News Articles Recommendation

Submitted by - Arch Desai, Sohil Parsana, Jay Shah

Motivation

Reading the news online has exploded as the web provides access to millions of news sources from around the world. The sheer volume of articles can be overwhelming to readers. **A key challenge of news service website is help users to find news articles that are interesting to read**. This is advantageous to both users and news service, as it enables the user to rapidly find what he or she needs and the news service to help retain and increase customer base.

Objective

Objective of the project is to build a hybrid-filtering personalized news articles recommendation system which can suggest articles from popular news service providers based on reading history of twitter users who share similar interests (Collaborative filtering) and content similarity of the article and user’s tweets (Content-based filtering).

This system can be very helpful to Online News Providers to target right news articles to right users.

But Why Twitter?

Statistics

- 74% of Twitter users say they use the network to get their news.
- 500 million tweets are sent each day.
- 24% of US adults use Twitter.

How Twitter can be used?

- Based on user's tweets we can know user's interests and can recommend personalized news articles which user would share on Twitter. This can increase news articles and news service's popularity.

Project Flow

1. Collect active Twitter users’ data
2. Analyze users’ tweets
3. Cluster users according to their interests
4. Perform sentiment analysis and topic modeling
5. Collect and analyze news articles
6. Get user’s Twitter handle & Recommend news articles

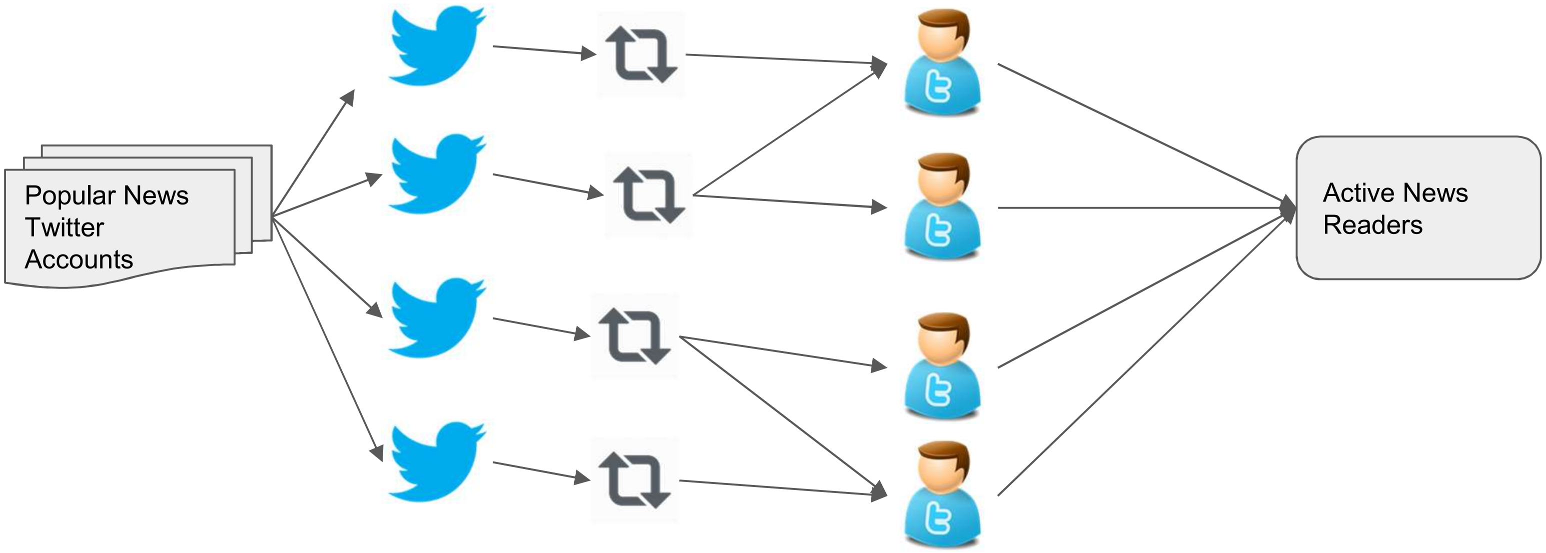
1. Collect active Twitter users’ data

As a first step, the engine identifies readers with similar news interests based on their behavior of retweeting articles posted on Twitter. (**Library** : Tweepy)

The flow of collecting active users’ data:

- Get Twitter users who retweet tweets of New York Times, Bloomberg, Washington Post. We identify them as active news readers.
- Create a popularity Index
  - Popularity = Number of Followers / Number of Friends
- Filter users based on their twitter activity and popularity
- Collect information from Twitter profiles of these filtered users

1. Collect active Twitter users' data - 2



2. Analyze users’ Tweets

The tweets contains URLs, Usernames, non-english words, punctuations and numbers. Sometimes whole tweets are in different languages. To get information from tweets, preprocessing is important. (**Library** : NLTK)

Preprocessing:

