Yoga-Veganism: Correlation Mining of Twitter Health Data

Tunazzina Islam

Ph.D. Student

Department of Computer Science

Purdue University, West Lafayette

islam32@purdue.edu https://tunazislam.github.io/ @Tunaz_Islam

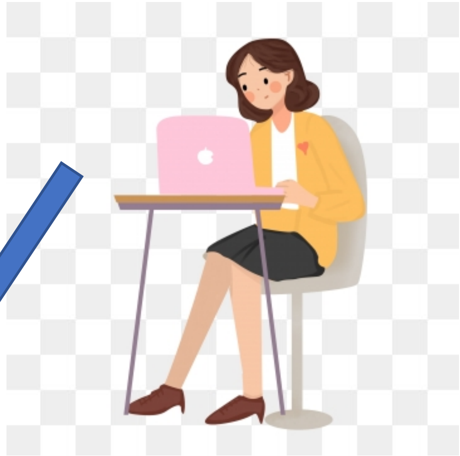
WISDOM'19@KDD'19, Anchorage, Alaska

Date: August 5, 2019

PURDUE UNIVERSITY

Motivation

Balanced diet



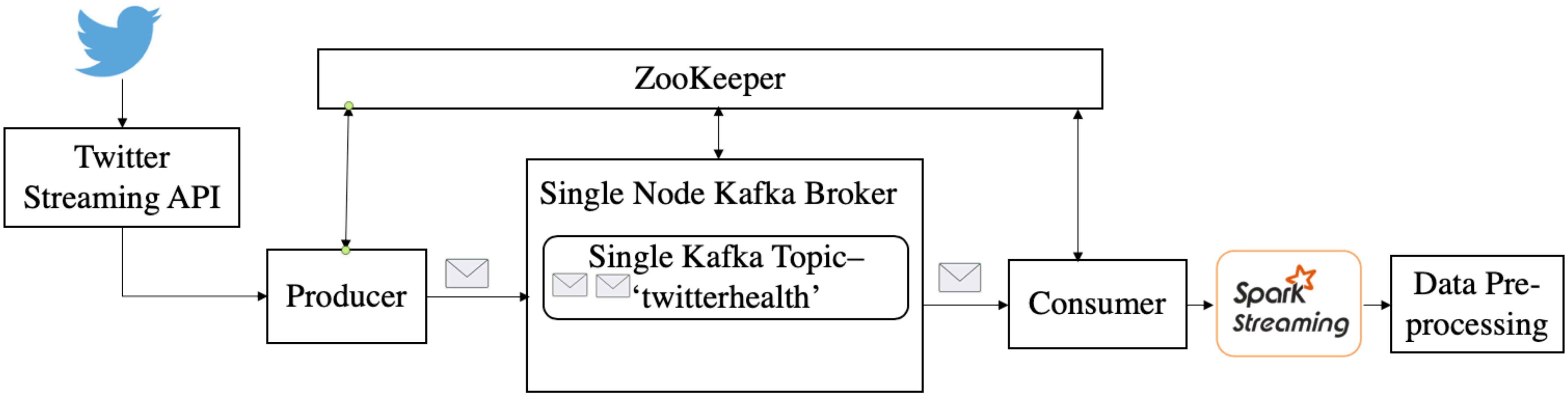
@vuthihoangquye1: RT @go1click: Ketogenic Diet The truth:> buff.ly/2NQr4jY #health #fitness #diet #healthy #fitness #weightloss #exercise #workout #sport #paleo #yoga #food #nutrition #fat #cbd #keto #wellness #news #ff #inspiration

