Json file**.
- Using the **recommend function** which take the **User ID** and **no. of recommendations** we want to appear for this user.



Deployment

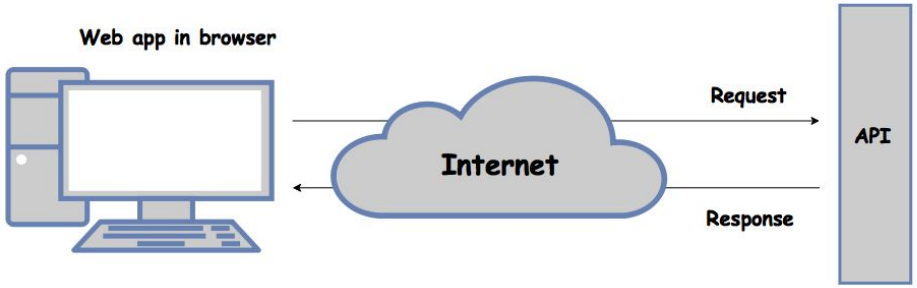
Deployment

- Used **FAST API** and **HTML Templates**.

- Exposing different APIs to generate predictions for users, by passing **User ID** as login and display **top recommended articles** for them.

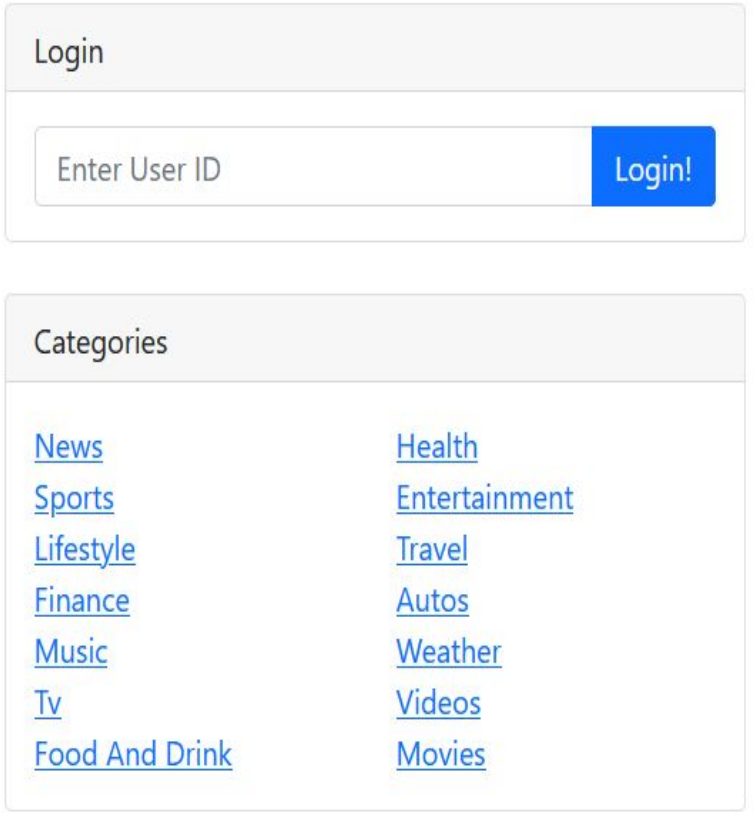
Deployment

- **Requests and Responses :-**
  - **/Home :-**
    - Shows the **top trending articles**.
  - **/Recommend :-**
    - Shows the **recommended articles** for the **logged-in** user if the user was found in dataset, **else** show top trending articles.
  - **/Category :-**
    - Shows top trending articles in the **clicked category**.



Deployment

- **Side Widgets :**
  - **Login:**
    - Entering **user ID** to show related user's predictions.
  - **History:**
    - To show the **logged-in** user's previous clicked articles.
  - **Categories:**
    - To show **top articles** in each category.



Demo

Demo

Autos

The Volkspod VW Beetle Minibikes Are Part Creepy, Part Adorable

Abstractive Summary:- The volkspods, a pair of tiny motorcycles built using volkswagen beetle parts, have gone on sale on ebay for $10,000 (â£6,000) and are already being snapped up by people on social media. and we can’t get enough of them.

Extractive Summary:- Weird things cycleweird the pulse autocycle cycleweird custom 1930 henderson kj streamline meet the volkspods a duo of tiny bikes built using volkswagen beetle parts—including the front fenders . a tiny 79cc volkswagen engine is mounted to an equally tiny tubular frame , topped with a modified first generation beetle fender .

Read more →

Demo

Foodanddrink

65 Best Fall Soups That Will Warm You and Your Family Up All Season Long

Abstractive Summary:- There's nothing quite like a hot bowl of soup to help you keep warm in the winter, and we've rounded up some of the best fall soups to try out this week. and you can find them all here... and here.
Extractive Summary:- Not to mention , there is really nothing like curling up on your couch with a hot bowl of brothy goodness , whether it includes sweet potatoes , butternut squash , or pumpkins—or maybe all three ahead , you will find soups that utilize all the best flavors of fall , plus other comforting additions like bacon , pasta , and more . whether you are in the mood for a classic chicken noodle soup or something a little more elevated , like a curry noodle bowl , you are bound to encounter a few fall soups that speak to your taste buds .

Read more →

Demo

Travel

FAA threatened to ground 38 Southwest Airlines jets over maintenance concerns, report says

Abstractive Summary:- Southwest airlines has been told it can keep dozens of used foreign-owned planes in the air while it completes a safety review, according to a report in the wall street journal, which cites government documents and people familiar with the matter. and.