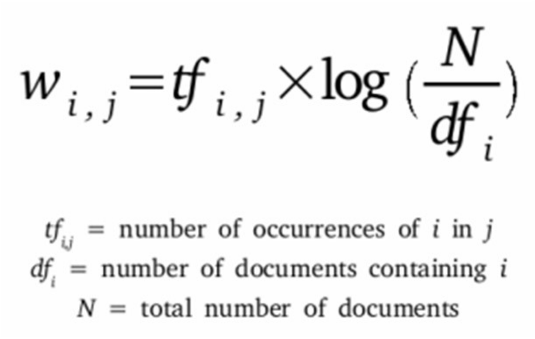
- Clean tweets
  - Removal of URLs, Usernames, numbers, non-english words, punctuations
- Tokenize tweets
  - Process of breaking stream of textual content into words
- Remove Stop words
  - Stop words: Most common words in a languages e.g the, is, am, are, there

- Stemming and Lemmatization
    - Both of these are text normalization techniques used to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.
    - **Stemming** : Chops off words without any context (**PortStemmer**)
        - walking : walk, smiled : smile, houses : house
    - **Lemmatization** : Finds the lemma of words with the use of a vocabulary and morphological analysis of words (**WordNetLemmatizer**)
        - better : good, are : be, ran : run
    - Difference:
        - Caring - Car : Stemming
        - Caring - Care : Lemmatization

3. Cluster users according to their interests

We can cluster users based on their similarity of interests retrieved from their tweets and that requires vectorized representation of tweets.

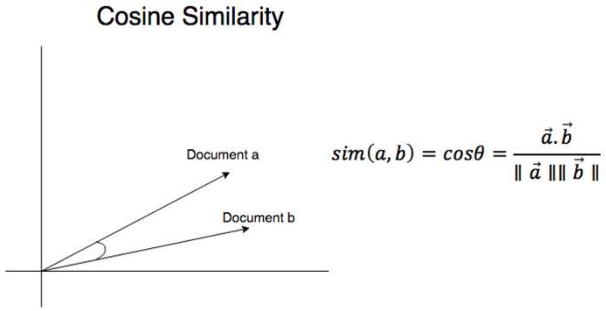
- Find TF-IDF matrix (**Library :** TfidfVectorizer from sklearn)
    - TF-IDF stands for Term Frequency- Inverse Document Frequency
    - TF : Gives frequency of words in each user's tweets
    - IDF : Calculates the weight of rare words across all users' tweets. The words that occur rarely in the corpus have a high IDF score.



$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

tf_{ij} = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

- TF-IDF is a weight that ranks the importance of a term in its contextual document corpus.
- Perform K-means clustering to cluster users based on tf-idf matrix.
- To reduce the dimension of Tf-Idf matrix we define error term, distance matrix
  - Distance Matrix = 1 - Cosine Similarity of users’ tweets
  - Cosine similarity = (dot product of two vectors) / (product of vectors’ magnitudes)
  - The cosine of the angle between the vectors is a good indicator of similarity

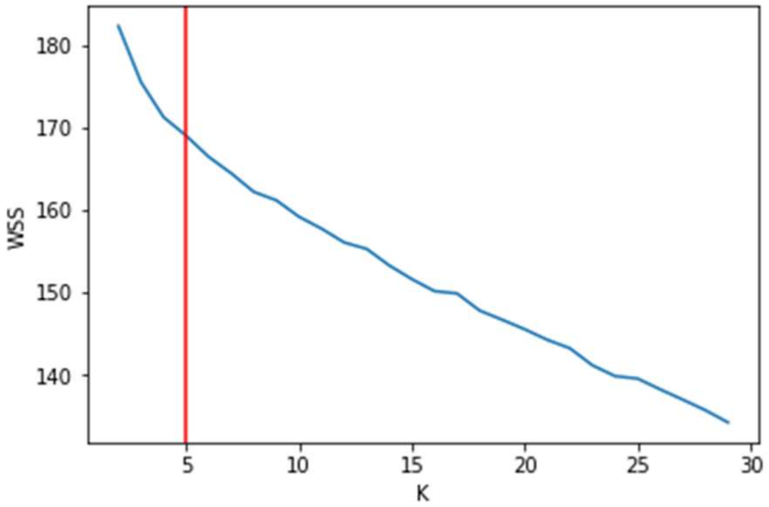

Cosine Similarity

$$sim(a,b) = cos\theta = \frac{\vec{a}.\vec{b}}{\parallel \vec{a} \parallel \parallel \vec{b} \parallel}$$

