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
Motivation

Balanced diet

Exercise

Running

Yoga

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

Twitter Streaming API

Producer

Single Node Kafka Broker

Single Kafka Topic– ‘twitterhealth’

ZooKeeper

Consumer

Data Pre-processing
Methodology of Correlation Mining

Twitter API

Streaming API i.e. Tweepy

Twitter

Extracted 40k Tweets (json format)

Parsed tweets

Raw tweets

Pre-processing

Cleaned Tweets

Manual Annotation 500 tweets (Train/Test)

Construct document-term matrix

Accuracy Comparison

Inferred Tweets

Topic Inference

Identified Topics

Topic Modeling

Bag-of-Words (BoW) model
Methodology of Correlation Mining

Twitter API
Streaming API i.e. Tweepy

Extracted 40k Tweets (json format)

Parsing tweets
Pre-processing

Raw tweets
Cleaned Tweets

Manual Annotation 500 tweets (Train/Test)

Accuracy Comparison
Inferred Tweets
Identified Topics

Topic Inference
Topic Modeling

Bag-of-Words (BoW) model
Construct document-term matrix
Methodology of Correlation Mining

1. **Streaming API** (e.g., Tweepy)
   - Extracted 40k Tweets (json format)

2. **Parsing tweets**
   - Raw tweets
   - Pre-processing

3. **Cleaned Tweets**
   - Construct document-term matrix

4. **Manual Annotation**
   - 500 tweets (Train/Test)

5. **Inferred Tweets**
   - Topic Inference

6. **Identified Topics**
   - Topic Modeling

7. **Bag-of-Words (BoW) model**

8. **Accuracy Comparison**
Methodology of Correlation Mining

1. **Streaming API (e.g., Tweepy)**
   - Twitter API

2. **Manual Annotation**
   - 500 tweets (Train/Test)

3. **Extracted 40k Tweets (json format)**

4. **Parsing tweets**
   - Raw tweets

5. **Pre-processing**
   - Cleaned Tweets

6. **Topic Modeling**
   - Construct document-term matrix

7. **Bag-of-Words (BoW) model**

8. **Identified Topics**

9. **Inferred Tweets**

10. **Accuracy Comparison**
Methodology of Correlation Mining

1. Twitter API (e.g., Tweepy)
2. Extracted 40k Tweets (json format)
3. Parsing tweets
4. Pre-processing
5. Cleaned Tweets
6. Manual Annotation: 500 tweets (Train/Test)
7. Topic Inference
8. Identified Topics
9. Inferred Tweets
10. Bag-of-Words (BoW) model
11. Construct document-term matrix
12. Accuracy Comparison

Flowchart:
- Streaming API
- Parsing tweets
- Pre-processing
- Cleaning Tweets
- Manual Annotation
- Topic Modeling
- Construct document-term matrix
Methodology of Correlation Mining

1. **Streaming API** (i.e., Tweepy)
   - Twitter API

2. **Extracted 40k Tweets** (json format)
   - Raw tweets

3. **Pre-processing**
   - Cleaned Tweets

4. **Manual Annotation**
   - 500 tweets (Train/Test)

5. **Accuracy Comparison**
6. **Inferred Tweets**
7. **Topic Inference**
8. **Identified Topics**
9. **Topic Modeling**
10. **Bag-of-Words (BoW) model**

**Construct document-term matrix**
Methodology of Correlation Mining

Twitter API

Streaming API i.e. Tweepy

Extracted 40k Tweets (json format)

Parsing tweets

Pre-processing

Raw tweets

Cleaned Tweets

Construct document-term matrix

Manual Annotation 500 tweets (Train/Test)

Inferred Tweets

Accuracy Comparison

Identified Topics

Topic Inference

Topic Modeling

Bag-of-Words (BoW) model
Overall Pipeline

Twitter Streaming API

Producer

Single Node Kafka Broker

Single Kafka Topic—‘twitterhealth’

Consumer

ZooKeeper

Accuracy Comparison with Ground Truth

Topic Inference of each Tweet

Spark Streaming

Topic Modeling
Topic Modeling Methodology

- Bag of Words (BoW)
- Topic Model

Topics

Tweets

Freq. of words in a topic
Topic Modeling Methodology

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

**Topic Model**
- Topics
- Freq. of words in a topic
- words

**Tweets**
Topic Modeling Methodology

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

**Topic Model**
- 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

**Tweets**

**Freq. of words in a topic**

**Words**
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
<table>
<thead>
<tr>
<th>Topic1</th>
<th>Topic2</th>
</tr>
</thead>
<tbody>
<tr>
<td>yoga</td>
<td>diet</td>
</tr>
<tr>
<td>everi</td>
<td>vegan</td>
</tr>
<tr>
<td>life</td>
<td>fit</td>
</tr>
<tr>
<td>job</td>
<td>day</td>
</tr>
<tr>
<td>remember</td>
<td>new</td>
</tr>
<tr>
<td>goe</td>
<td>like</td>
</tr>
<tr>
<td>woman</td>
<td>beyonce</td>
</tr>
<tr>
<td>everyone</td>
<td>amp</td>
</tr>
<tr>
<td>cook</td>
<td>eat</td>
</tr>
<tr>
<td>therapy</td>
<td>workout</td>
</tr>
</tbody>
</table>
Topics using LSA