Extractive Summary:- Aviation regulators threatened to ground 38 boeing 737 jets operated by southwest airlines last month because the carrier is not able to verify that the aircraft meet all mandatory safety standards , the wall street journal reported , citing government documents . southwest told the wsj it was in compliance with the long term inspection program covering the previously foreign owned 737 used jets , and that none of the discrepancies threatened flight safety .

Read more →

Demo

News Feed Summarization And Recommendation

Conclusion

Conclusion

- **Extractive Summary** was better as our summary reference was extractive (Abstract Summary).
- **Recommendation** Precision nearly equals **20%**.
  - **<u>Results through our testing have proved to provide:</u>**
    - Diversity (Different)
    - Novelty (New)
    - Serendipity (Being Interested for user)
    - Relevance (Connect)

Future Work

Future Work

- Adding a login forum with a database where new users can **register**.
- Adding **search** capabilities to the application where user can search articles by **keywords**.
- Enable user to select his **preferences** among different **subcategories**.
- Building a more **interactive UI**.

Appendix: Github

Github ReadMe

News Feed Summarization and Recommendation

- Summarize key information of a given News feed and Recommend to user a set of summarized articles matching his/her interest.

- Instructed By Artificial Intelligence Professional Training Program (AI-Pro) - Information Technology Institute (ITI)

- Supervised By Orange Innovation Egypt Project

- Presentation Link

Team Members :-

1. Karim Atef Henry
2. Mohamed Ahmed Hassan
3. Fady Nasser Fawzy
4. Hoyam Nabil

Github ReadMe

About Dataset :-

- The MIND dataset for news recommendation was collected from anonymize behavior logs of Microsoft News website.
- The data randomly sampled 1 million users who had at least 5 news clicks during 6 weeks from October 12 to November 22, 2019.
- To protect user privacy, each user is de-linked from the production system when securely hashed into an anonymized ID.
- Also collected the news click behaviors of these users in this period, which are formatted into impression logs.
- The impression logs have been used in the last week for test, and the logs in the fifth week for training.
- For samples in training set, used the click behaviors in the first four weeks to construct the news click history for user modeling.
- Among the training data, the samples in the last day of the fifth week used as validation set.
- This dataset is a small version of MIND (MIND-small), by randomly sampling 50,000 users and their behavior logs.
- Only training and validation sets are contained in the MIND-small dataset.

Technologies Used :-

- **Summarization Techniques:-** Extractive and Abstractive Summarization.
- **Recommendation Techniques:-** Matrix Factorization and Surprise Libraries.

Github ReadMe

Instructions :-

1. Install "requirement.txt" Packages :-
    - "pip install -r requirements.txt"
2. Download "News Dataset", "Behaviour Dataset" and "LexRankSummary Dataset" Then Extract.
    - News Dataset
    - Behaviour Dataset
    - Summary Dataset
3. Place "News.tsv" and "Behaviour.tsv" in "../Data/raw/".
4. Rename "Summary Dataset" to "Summary.p" and Place it in "../Data/Summary/".
5. Run "Recommender_Trainer.py" in "../Recommendation/" to train Recommender Models.
6. Run "Main.py".

Github ReadMe

Project Organization :-

```
├── Data
│   ├── prediction          -> Prediction for all users
│   │   ├── prediction_surprise
│   │   │   ├── predictions_articles.json
│   │   │   └── predictions_subcategories.json
│   │   │
│   │   └── prediction_matrix_factorization
│   │
│   ├── processed           -> Dataset after preprocessing
│   │   ├── Category_df.csv
│   │   ├── df_unpivoted.csv
│   │   ├── SubCat_Df.csv
│   │   └── news_articles.p
│   │
│   ├── raw                 -> The original MIND dataset
│   │   ├── behaviours.tsv
│   │   └── news.tsv
│   │
│   └── summary             -> Generated Summary using Lex Rank
│       └── Summary.p
│
├── Data_Preprocessing      -> Scripts to Preprocessing raw data for modeling
│   └── DataProcessing.py
│
├── Recommendation          -> Scripts to generate the model and predicted json arrays
│   ├── matrix_factorization_recommender.py
│   ├── surprise_recommender.py
│   ├── Router.py
│   └── Recommender_Trainer.py
│
├── Summarization           -> Scripts to generate the model and predicted json arrays
│   ├── app.py
│   ├── summarizer.py
│   └── utils.py
│
├── static                  -> Template Assets,CSS and JS
│   ├── assets
│   │   └── favicon.ico
│   ├── css
│   │   └── styles.css
│   └── js
│       └── scripts.js
│
├── storage                 -> contains Generated model
│   ├── storage_surprise
│   │   ├── model_pickle_sub.joblib
│   │   ├── model_pickle_aricles.joblib
│   │   ├── model_status_sub.json
│   │   └── model_status_articles.json
│   │
│   └── matrix_fact
│       ├── model_pickle_sub.joblib
│       ├── model_pickle_aricles.joblib
│       ├── model_status_sub.json
│       └── model_status_articles.json
│
├── templates               -> contains .html files
│   ├── home.html
│   └── index.html
│
├── main.py                 -> contains Main run file
│
├── README.md               -> The top-level README for developers using this project.
│
├── requirements.txt        -> contains Requirement libraries to run this project
│
└── run.sh                  -> uvicorn redirecting script
```

Github Link

- Github.com/New_feed_Summarization_Recommendation


Thank you.