Exercise

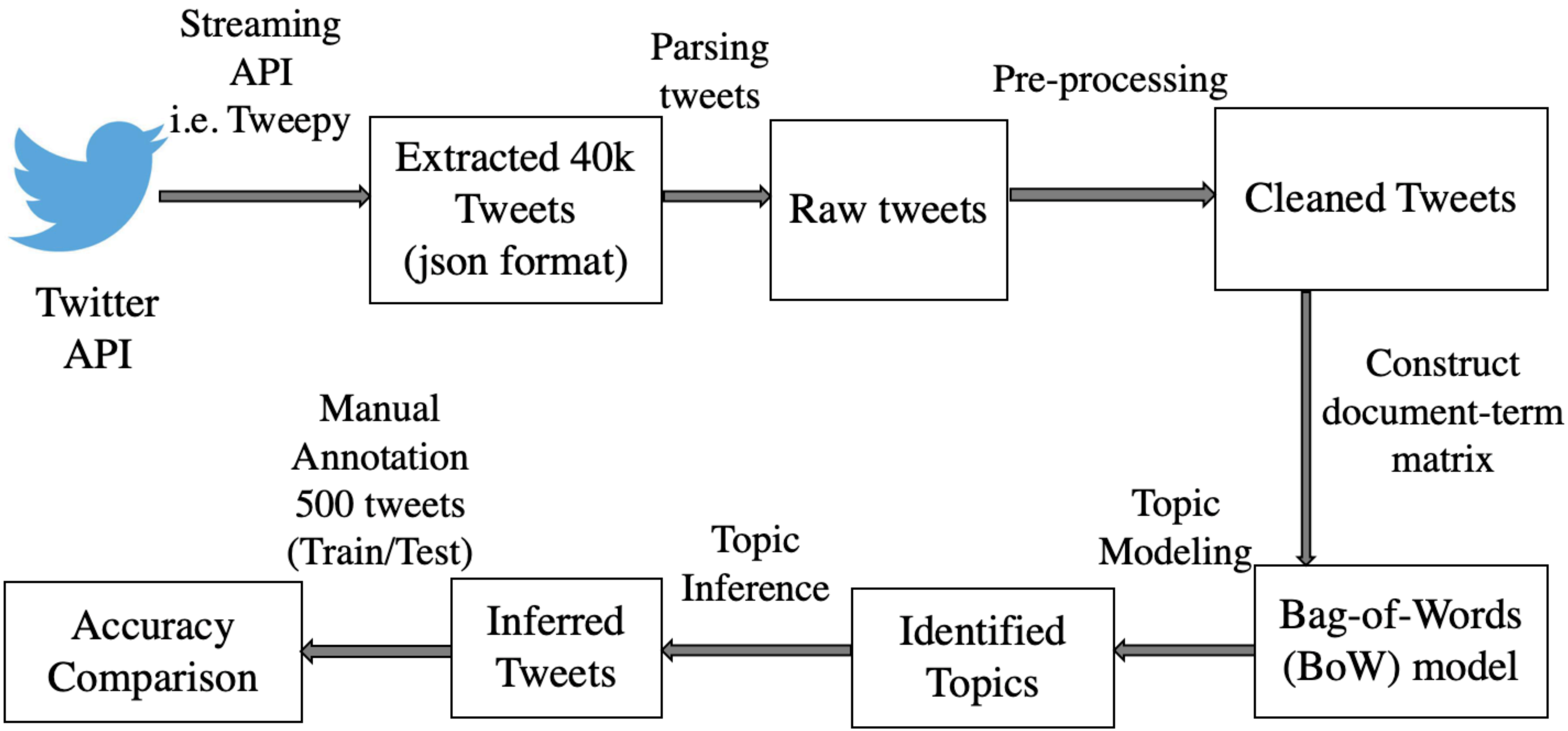
Running

Yoga

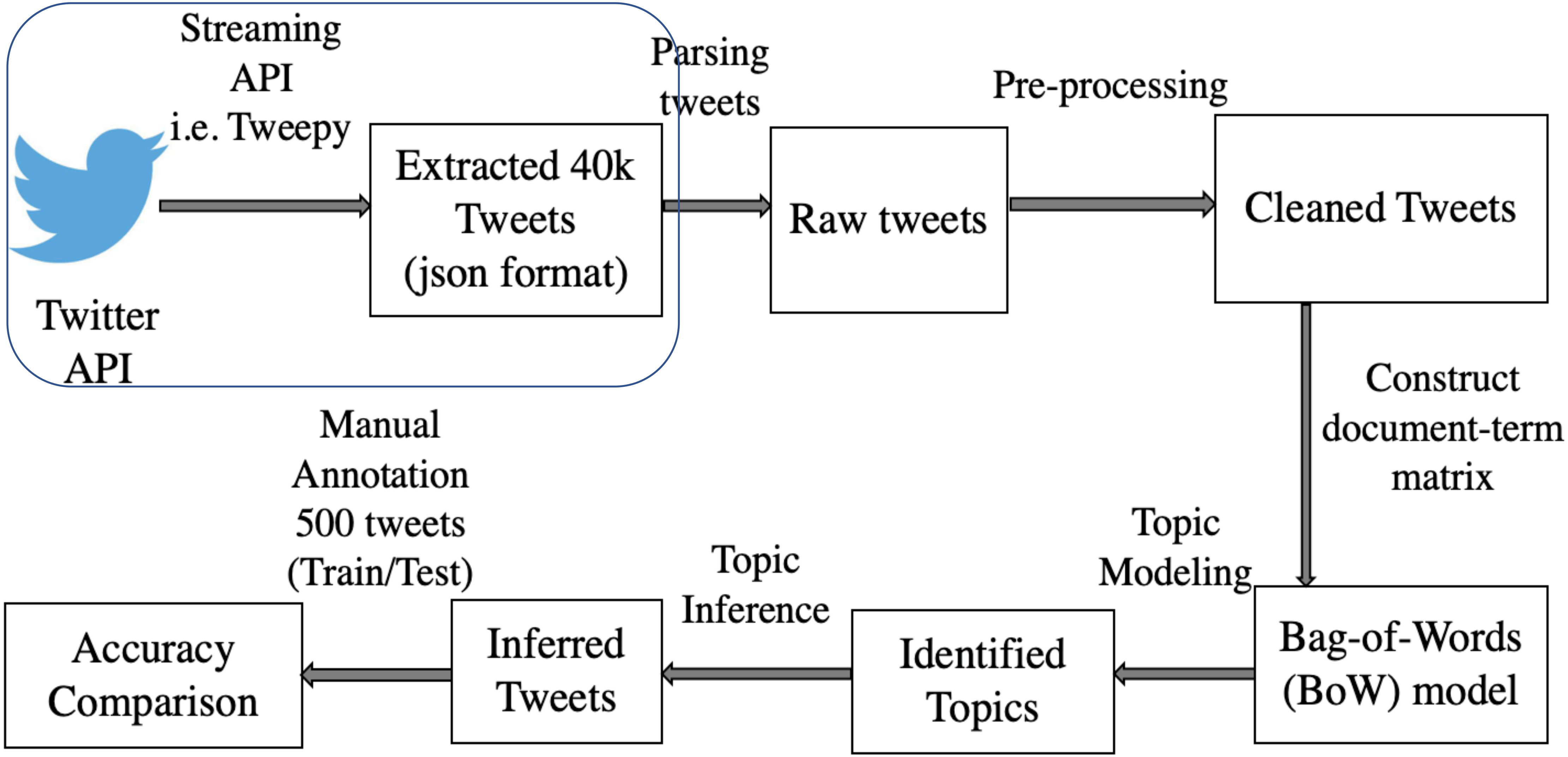
Twitter Data Collection



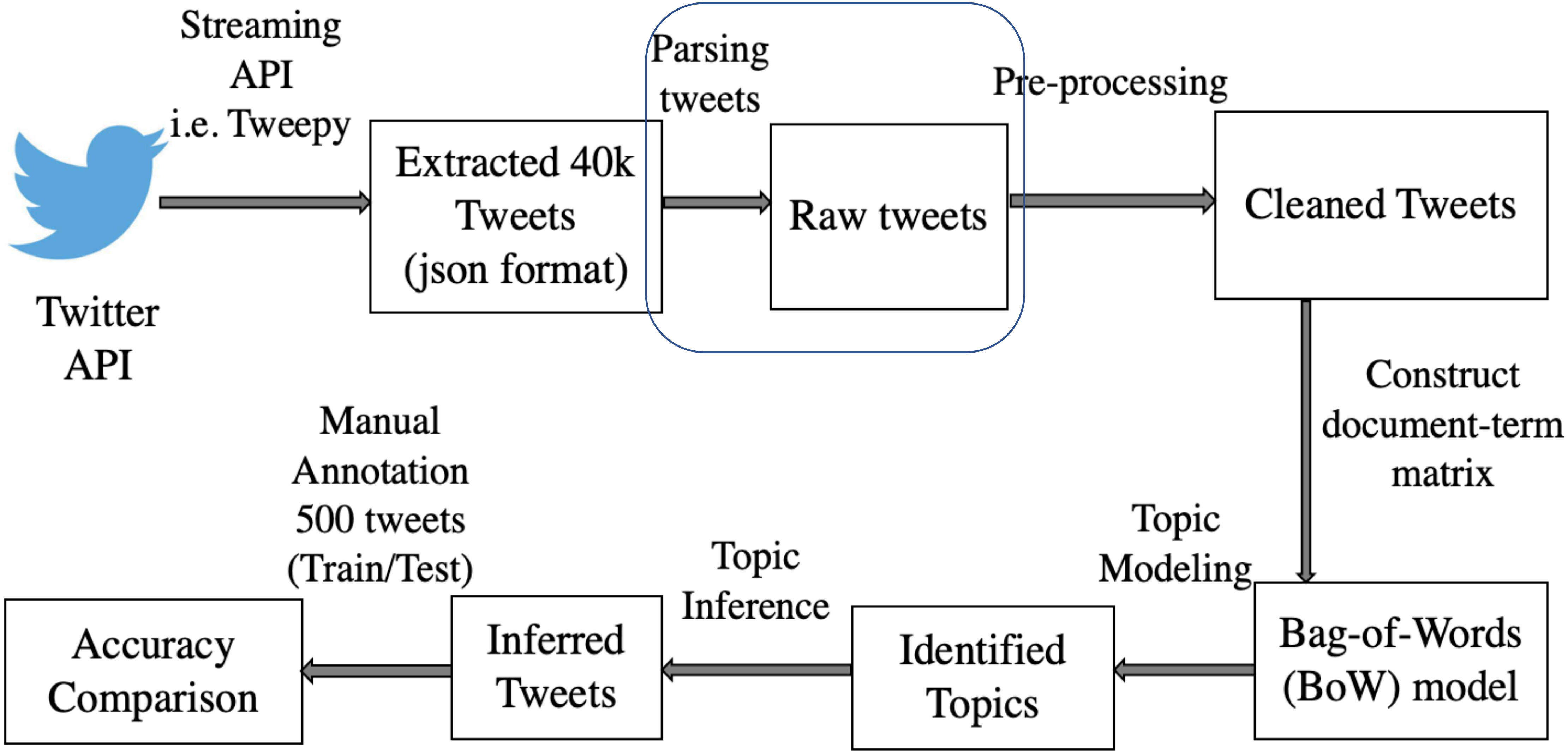
Methodology of Correlation Mining



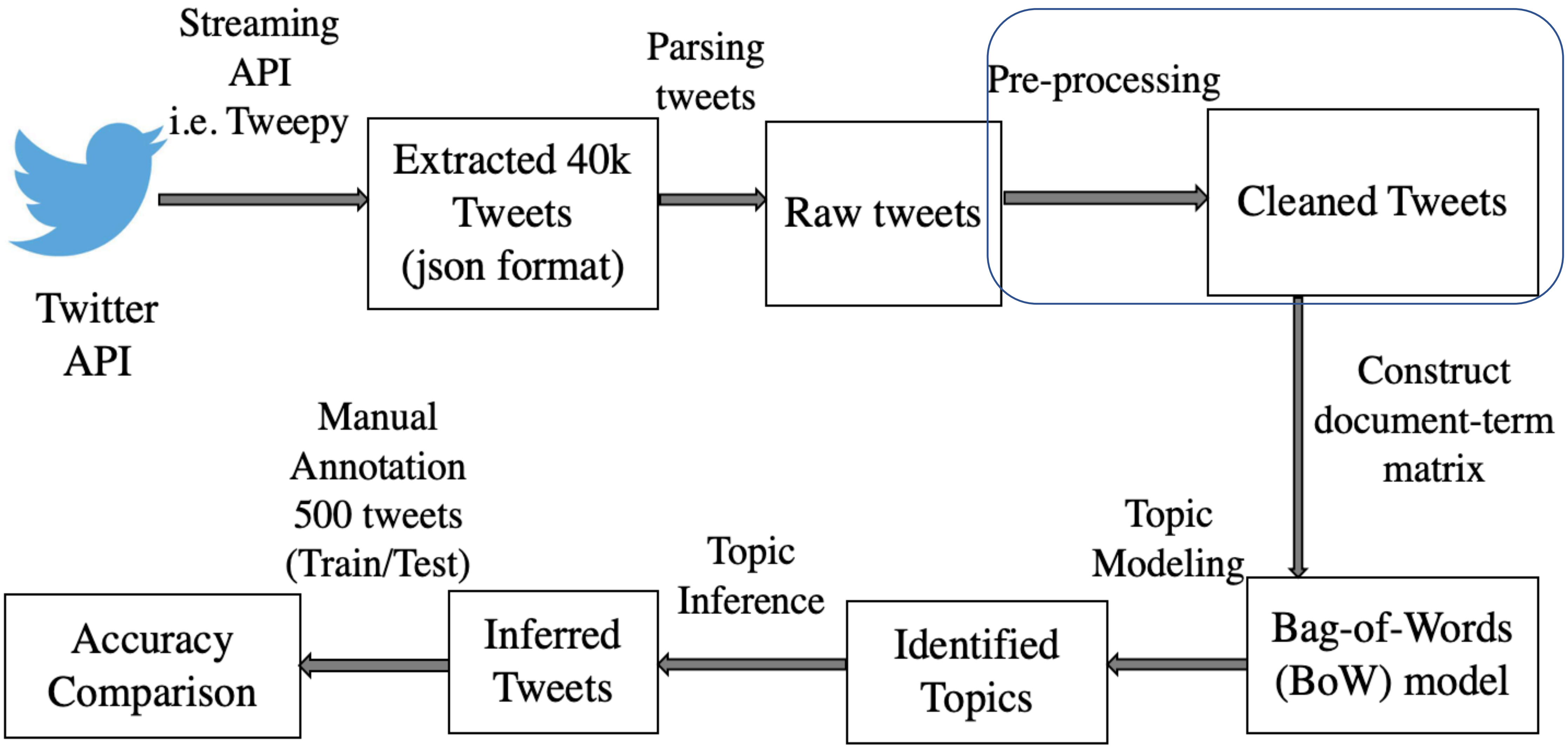
Methodology of Correlation Mining



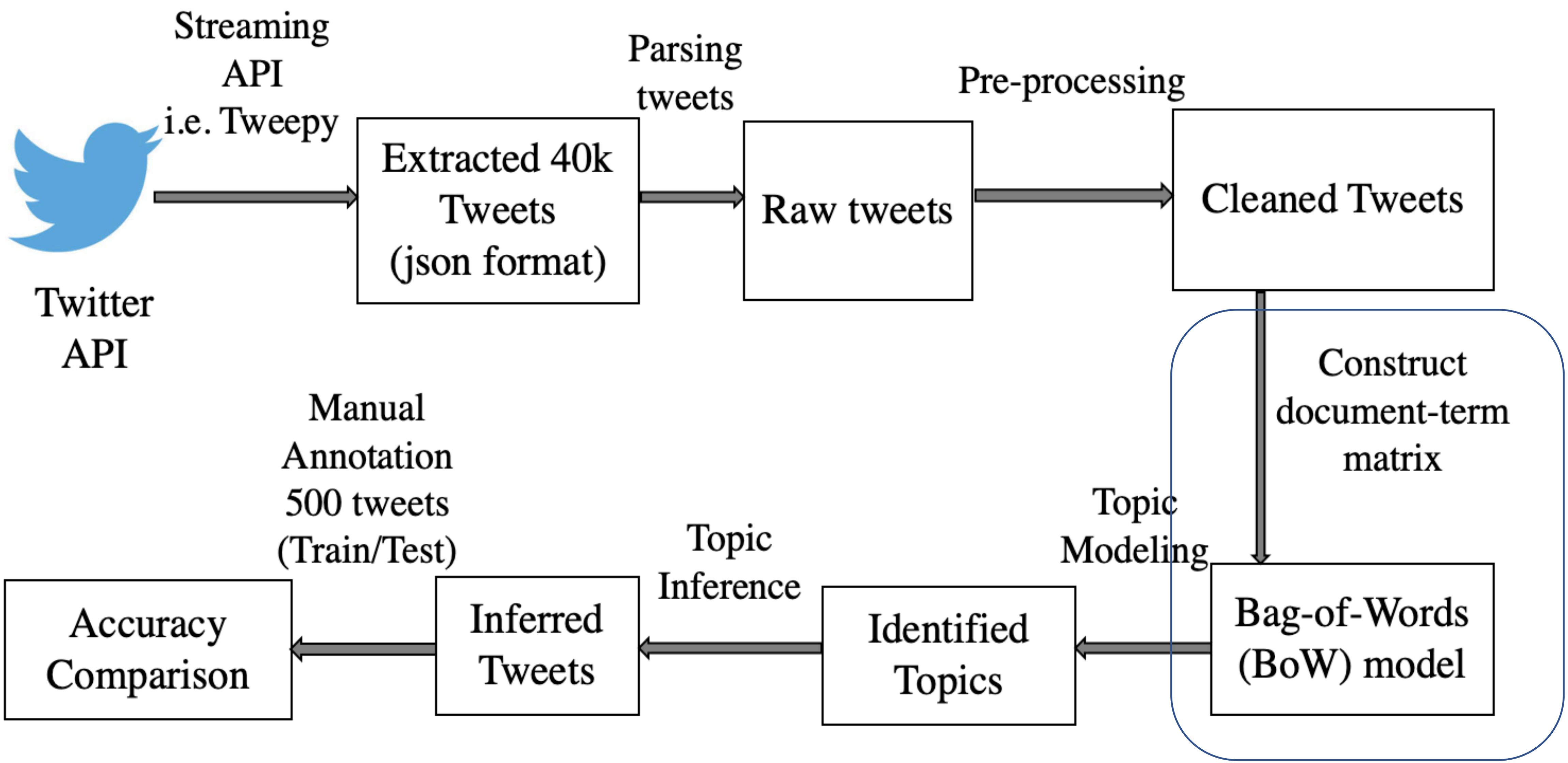
Methodology of Correlation Mining



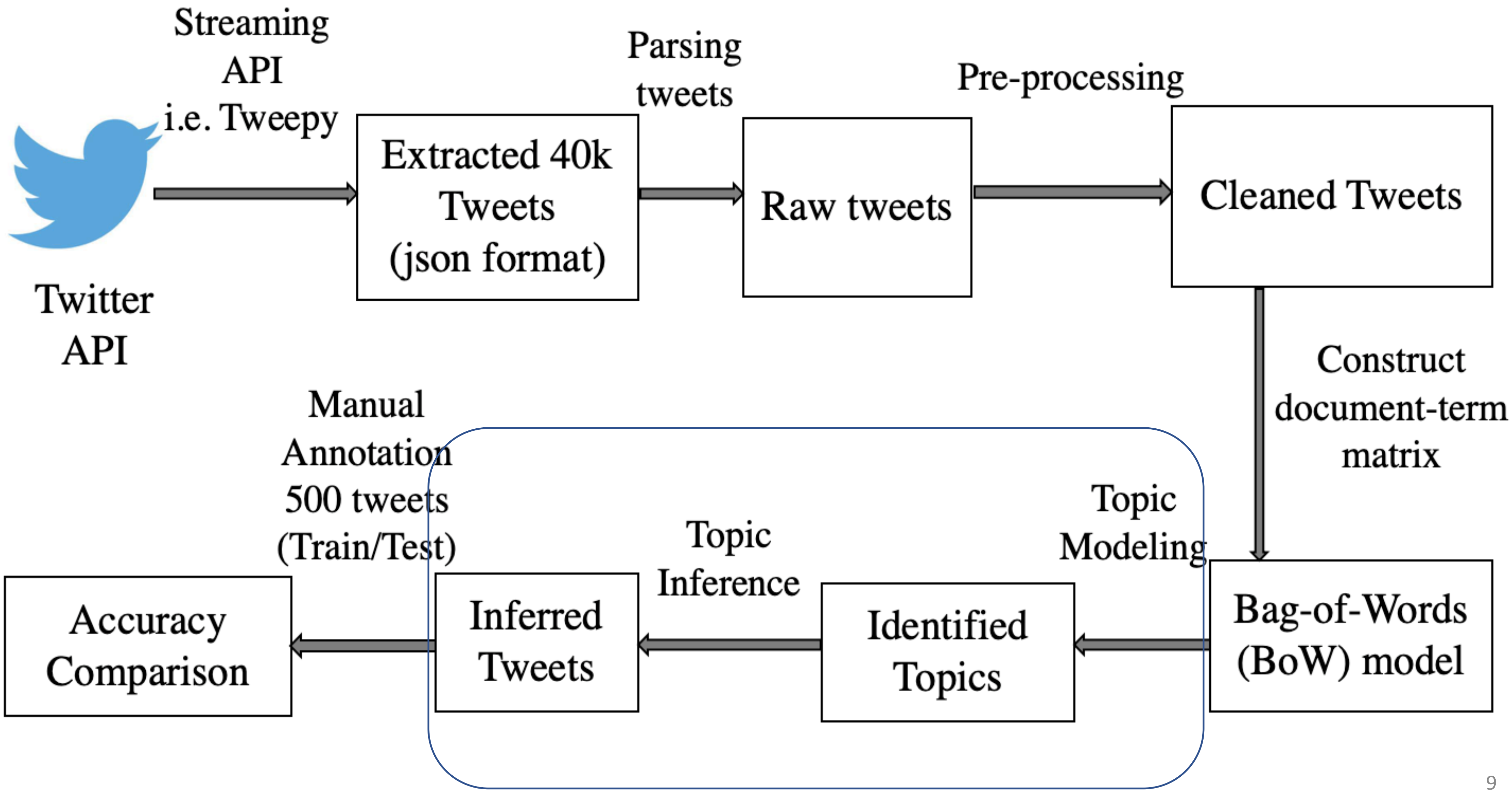
Methodology of Correlation Mining