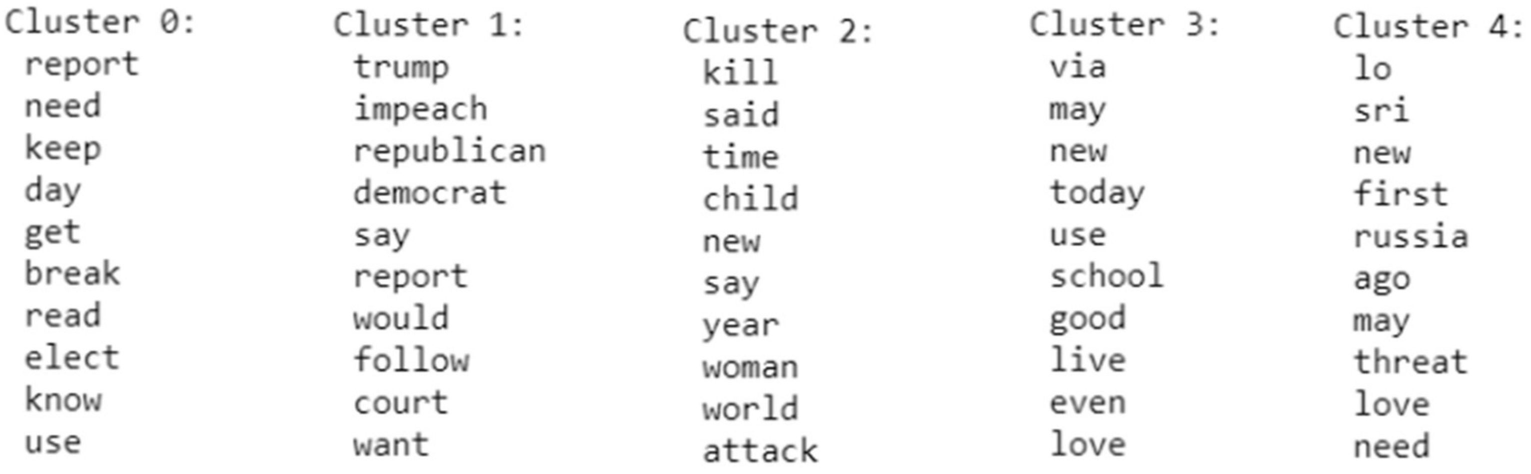
- Reduce dimension matrix using multi-dimension-scaling. (**Library:** Skleran’s MDS)

Selection of optimal K with Elbow Method

The elbow method, in which the sum of squares at each number of clusters is calculated and graphed, and the user looks for a change of slope from steep to shallow (an elbow) to determine the optimal number of clusters.



- We created 5 clusters and top words in 5 clusters are shown below.



```
Cluster 0:        Cluster 1:        Cluster 2:        Cluster 3:        Cluster 4:
 report            trump             kill              via               lo
 need              impeach           said              may               sri
 keep              republican        time              new               new
 day               democrat          child             today             first
 get               say               new               use               russia
 break             report            say               school            ago
 read              would             year              good              may
 elect             follow            woman             live              threat
 know              court             world             even              love
 use               want              attack            love              need
```