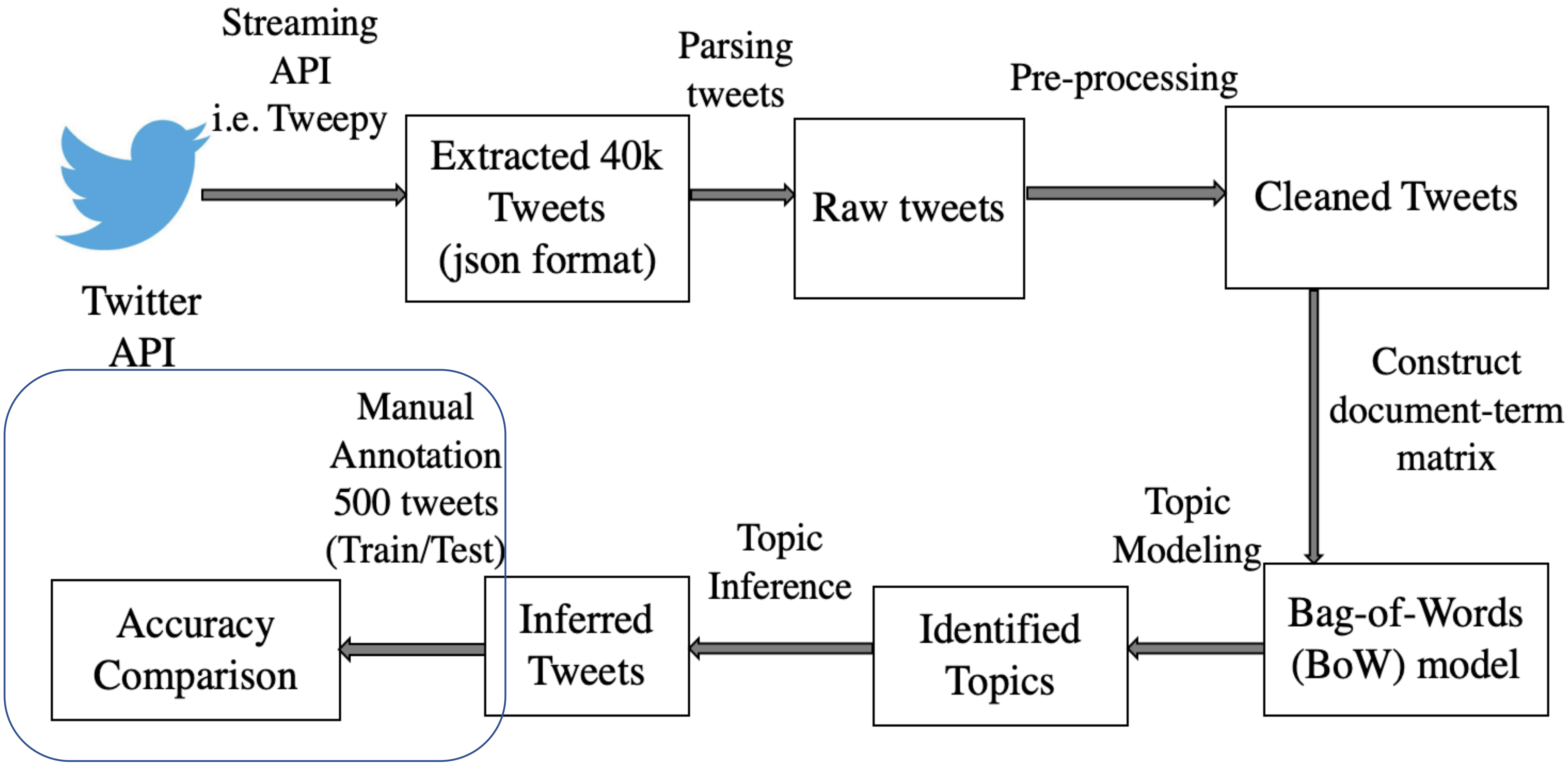
Methodology of Correlation Mining



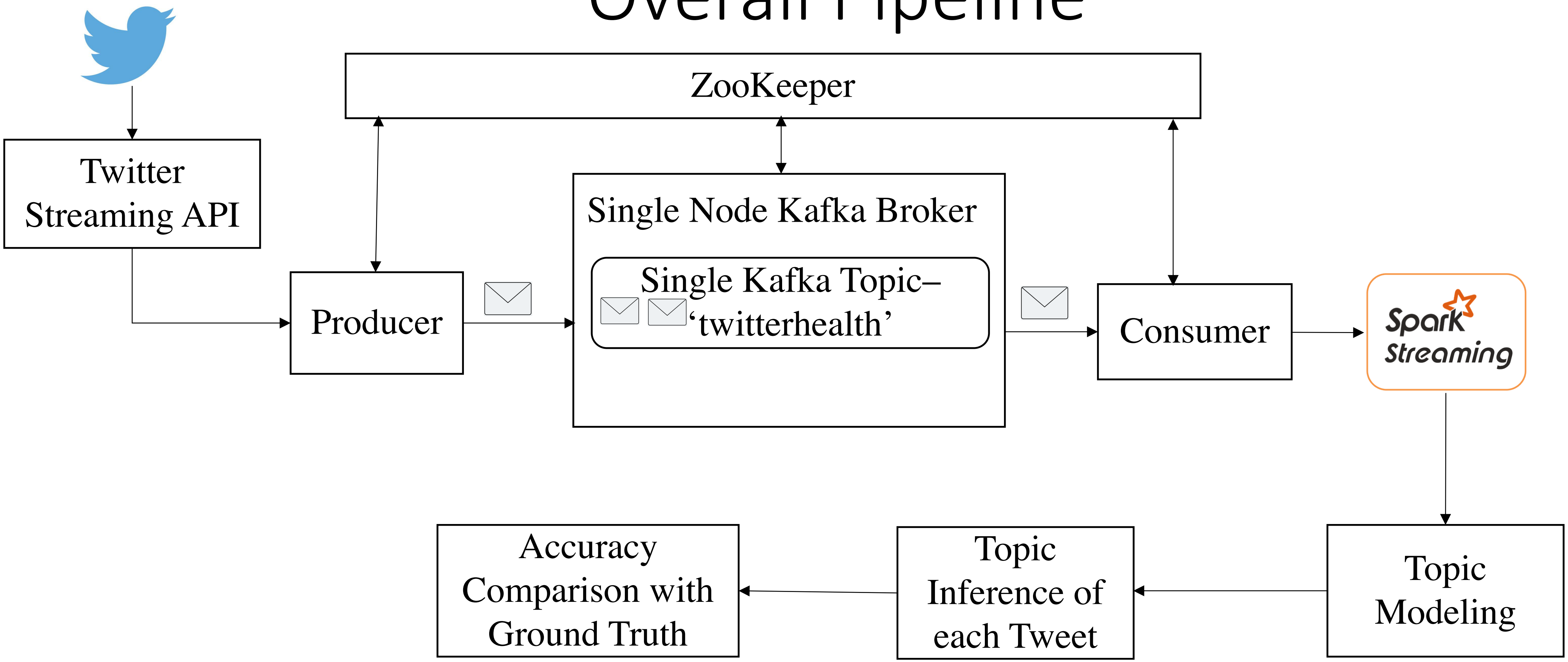
Methodology of Correlation Mining



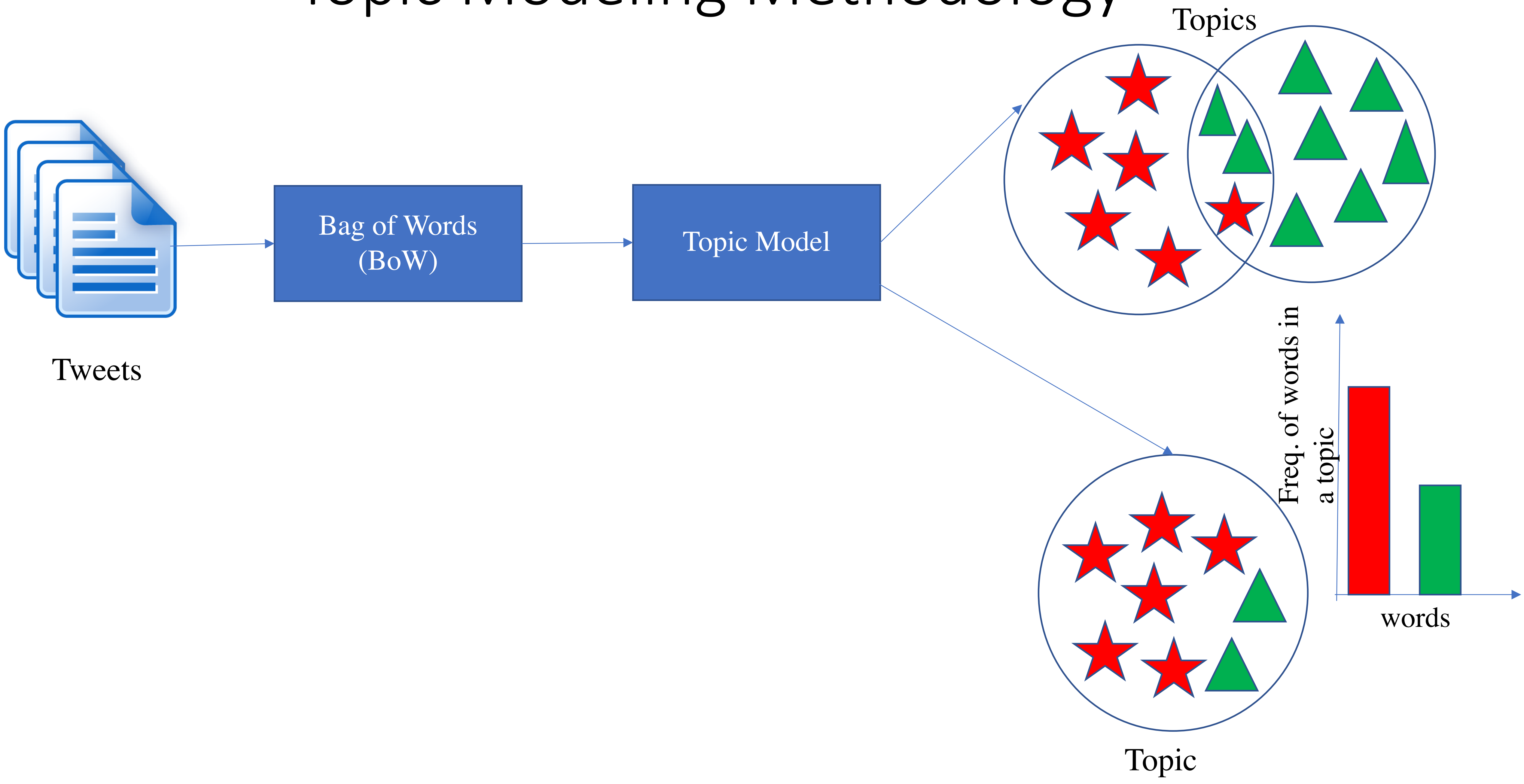
Methodology of Correlation Mining



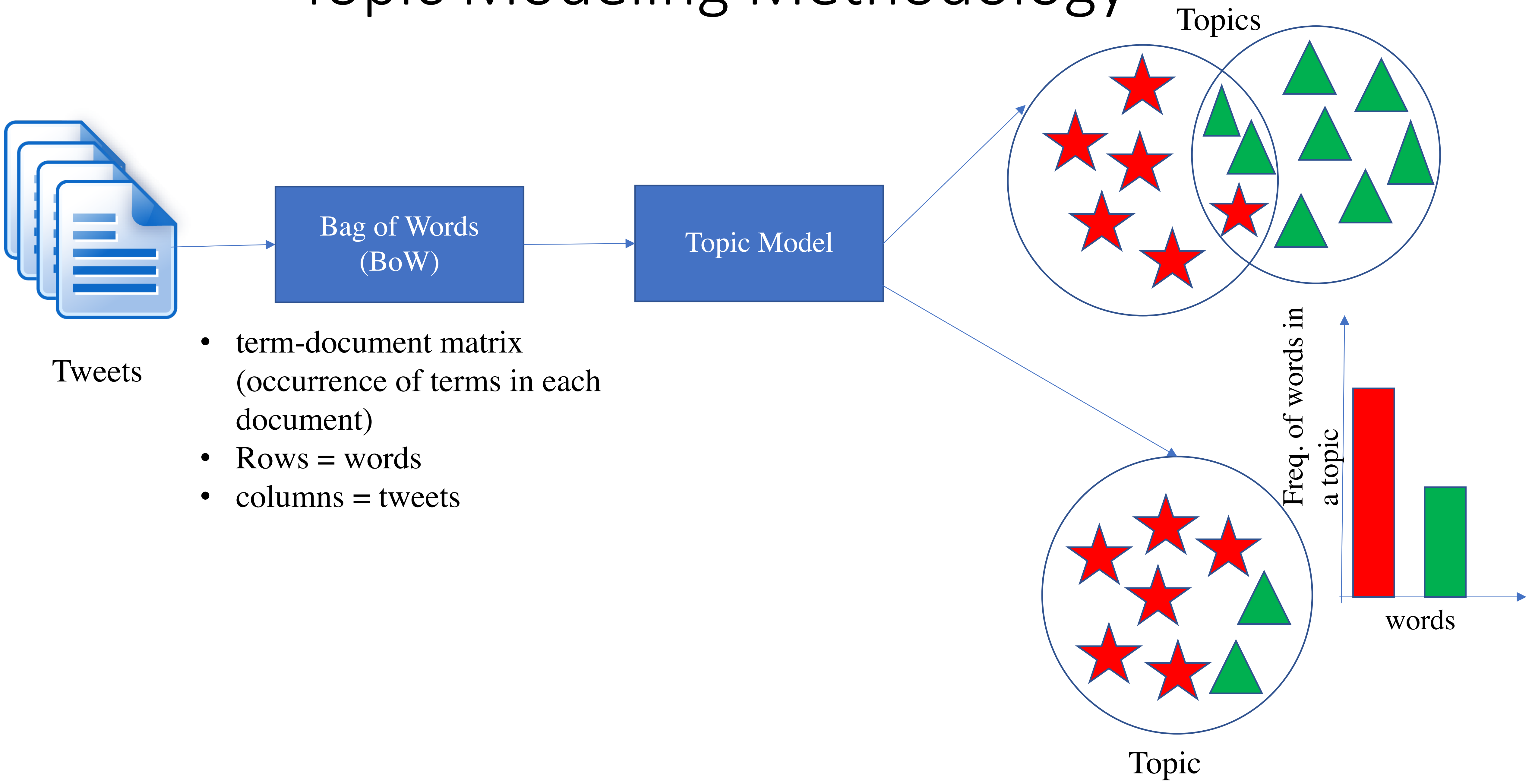
Overall Pipeline



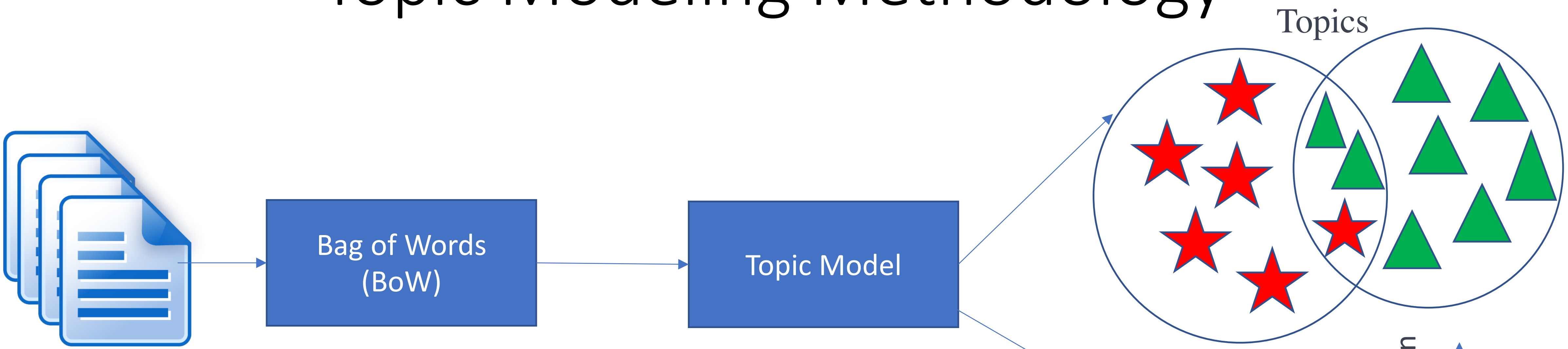
Topic Modeling Methodology



Topic Modeling Methodology



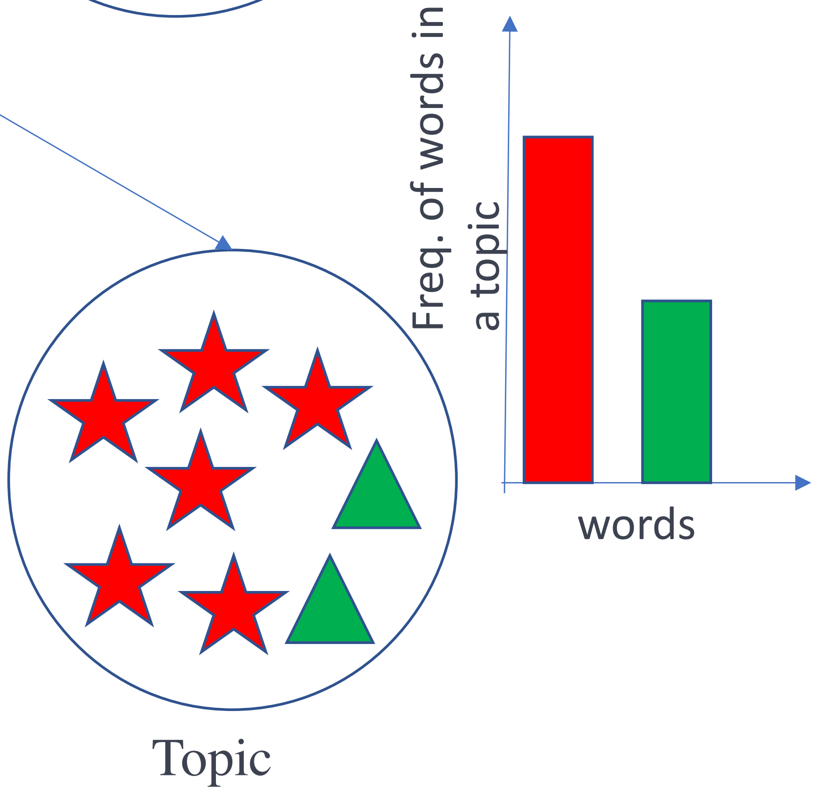