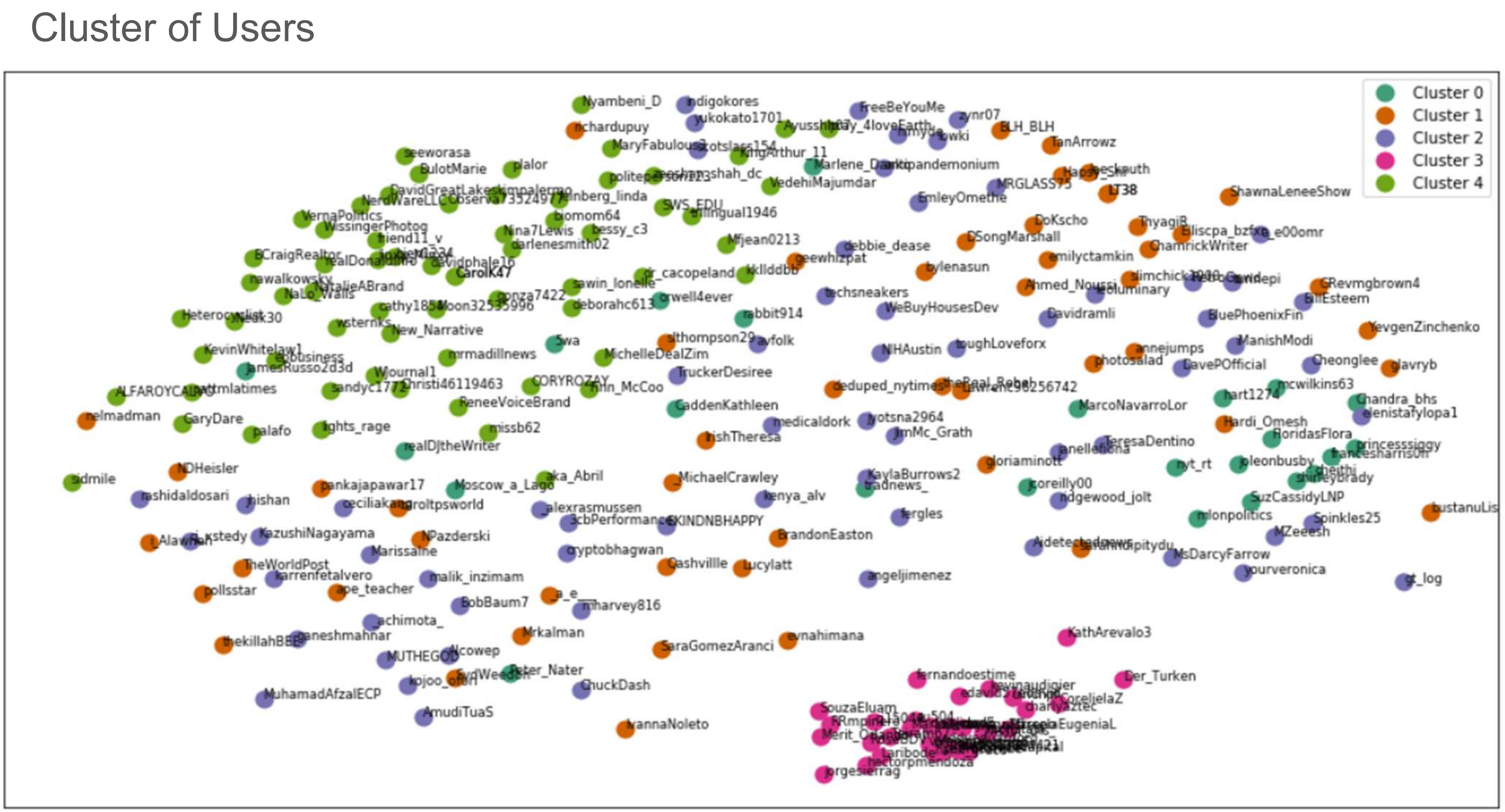
Earth Day

Trump News

Terrorist Attack

Daily Top News

World News


Cluster of Users
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Nyambeni_D
indigokores
yukokato1701
FreeBeYouMe
zynr07
richardupuy
BLH_BLH
TanArrowz
MaryFabulous
KingArthur_11
Marlene_Dan
pandemonium
seeworasa
BulotMarie
plalor
VedehiMajumdar
MRGLASS75
LT38
ShawnaLeneeShow
DavidGreatLakes
NerdWareLLC
linberg_linda
SWS_EDU
EmleyOmethe
VernaPolitics
WissingerPhotog
biomom64
bilingual1946
DoKscho
ThyagiB
Nina7Lewis
bessy_c3
DSongMarshall
friend11_v
darlenesmith02
Mjean0213
debbie_dease
ChamrickWriter
BCraigRealtor
realDonaldTrump
davidphale16
geewhizpat
emilyctamkin
CarolK47
dr_cacopeland
bylenasun
nawalkowsky
NatalieABrand
sawin_lonelle
Ahmed_Noussi
Revmgbrown4
NaLo_Walls
orwell4ever
techsneakers
BillEsteem
Heterocyclist
cathy1854
Moon32535996
deborahc613
rabbit914
WeBuyHousesDev
Davidramli
BluePhoenixFin
wsternks
New_Narrative
YevgenZinchenko
Swa
lthompson29
avfolk
NHAustin
toughLoveforx
annejumps
ManishModi
KevinWhitelaw1
mrmadillnews
MichelleDealZim
photosalad
DavePOfficial
Cheonglee
glavryb
JamesRusso2d3d
Wjournal1
TruckerDesiree
ALFAROYCALVO
sandyc1772
Christi46119463
CORYROZAY
deduped_nytimes
theReal_Rebel
mcwilkins63
hart1274
Chandra_bhs
nelmadman
GaryDare
ReneeVoiceBrand
CaddenKathleen
MarcoNavarroLor
elenistafylopa1
medicaldork
jyotsna2964
Hardi_Omesh
palafo
lights_rage
missb62
JimMc_Grath
FloridasFlora
IrishTheresa
realDJtheWriter
TeresaDentino
princesssiggy
janellefiona
nyt_rt
jleonbusby
NDHeisler
gloriaminott
sidmile
pankajapawar17
aka_Abril
MichaelCrawley
KaylaBurrows2
shirleybrady
Moscow_a_Lago
coreilly00
rashidaldosari
jhishan
ceciliakang
roltpsworld
alexrasmussen
kenya_alv
ridgewood_jolt
SuzCassidyLNP
ferglles
mlonpolitics
Spinkles25
bustanuLis
3cbPerformance
SKINDNBHAPPY
KazushiNagayama
NPazderski
BrandonEaston
MZeeesh
Marissane
cryptobhagwan
MsDarcyFarrow
TheWorldPost
Cashvillle
Lucylatt
karrenfetalvero
malik_inzimam
angeljimenez
yourveronica
gt_log
pollsstar
ape_teacher
a_e
BobBaum7
mharvey816
achimota_
Mrkalman
evnahimana
KathArevalo3
thekillahBEE
ganeshmahnar
MUTHEGOD
Alcowep
SaraGomezAranci
Peter_Nater
kojoo_ofori
fernandoestime
Der_Turken
ChuckDash
MuhamadAfzalECP
SouzaEluam
CoreliaelaZ
AmudiTuaS
IvannaNoleto
Merit_Oa
EugeniaL
hectorpmendoza
jorgesierrag

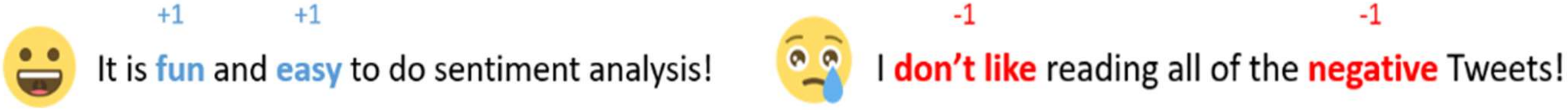
4. Perform Sentiment Analysis and Topic Modelling

Sentiment analysis -Computational study of opinions, sentiments, evaluations, attitudes, appraisal, affects, views, emotions, subjectivity, etc., expressed in text.