Topic Modeling Methodology



Tweets

- term-document matrix (occurrence of terms in each document)
- Rows = words
- columns = tweets

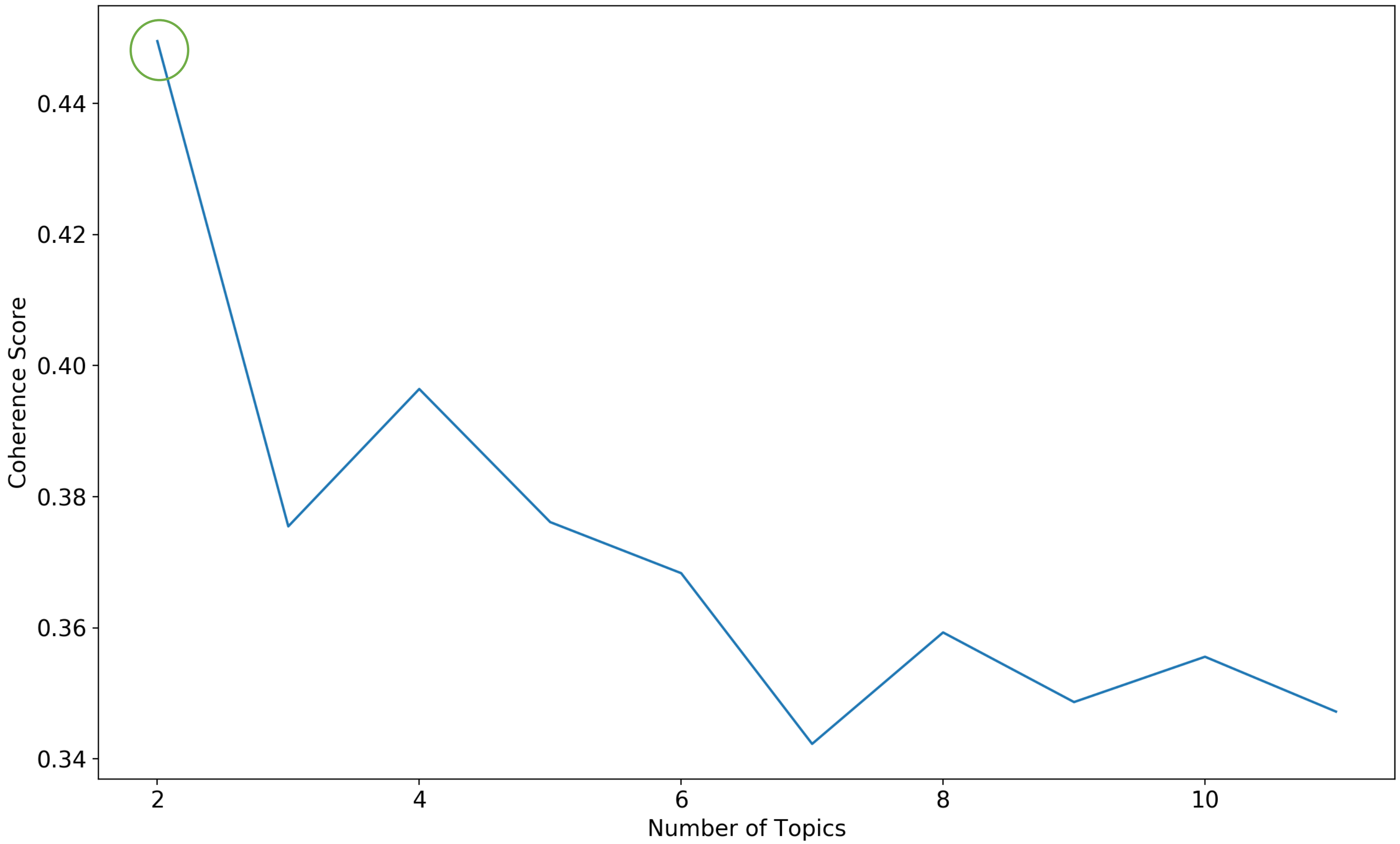
- Latent Semantic Analysis
  - Singular value decomposition
- Non-negative Matrix Factorization
  - Matrices are non-negative
  - Normalization with TF-IDF to give more weight to the "more" important terms
- Latent Dirichlet Allocation
  - Dirichlet distribution



How to choose optimal Number of Topics?

- Build many LSA, LDA, NMF models with different values of number of topics (k).
- pick k with highest coherence value.

Optimal Number of Topics vs Coherence Score LSA



K = 2
Coherence Value = 0.4495

Topics using LSA

Topic1

yoga
everi
life
job
remember
goe
woman
everyone
cook
therapy

Topic2

diet
vegan
fit
day
new
like
beyonce
amp
eat
workout

Topics using LSA

Topic1

yoga
everi
life
job
remember
goe
woman
everyone
cook
therapy

Topic2

diet
vegan
fit
day
new
like
beyonce
amp
eat
workout

- *highly dense matrix*.

Topics using LSA

Topic1

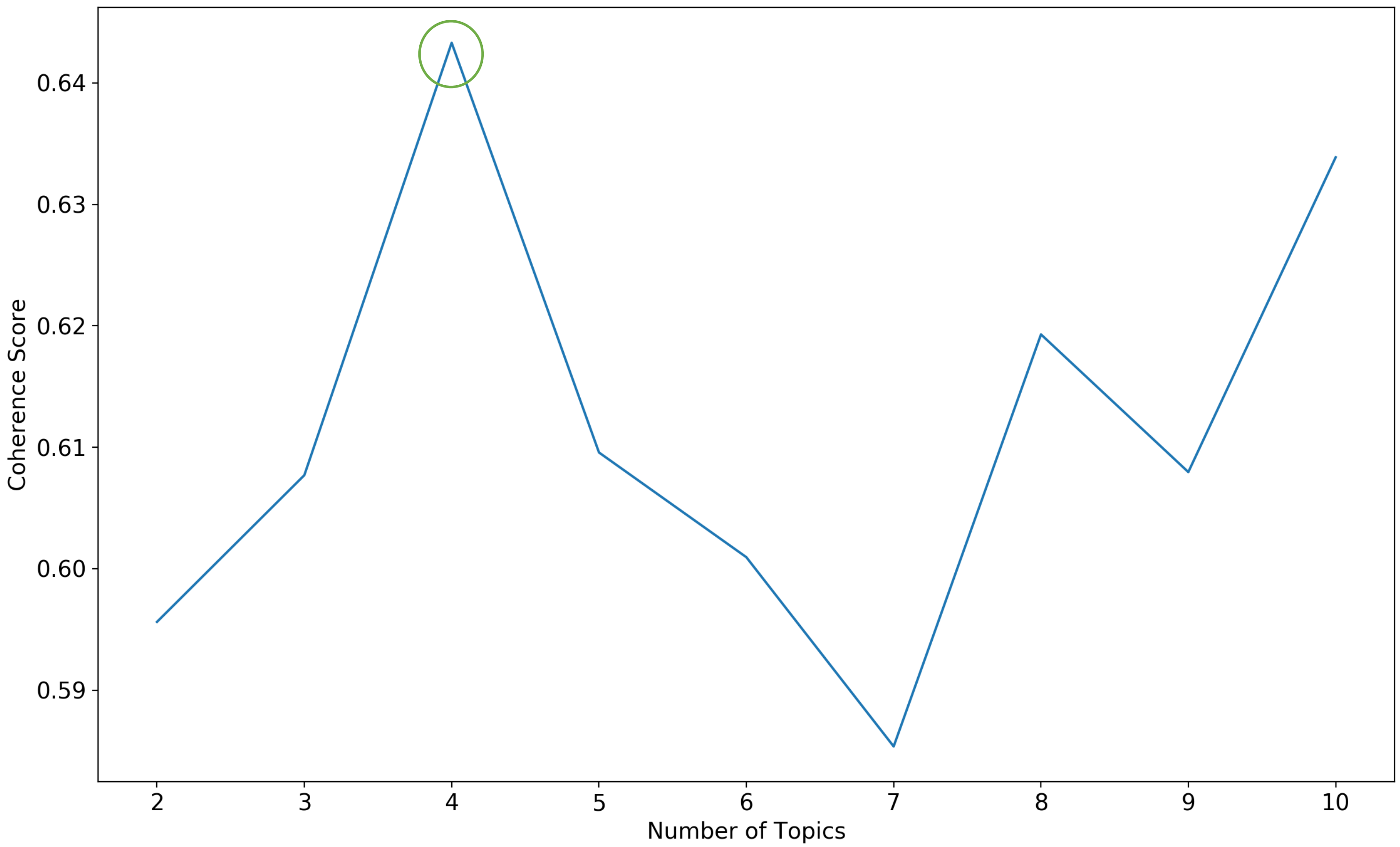
yoga
everi
life
job
remember
goe
woman
everyone
cook
therapy

Topic2

diet
vegan
fit
day
new
like
beyonce
amp
eat
workout

- *highly dense matrix*
- *unable to capture the meanings of words.*
- *lower accuracy*

Optimal Number of Topics vs Coherence Score NMF



K = 4
Coherence Value = 0.6404

Topic coherence measure TC-W2V

Topics using NMF

Topic1	Topic2	Topic3	Topic4
Yoga	diet	vegan	fitness
job	beyonce	go	workout
every_woman	new	eat	go
cooks_goe	bitch	make	good
therapy_remember	ciara_prayer	food	amp
life_juggl	day	day	day
everyone_birthday	eat	amp	yoga
boyfriend	go	shit	health
hot	fat	meat	gym
know	keto	vegetarian	today

Topics using NMF

- *sparse representations*

Topic1	Topic2	Topic3	Topic4
Yoga	diet	vegan	fitness
job	beyonce	go	workout
every_woman	new	eat	go
cooks_goe	bitch	make	good
therapy_remember	ciara_prayer	food	amp
life_juggl	day	day	day
everyone_birthday	eat	amp	yoga
boyfriend	go	shit	health
hot	fat	meat	gym
know	keto	vegetarian	today

Topics using NMF

- *sparse representations*
- *same keywords are repeated in multiple topics.*

Topic1	**Topic2**	**Topic3**	**Topic4**
Yoga	diet	vegan	fitness
job	beyonce	go	workout
every_woman	new	eat	go
cooks_goe	bitch	make	good
therapy_remember	ciara_prayer	food	amp
life_juggl	day	day	day
everyone_birthday	eat	amp	yoga
boyfriend	go	shit	health
hot	fat	meat	gym
know	keto	vegetarian	today

Topics using NMF

- *sparse representations*
- *same keywords are repeated in multiple topics.*

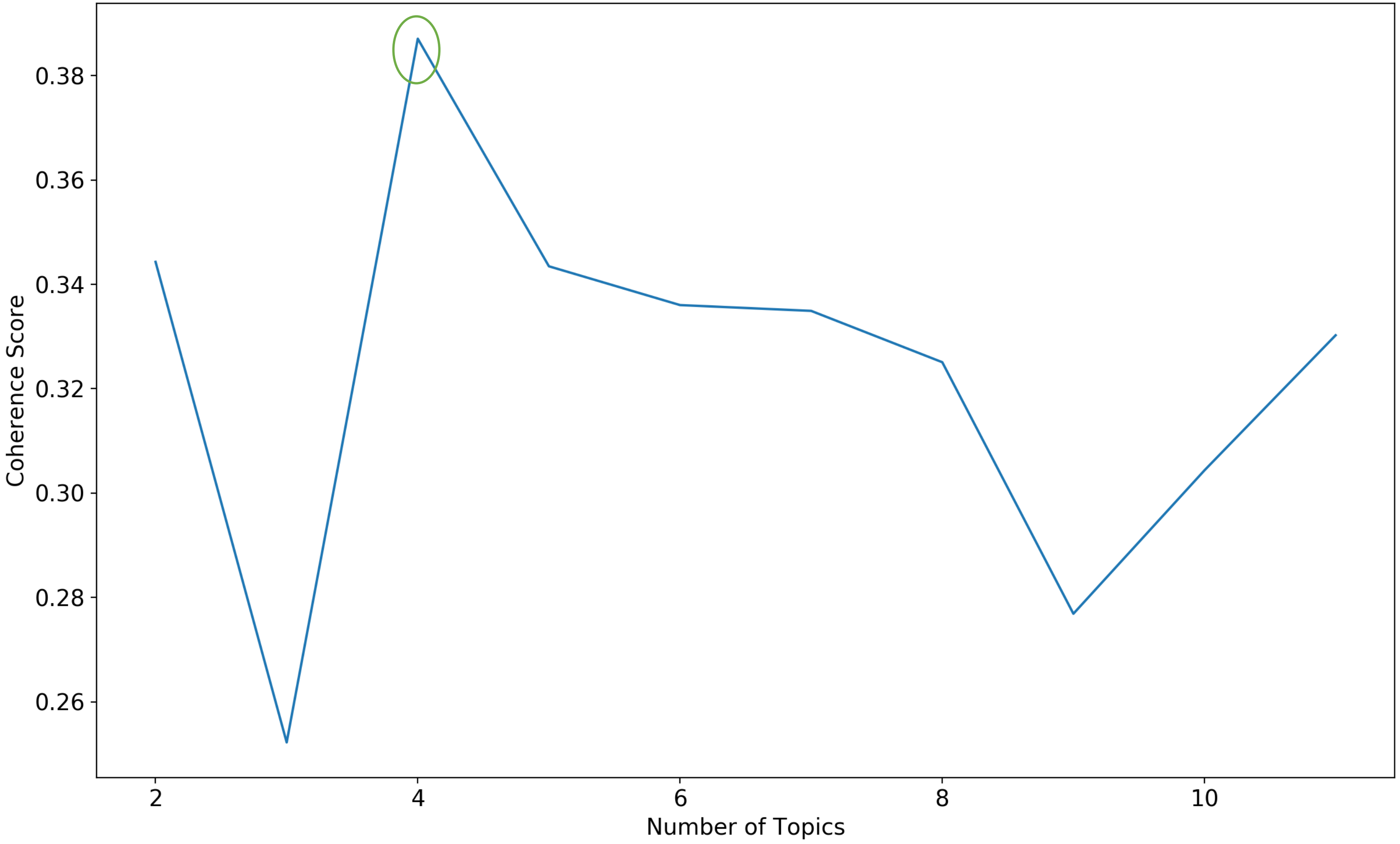
Topic1	Topic2	Topic3	Topic4
Yoga	diet	vegan	fitness
job	beyonce	go	workout
every_woman	new	eat	go
cooks_goe	bitch	make	good
therapy_remember	ciara_prayer	food	amp
life_juggl	day	day	day
everyone_birthday	eat	amp	yoga
boyfriend	go	shit	health
hot	fat	meat	gym
know	keto	vegetarian	today