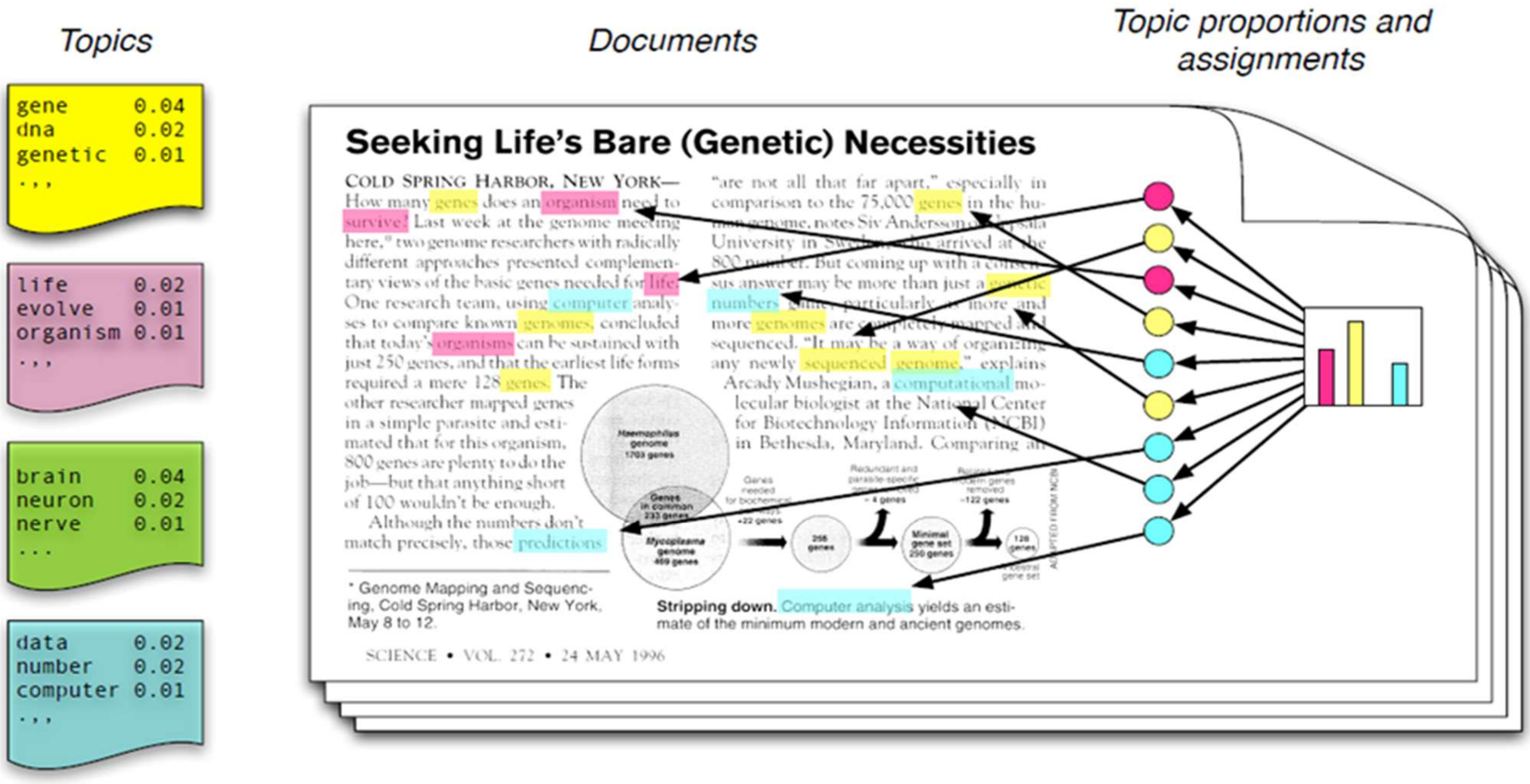
It is also called opinion mining

We have used pretrained model from Textblob library that gives two results:

- Subjectivity and Polarity
- **Polarity** is score between [-1,1], where 0 indicates neutral, +1 indicates a very positive sentiment and -1 represents a very negative sentiment.
- **Subjectivity** is score between [0,1], where 0.0 is very objective and 1.0 is very subjective. Subjective sentence expresses some personal feelings, views, beliefs, opinions, allegations, desires, beliefs, suspicions, and speculations where as Objective sentences are factual.


+1
+1
It is fun and easy to do sentiment analysis!
-1
-1
I don't like reading all of the negative Tweets!

Topic Modelling Motivation



gene 0.04
dna 0.02
genetic 0.01
...

life 0.02
evolve 0.01
organism 0.01
...

brain 0.04
neuron 0.02
nerve 0.01
...

data 0.02
number 0.02
computer 0.01
...