Topics using NMF

- *sparse representations*
- *same keywords are repeated in multiple topics.*

Topic1	Topic2	Topic3	Topic4
Yoga	diet	vegan	fitness
job	beyonce	go	workout
every_woman	new	eat	go
cooks_goe	bitch	make	good
therapy_remember	ciara_prayer	food	amp
life_juggl	day	day	day
everyone_birthday	eat	amp	yoga
boyfriend	go	shit	health
hot	fat	meat	gym
know	keto	vegetarian	today

Optimal Number of Topics vs Coherence Score LDA



K = 4
Coherence Value = 0.3871

Topics using LDA

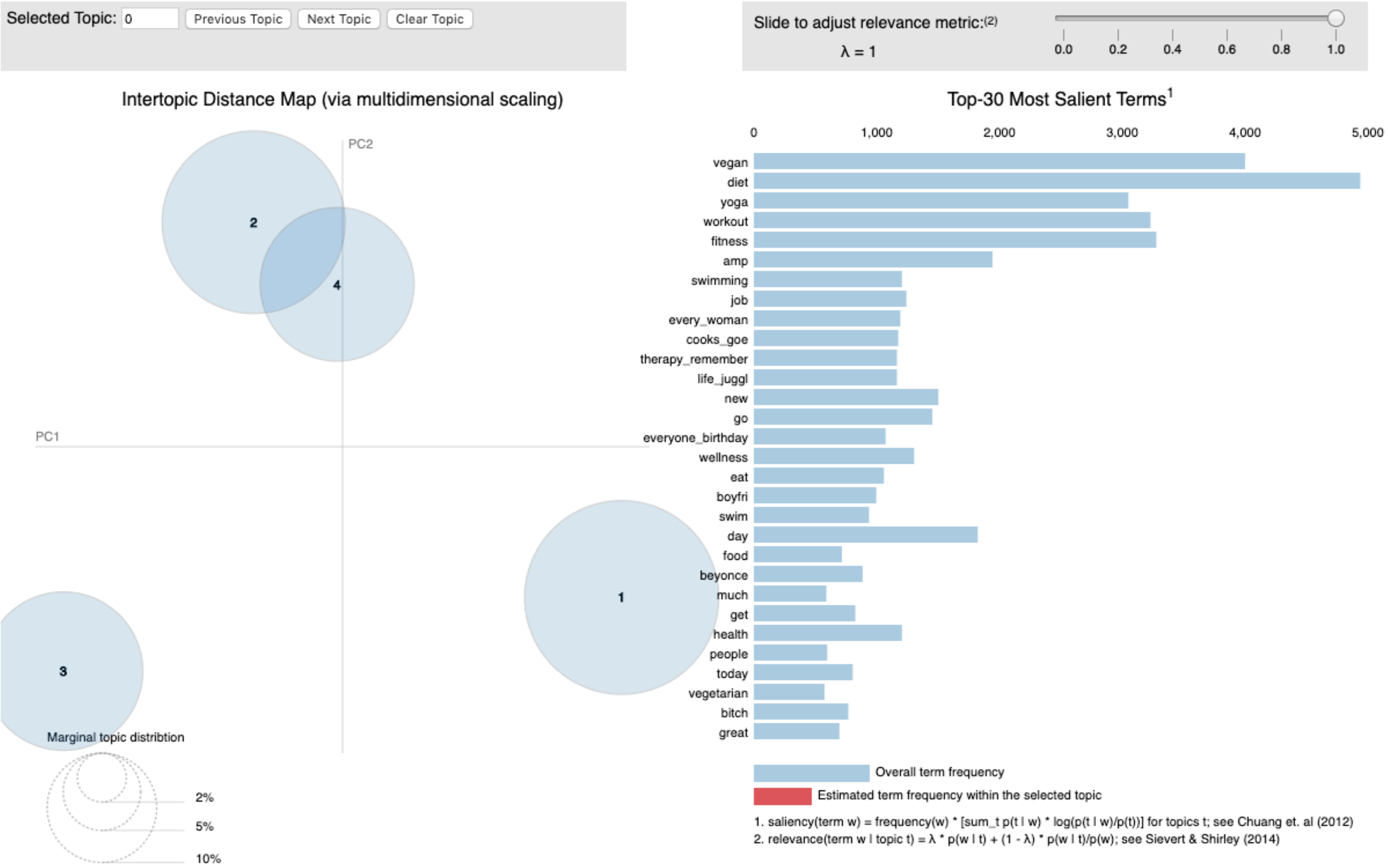
Topic1	Topic2	Topic3	Topic4
diet	vegan	swimming	fitness
workout	yoga	swim	amp
new	job	day	wellness
go	every_woman	much	health
day	cooks_goe	support	time
beyonce	therapy_remember	really	great
get	life_juggle	try	look
today	everyone_birthday	always	hiking
bitch	eat	relationship	make
gym	boyfriend	pool	love

Topics using LDA

- *coherent topics*

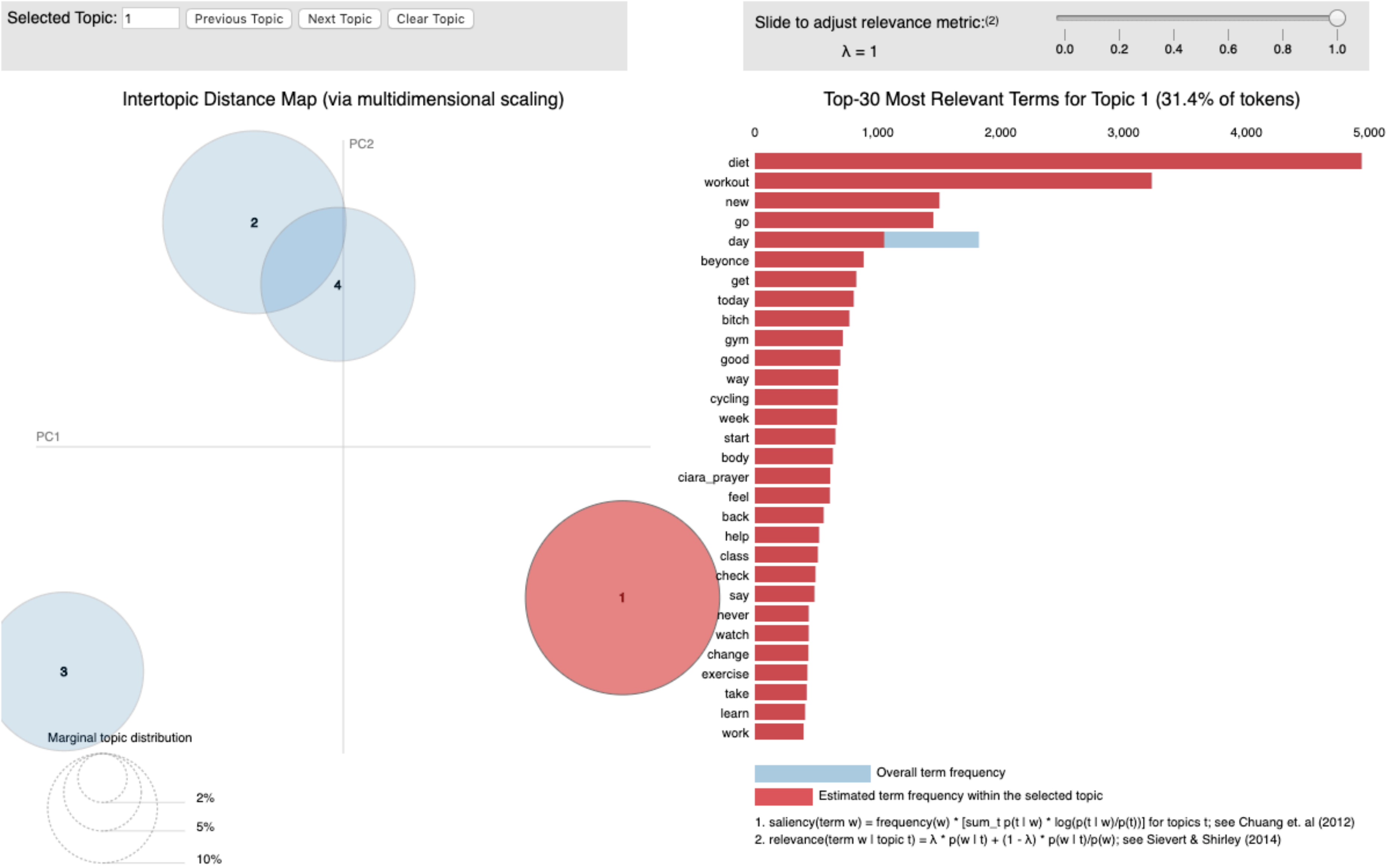
Topic1	Topic2	Topic3	Topic4
diet	vegan	swimming	fitness
workout	yoga	swim	amp
new	job	day	wellness
go	every_woman	much	health
day	cooks_goe	support	time
beyonce	therapy_remember	really	great
get	life_juggle	try	look
today	everyone_birthday	always	hiking
bitch	eat	relationship	make
gym	boyfriend	pool	love

Visualization of Topics- pyLDAVIS



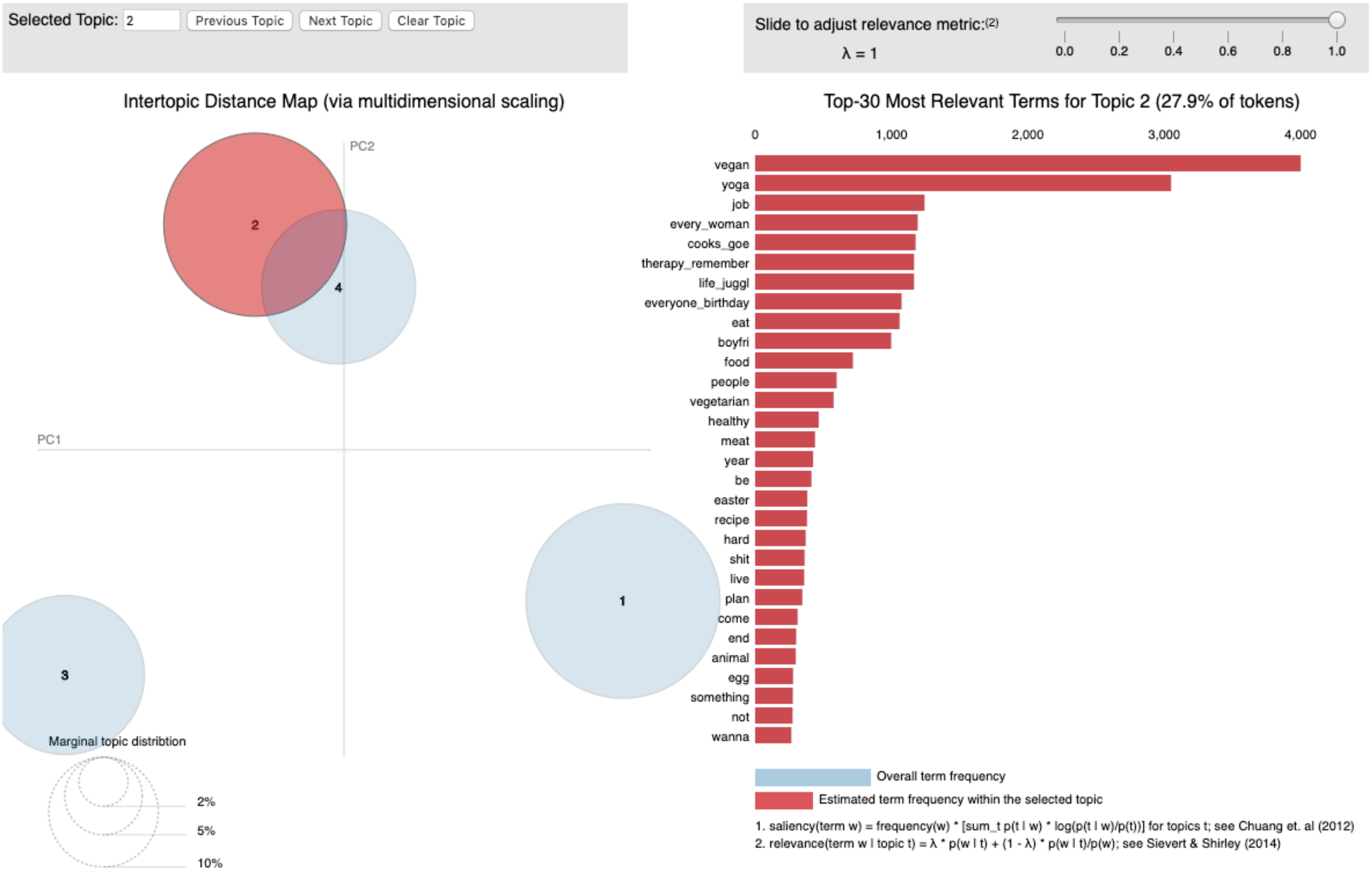
Online link: https://tunazislam.github.io/files/LDA_Visualization_t4.html