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

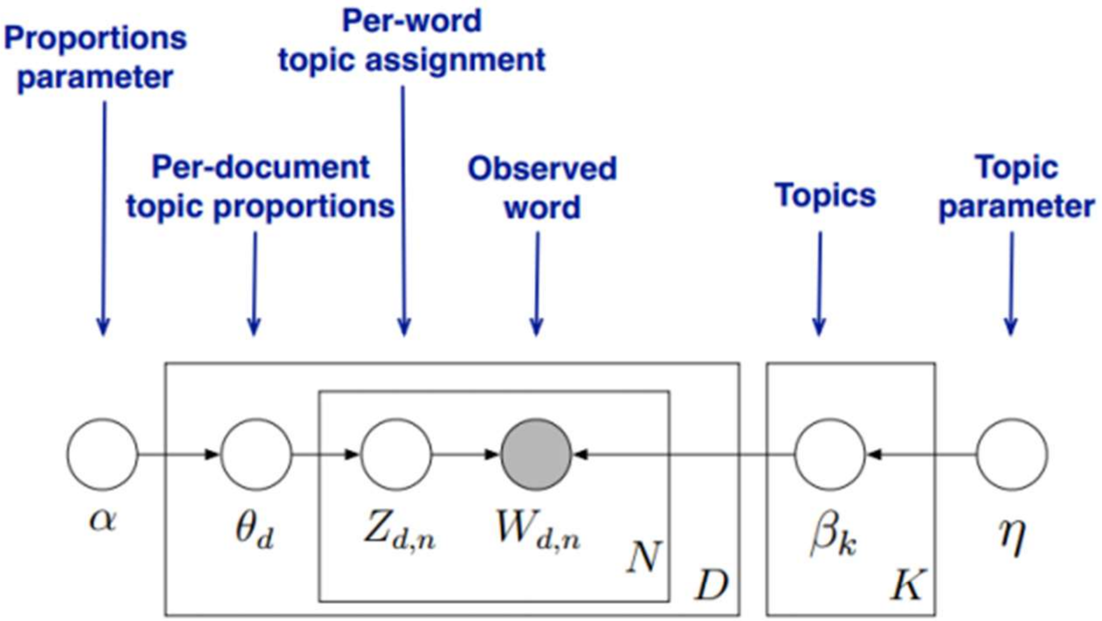
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Model Description



1) Draw each topic $\beta_i \sim Dir(\eta)$ for $i = 1, \dots, K$

2) For each document:

First, Draw topic proportions $\theta_d \sim Dir(\alpha)$

For each word within the document:
a) Draw $Z_{d,n} \sim Multi(\theta_d)$
b) Draw $W_{d,n} \sim Multi(\beta_{z_{d,n}})$

$$p(\beta, \theta, \mathbf{z}, \mathbf{w}) = \left(\prod_{i=1}^{K} p(\beta_i \mid \eta) \right) \left(\prod_{d=1}^{D} p(\theta_d \mid \alpha) \prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right)$$

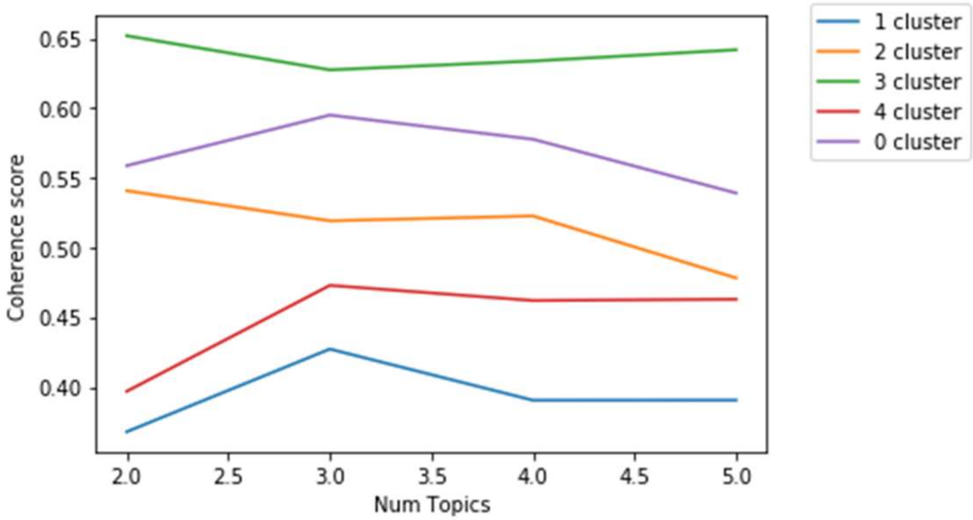
Application in our problem

1. Using LDA Mallet model for topic modelling of each individual cluster. It is more efficient than Gensim’s LDA package requiring *O(corpus).*
2. Tuning of number of topic for each cluster accomplished using the coherence measure: using C_v measure (combining normalized pointwise similarity and cosine similarity)

Topic Model Visualization

Interactive Viz-1

Interactive Viz-2



5. Collect articles

We scraped recent news articles from different news channels using python package : Newspaper3k.

They have different categories so we can train our algorithm using all topics. And our algorithm can satisfy wide range of topics giving good and similar recommendations.

6. Get User’s Twitter Handle and Recommend News articles

There are two main types of collaborative filtering: user-based and item-based. Note that the two are entirely symmetric (or more precisely the transpose of each other)

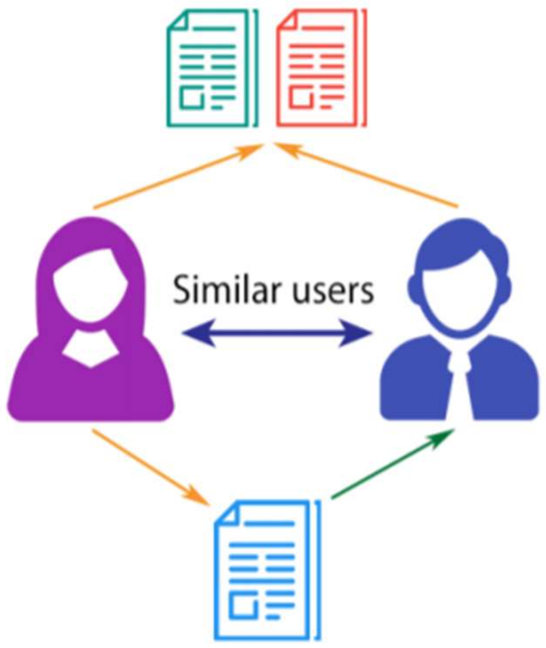
(1) Content-based Filtering

- Based on the tweets of a user, we can identify his or her interests.
- Based on the similarity of user’s interests and news article’s content/tags/headlines, we can recommend news articles.
- The approach has intuitive appeal: If a user posts ten tweets having the word “Clinton,” user would probably like future “Clinton”-tagged news articles.
- Example- Amazon(Recommendation based on recently viwed items)
- Shortcomings of this method:
    - Since, rare words have large weightage in the algorithm it sometimes degrades the performance.
    - For example, if a user’s one tweet contains word “election”, he would get recommendation of news articles tagged election as it is a rare word and higher weightage is given to it.

(2) Collaborative Filtering

- Based on tweets of a user, we can identify a cluster in which user belongs to.
- Based on the topics of each cluster, we can recommend news article to a user
- For example, If a user tweets about election, he or she can be assigned to a cluster of users who have read and retweeted news articles that our user isn't aware of on the topic of election and we can recommend it to a user
- Example - Amazon (Customer who bought this item also bought)
- Shortcoming:
    - this approach fails at recommending newly-published, unexplored articles: articles that are relevant to groups of readers but hadn’t yet been read by any reader in that group.

COLLABORATIVE FILTERING

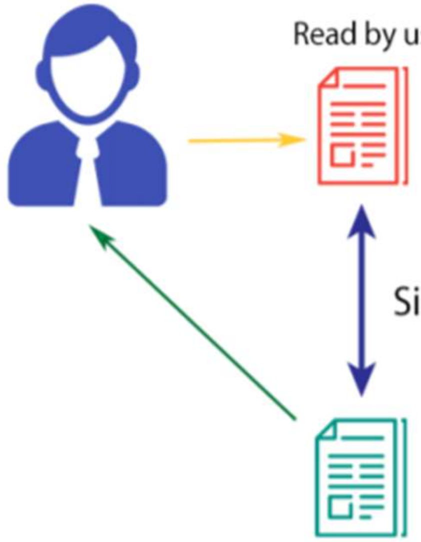


Read by both users

Similar users

Read by her,
recommended to him!

CONTENT-BASED FILTERING



Read by user

Similar articles

Recommended
to user

Hybrid-filtering

In these methods, a combination of both recommendation algorithms are used to maximize advantage and minimize the drawbacks of both algorithms. The different methods for hybridization are shown below.

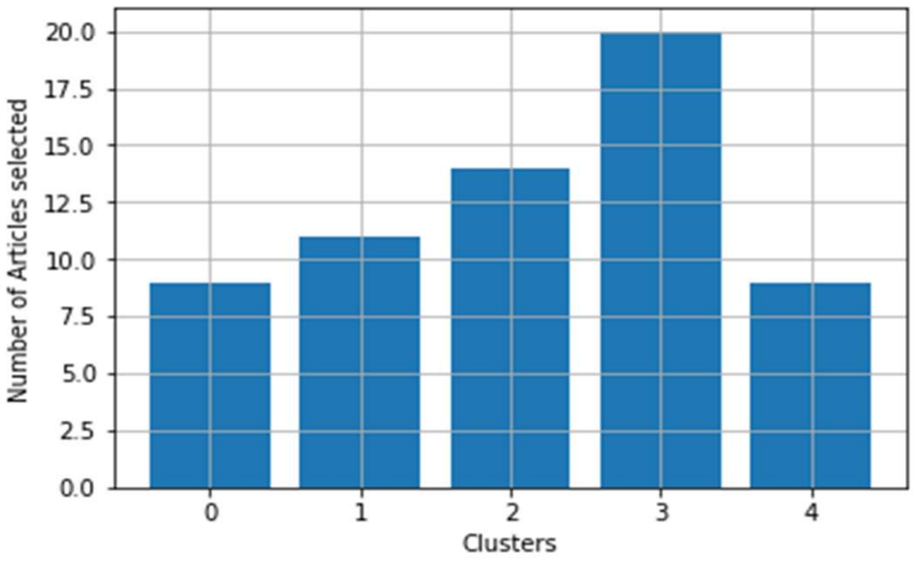
Method	**Description**
Weighted	The scores (or votes) of several recommendation techniques are combined together to produce a single recommendation
Switching	The system switches between recommendation techniques depending on the current situation.For example, in earlier phases, one might use a knowledge-based recommender system to avoid cold-start issues
Mixed	Recommendations from several different recommenders are presented at the same time
Feature combination	Features from different recommendation data sources are thrown together into a single recommendation algorithm. Similarity with Stacking

Our Approach (Feature Combination)

- We get the news articles based on the topics of a cluster in which user belongs to.
- We compare the content of news articles to the content of a user’s tweets and find similarity
- We rank the articles based on similarity with users personal interest and recommend to users.