Visualization of Topics- pyLDAVIS



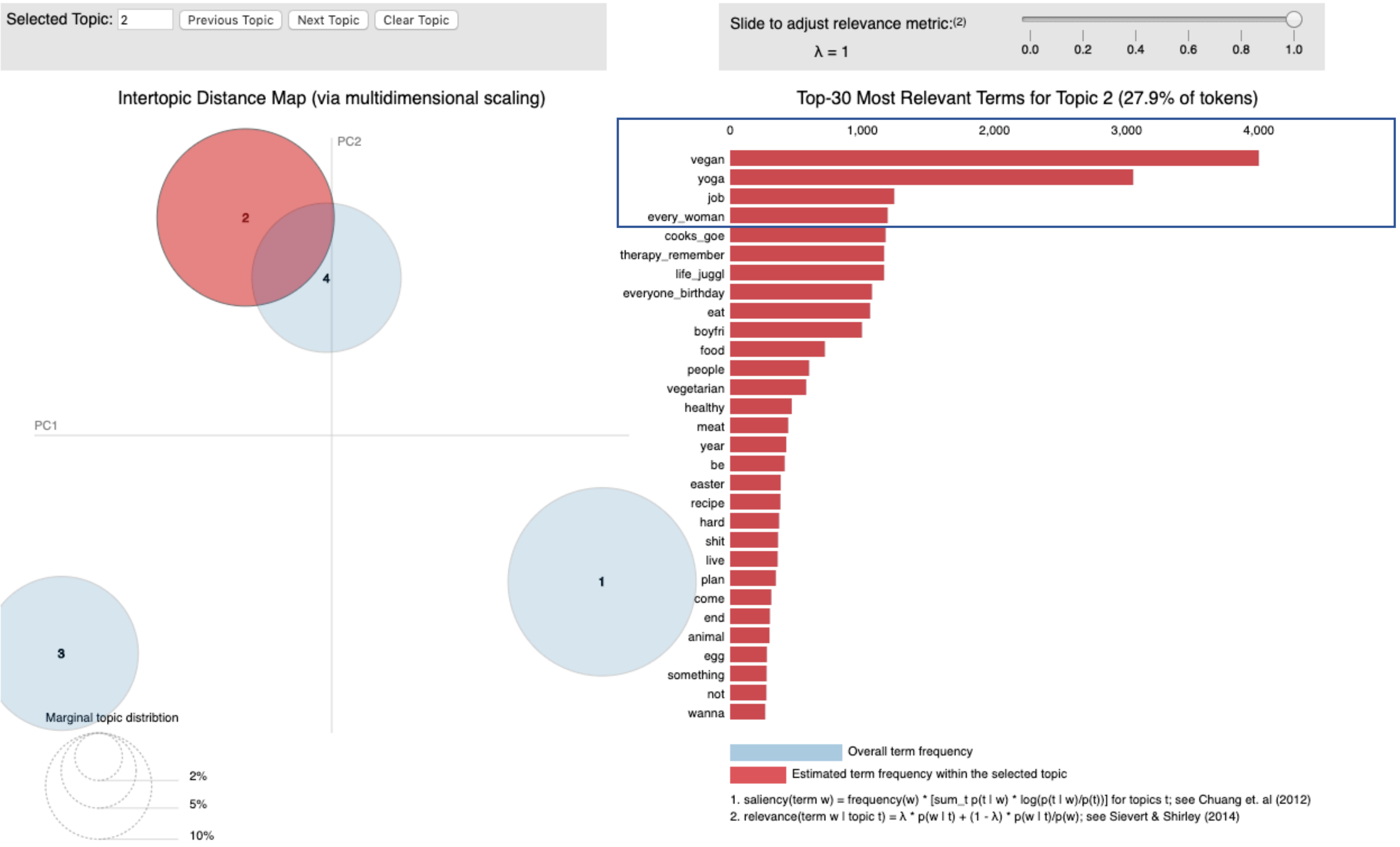
Online link: https://tunazislam.github.io/files/LDA_Visualization_t4.html

Visualization of Topics- pyLDAVIS



Online link: https://tunazislam.github.io/files/LDA_Visualization_t4.html

Visualization of Topics- pyLDAVIS



Top-4 co-occurring keywords

vegan

yoga

job

every_woman

Online link: https://tunazislam.github.io/files/LDA_Visualization_t4.html

Topic Inference (Train data)

- Observing dominant topic, 2nd dominant topic and its percentage of contribution in each Tweet.

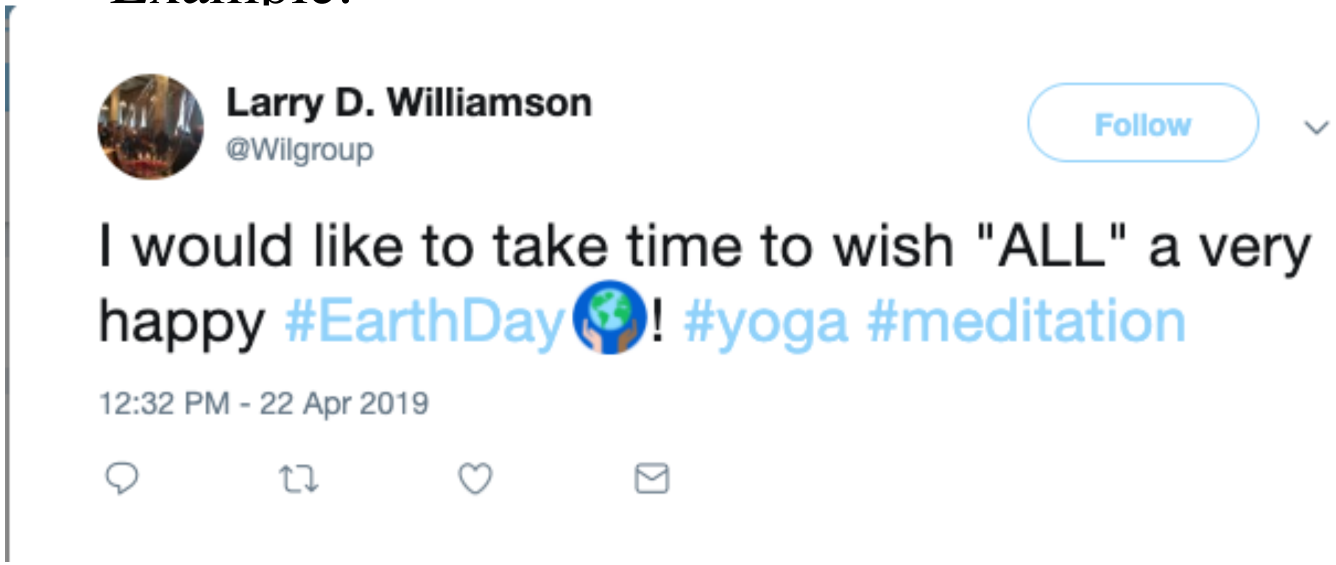
Example:

Dominant Topic	**2nd Dominant Topic**
Topic 2	Topic 1
vegan	*diet*
yoga	*workout*
job	*new*
every_woman	*go*
cooks_goe	*day*
therapy_remember	*beyonce*
life_juggle	*get*
everyone_birthday	*today*
eat	*bitch*
boyfriend	*gym*
61%	**18%**

Topic Inference on **New** Tweets (Test data)

- Observing dominant topic, 2nd dominant and its percentage of contribution to **new** Tweet.

Example:



Dominant Topic	**2nd Dominant Topic**
Topic 2	Topic 4
vegan	*fitness*
yoga	*amp*
job	*wellness*
every_woman	*health*
cooks_goe	*time*
therapy_remember	*great*
life_juggle	*look*
everyone_birthday	*hiking*
eat	*make*
boyfriend	*love*
33%	**32%**

Manual Annotation (Train/Test data)

- **100, 200, 300, 400, and 500** tweets from train data
- **New** tweets for test data
- Calculate accuracy with ground truth

Manual Annotation

- Intent of tweets.
- For example:
  - **Tweet 1:** *Learning some traditional yoga with my good friend*.

  - **Tweet 2:** *Why You Should #LiftWeights to Lose #BellyFat #Fitness #core #abs #diet #gym #bodybuilding #workout #yoga*

Manual Annotation

- Intent of tweets.
- For example:
  - **Tweet 1:** *Learning some traditional yoga with my good friend*.

    Yoga activity

  - **Tweet 2:** *Why You Should #LiftWeights to Lose #BellyFat #Fitness #core #abs #diet #gym #bodybuilding #workout #yoga*

    Workout, Diet

Manual Annotation

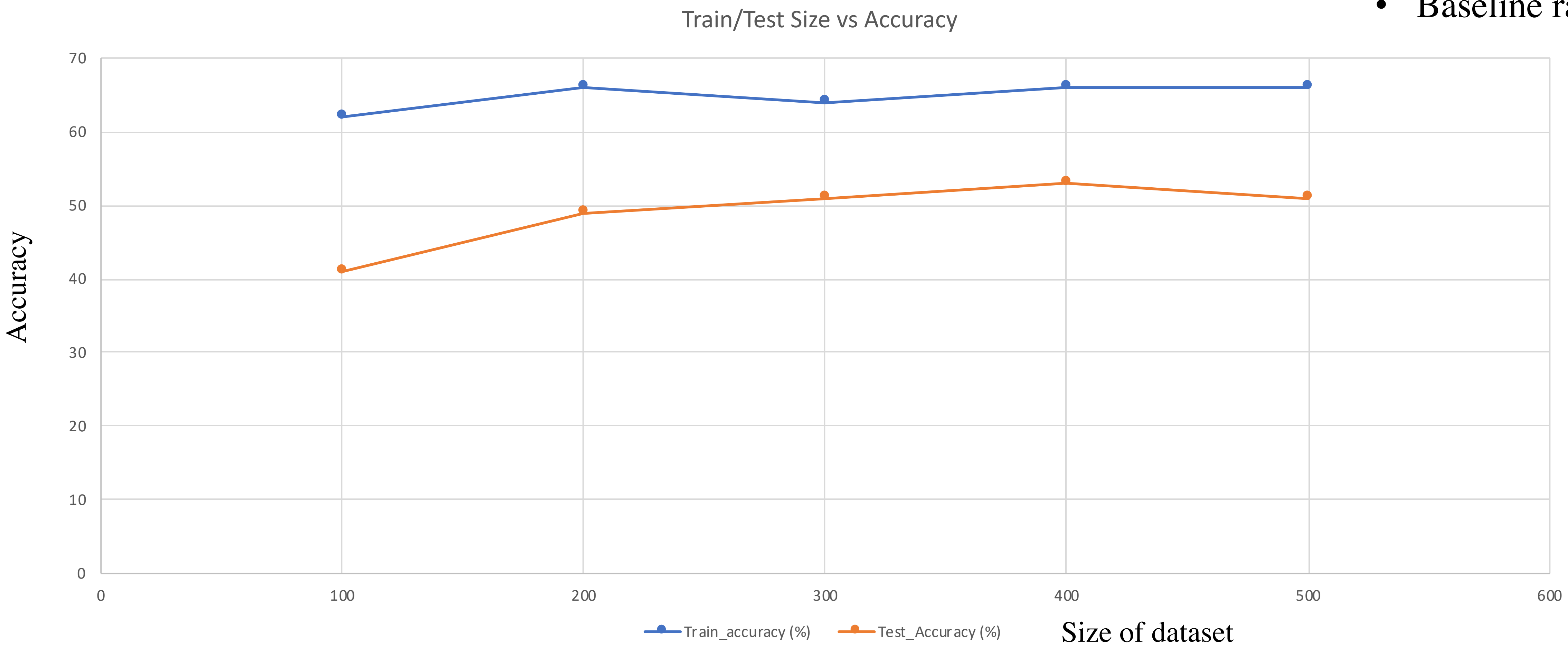
- Intent of tweets.
- For example:
  - **Tweet 1:** *Learning some traditional yoga with my good friend*.

    Topic 2
  - **Tweet 2:** *Why You Should #LiftWeights to Lose #BellyFat #Fitness #core #abs #diet #gym #bodybuilding #workout #yoga*

    Topic 1

Train/Test Accuracy with Ground Truth

- Train: 66%
- Test: 51%
- Baseline random: 25%



Observation 1

Dominant Topic	**2^{nd} Dominant Topic**
Topic 2	Topic 3
vegan	*swimming*
yoga	*swim*
job	*day*
every_woman	*much*
cooks_goe	*support*
therapy_remember	*really*
life_juggle	*try*
everyone_birthday	*always*
eat	*relationship*
boyfriend	*pool*
43%	**23%**

Observation 1

Miss Kate
@KateHagans
Follow

This morning I packed myself a salad. Went to yoga during lunch. And then ate my salad with water in hand.

I'm feeling so healthy I don't know what to even do with myself. Like maybe I should eat a bag of chips or something...

12:32 PM - 22 Apr 2019

17 Likes

17

Misleading topic

Dominant Topic	**2nd Dominant Topic**
Topic 2	Topic 3
vegan	*swimming*
yoga	*swim*
job	*day*
every_woman	*much*
cooks_goe	*support*
therapy_remember	*really*
life_juggle	*try*
everyone_birthday	*always*
eat	*relationship*
boyfriend	*pool*
43%	**23%**

Observation 1

Diet related topic (Topic 1)

Dominant Topic	**2nd Dominant Topic**
Topic 2	Topic 3
vegan	*swimming*
yoga	*swim*
job	*day*
every_woman	*much*
cooks_goe	*support*
therapy_remember	*really*
life_juggle	*try*
everyone_birthday	*always*
eat	*relationship*
boyfriend	*pool*
43%	**23%**

Observation 2


Jimmy from the BX @BloodwingBX · Apr 22
Replying to @HoarseWisperer @TheRickWilson
My extra sweet halfcaf double vegan soy chai pumpkin latte was 2 degrees hotter than it should have been and the foam wasn't very foamy. And they spelled my name Jimothy, "Jim" on the cup... it's a living hell here.
9 17 211

Dominant Topic	**2nd Dominant Topic**
Topic 3	Topic 2
swimming	*vegan*
swim	*yoga*
day	*job*
much	*every_woman*
support	*cooks_goe*
really	*therapy_remember*
try	*life_juggle*
always	*everyone_birthday*
relationship	*eat*
pool	*boyfriend*
37%	**33%**

Observation 2

Unrelated topic	Related topic
Dominant Topic	**2nd Dominant Topic**
Topic 3	Topic 2
swimming	*vegan*
swim	*yoga*
day	*job*
much	*every_woman*
support	*cooks_goe*
really	*therapy_remember*
try	*life_juggle*
always	*everyone_birthday*
relationship	*eat*
pool	*boyfriend*
37%	**33%**

Still Questionable!

- Why does the model give Misleading topic?
- Why does the model give Unrelated topic?
- Is there bias in data?

Still Questionable!

- Why does the model give Misleading topic?
- Why does the model give Unrelated topic?
- Is there bias in data?

Interpretability
&
Explainability

Still Questionable!

- Why does the model give Misleading topic?
- Why does the model give Unrelated topic?
- Is there bias in data?

Interpretability
&
Explainability

Future Work

Still Questionable!

- Why does the model give Misleading topic?
- Why does the model give Unrelated topic?
- Is there bias in data?

Interpretability
&
Explainability

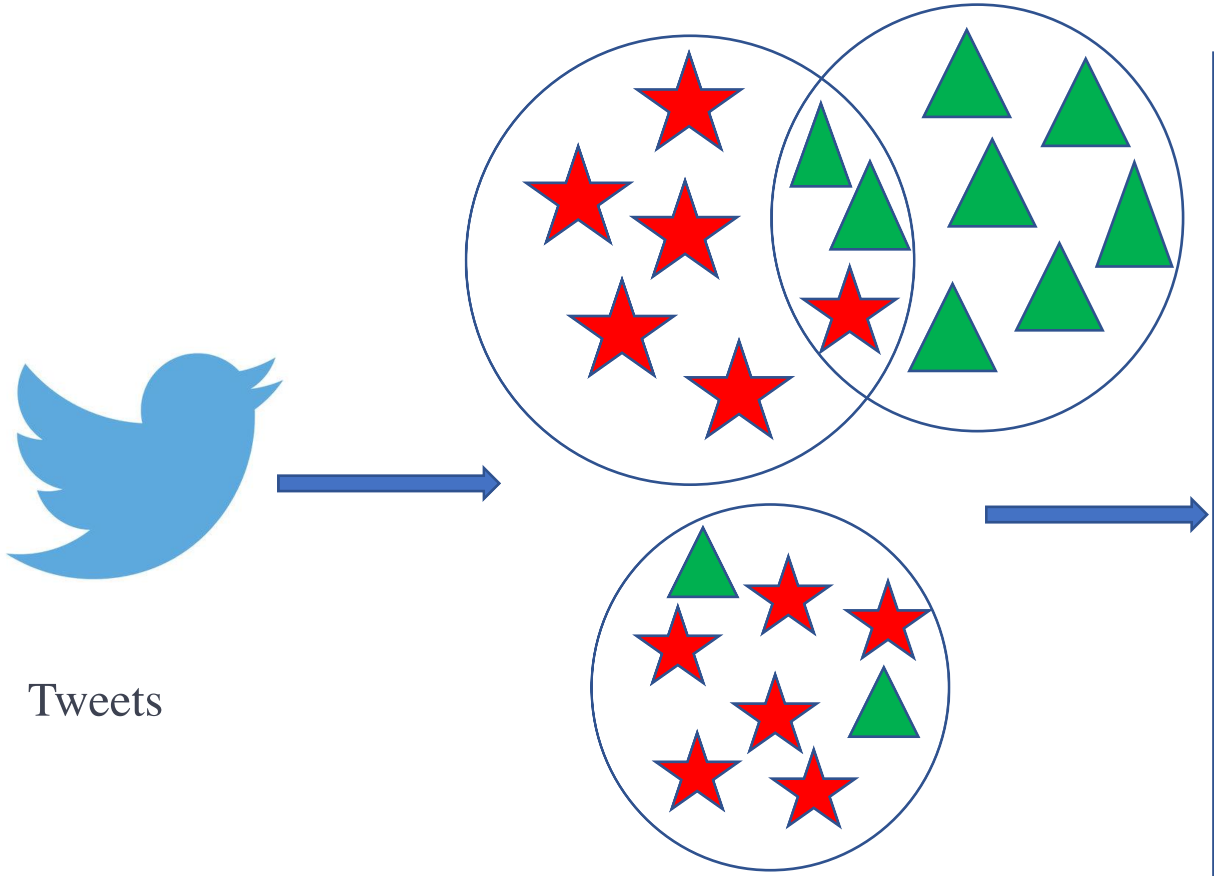
Future Work

- Analyze the Model interpretability

LIME: Local Interpretable model-agnostic Explanation

Ex-Twit: Explainable Twitter Mining on Health Data – Tunazzina Islam. Social NLP 2019 @IJCAI 2019.
Pre-print: https://arxiv.org/abs/1906.02132

Summary



Topic Modeling

LSA
NMF
LDA

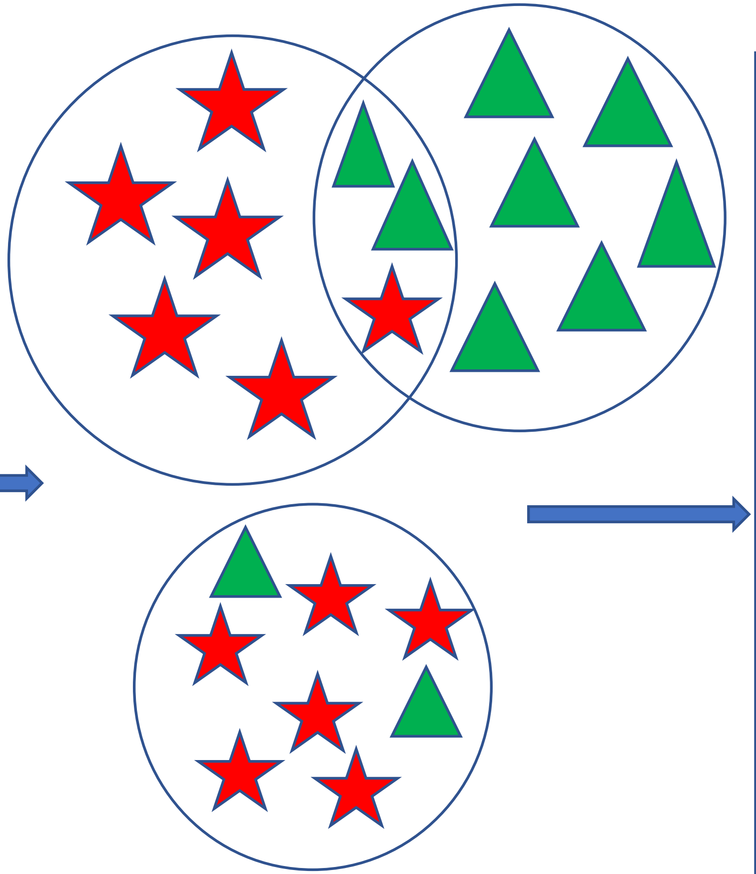
- Finding out dominant and 2nd dominant topic of each tweet (train data)
- Observing percentage of contribution of topic in each tweet
- Topic inference on new tweets (test data)
- Manual annotation both for train and test data to observe accuracy.
- Discovering interesting correlation

i.e. **Veganism and Yoga**

Topic Inference and Correlation Mining

QUESTION?

Tweets



Topic Modeling

LSA
NMF
LDA

- Finding out dominant and 2nd dominant topic of each tweet (train data)
- Observing percentage of contribution of topic in each tweet
- Topic inference on new tweets (test data)
- Manual annotation both for train and test data to observe accuracy.
- Discovering interesting correlation

i.e. **Veganism and Yoga**

Topic Inference and
Correlation Mining

THANK YOU

Tunazzina Islam

Ph.D. Student

Department of Computer Science

Purdue University, West Lafayette

PURDUE UNIVERSITY

islam32@purdue.edu https://tunazislam.github.io/ @Tunaz_Islam