Collaborative Filtering

		Example Recommended Article	
	Cluster	CNN	NYTimes
Earth Day	0	https://www.cnn.com/2019/04/24/entertainment/top-credit-cards-for-those-with-excellent-credit?	https://www.nytimes.com/2019/04/24/opinion/california-wildfire-climate.html
Trump News	1	https://www.cnn.com/2019/04/24/politics/presidential-tax-returns-states-2020-trump/index.html	http://www.nytimes.com/2019/04/24/us/politics/russia-2020-election-trump.html#commentsContainer
Terrorist Attack	2	https://www.cnn.com/2019/04/24/investing/ford-rivian/index.html	http://www.nytimes.com/interactive/2019/04/23/world/asia/sri-lanka-isis-religious-ethnic-tensions-map.html
Entertainment News	3	https://www.cnn.com/2019/01/30/business/kohls-weight-watchers/index.html	http://www.nytimes.com/2019/04/24/sports/damian-lillard-portland-trail-blazers.html
World News	4	https://www.cnn.com/2019/04/15/australia/australia-racism-media-christchurch-attack-intl/index.html	http://www.nytimes.com/2019/04/24/opinion/rwanda-genocide.html



Hybrid Filtering

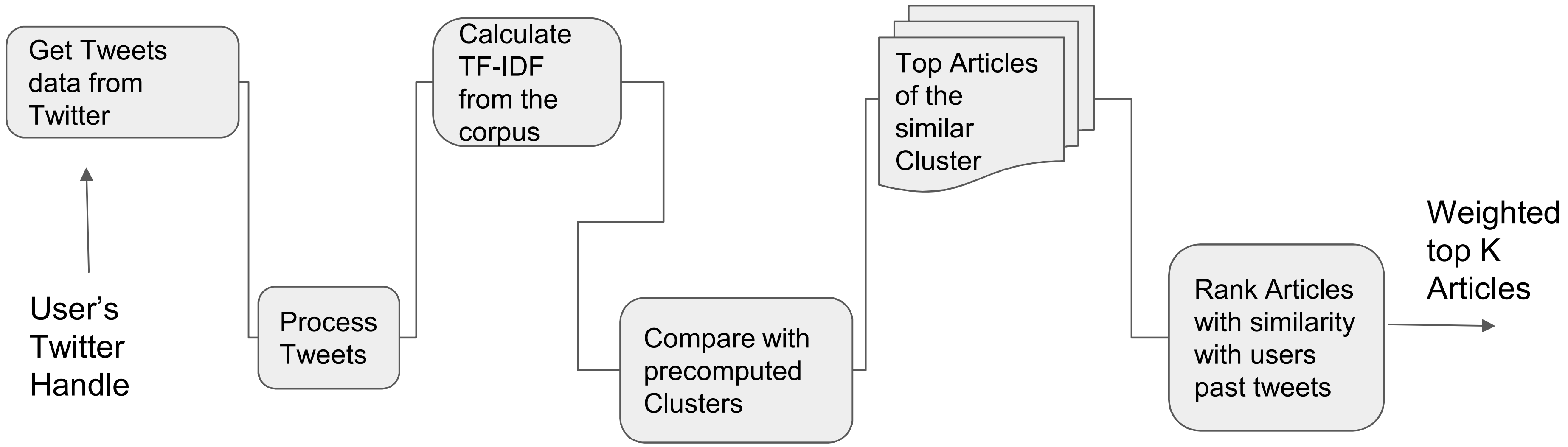
Here we added weight to our recommendation personalized by individual user

After user has been classified into clusters, we will calculate similarity score of user’s interest with identified articles within each clusters.

Weight = Topic Modeling Normalized Prob(80%) + Sentiment Score (20%)

Based on this weight criteria we will rank the articles personalized for each user.

Final Pipeline


User's Twitter Handle
Get Tweets data from Twitter
Process Tweets
Calculate TF-IDF from the corpus
Compare with precomputed Clusters
Top Articles of the similar Cluster
Rank Articles with similarity with users past tweets
Weighted top K Articles

Project Flow Covered

1. Collect active Twitter users' data ✓
2. Analyze users' tweets ✓
3. Cluster users according to their interests ✓
4. Perform sentiment analysis and topic modeling ✓
5. Collect and analyze news articles ✓
6. Get user's Twitter handle & Recommend news articles ✓

Current Challenges:

- How to recommend news articles to a user if he or she does not have any tweets? (Cold Start Problem)
- How to evaluate the performance of a recommendation system?
- Topic Modelling and retrieving users’ data are bottlenecks as they need to be updated very frequently.
- Scale this model on large dataset

Thank You.