<table>
<thead>
<tr>
<th>Topic1</th>
<th>Topic2</th>
<th>• highly dense matrix.</th>
</tr>
</thead>
<tbody>
<tr>
<td>yoga</td>
<td>diet</td>
<td></td>
</tr>
<tr>
<td>everi</td>
<td>vegan</td>
<td></td>
</tr>
<tr>
<td>life</td>
<td>fit</td>
<td></td>
</tr>
<tr>
<td>job</td>
<td>day</td>
<td></td>
</tr>
<tr>
<td>remember</td>
<td>new</td>
<td></td>
</tr>
<tr>
<td>goe</td>
<td>like</td>
<td></td>
</tr>
<tr>
<td>woman</td>
<td>beyonce</td>
<td></td>
</tr>
<tr>
<td>everyone</td>
<td>amp</td>
<td></td>
</tr>
<tr>
<td>cook</td>
<td>eat</td>
<td></td>
</tr>
<tr>
<td>therapy</td>
<td>workout</td>
<td></td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Topic1</th>
<th>Topic2</th>
<th>Topic3</th>
<th>Topic4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yoga</td>
<td>diet</td>
<td>vegan</td>
<td>fitness</td>
</tr>
<tr>
<td>job</td>
<td>beyonce</td>
<td>go</td>
<td>workout</td>
</tr>
<tr>
<td>every_woman</td>
<td>new</td>
<td>eat</td>
<td>go</td>
</tr>
<tr>
<td>cooks_goe</td>
<td>bitch</td>
<td>make</td>
<td>good</td>
</tr>
<tr>
<td>therapy_remember</td>
<td>ciara_prayer</td>
<td>food</td>
<td>amp</td>
</tr>
<tr>
<td>life_juggl</td>
<td>day</td>
<td>day</td>
<td>day</td>
</tr>
<tr>
<td>everyone_birthday</td>
<td>eat</td>
<td>amp</td>
<td>yoga</td>
</tr>
<tr>
<td>boyfriend</td>
<td>go</td>
<td>shit</td>
<td>health</td>
</tr>
<tr>
<td>hot</td>
<td>fat</td>
<td>meat</td>
<td>gym</td>
</tr>
<tr>
<td>know</td>
<td>keto</td>
<td>vegetarian</td>
<td>today</td>
</tr>
<tr>
<td>Topic1</td>
<td>Topic2</td>
<td>Topic3</td>
<td>Topic4</td>
</tr>
<tr>
<td>------------------------</td>
<td>------------------------------</td>
<td>-----------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>Yoga</td>
<td>diet</td>
<td>vegan</td>
<td>fitness</td>
</tr>
<tr>
<td>job</td>
<td>beyonce</td>
<td>go</td>
<td>workout</td>
</tr>
<tr>
<td>every_woman</td>
<td>new</td>
<td>eat</td>
<td>go</td>
</tr>
<tr>
<td>cooks_goe</td>
<td>bitch</td>
<td>make</td>
<td>good</td>
</tr>
<tr>
<td>therapy_remember</td>
<td>ciara_prayer</td>
<td>food</td>
<td>amp</td>
</tr>
<tr>
<td>life_juggl</td>
<td>day</td>
<td>day</td>
<td>day</td>
</tr>
<tr>
<td>everyone_birthday</td>
<td>eat</td>
<td>amp</td>
<td>yoga</td>
</tr>
<tr>
<td>boyfriend</td>
<td>go</td>
<td>shit</td>
<td>health</td>
</tr>
<tr>
<td>hot</td>
<td>fat</td>
<td>meat</td>
<td>gym</td>
</tr>
<tr>
<td>know</td>
<td>keto</td>
<td>vegetarian</td>
<td>today</td>
</tr>
</tbody>
</table>
## Topics using NMF

### Topic1
- Yoga
- job
- every_woman
- cooks_goe
- therapy_remember
- life_juggl
- everyone_birthday
- boyfriend
- hot
- know

### Topic2
- diet
- beyonce
- new
- bitch
- ciara_prayer
- day
- eat
- go
- fat
- keto

### Topic3
- vegan
- go
- eat
- make
- food
- day
- amp
- shit
- meat
- vegetarian

### Topic4
- fitness
- go
- workout
- eat
- good
- amp
- day
- yoga
- health
- gym
- today

- **sparse representations**
- **same keywords are repeated in multiple topics.**
<table>
<thead>
<tr>
<th>Topic1</th>
<th>Topic2</th>
<th>Topic3</th>
<th>Topic4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yoga</td>
<td>diet</td>
<td>vegan</td>
<td>fitness</td>
</tr>
<tr>
<td>job</td>
<td>beyonce</td>
<td>go</td>
<td>workout</td>
</tr>
<tr>
<td>every_woman</td>
<td>new</td>
<td>eat</td>
<td>go</td>
</tr>
<tr>
<td>cooks_goe</td>
<td>bitch</td>
<td>make</td>
<td>good</td>
</tr>
<tr>
<td>therapy_remember</td>
<td>ciara_prayer</td>
<td>food</td>
<td>amp</td>
</tr>
<tr>
<td>life_juggl</td>
<td>day</td>
<td>day</td>
<td>day</td>
</tr>
<tr>
<td>everyone_birthday</td>
<td>eat</td>
<td>amp</td>
<td>yoga</td>
</tr>
<tr>
<td>boyfriend</td>
<td>go</td>
<td>shit</td>
<td>health</td>
</tr>
<tr>
<td>hot</td>
<td>fat</td>
<td>meat</td>
<td>gym</td>
</tr>
<tr>
<td>know</td>
<td>keto</td>
<td>vegetarian</td>
<td>today</td>
</tr>
</tbody>
</table>

- sparse representations
- same keywords are repeated in multiple topics.
### Topics using NMF

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

- **Sparse representations**
- **Same keywords are repeated in multiple topics.**
Optimal Number of Topics vs Coherence Score LDA

$K = 4$

Coherence Value = 0.3871
<table>
<thead>
<tr>
<th>Topic1</th>
<th>Topic2</th>
<th>Topic3</th>
<th>Topic4</th>
</tr>
</thead>
<tbody>
<tr>
<td>diet</td>
<td>vegan</td>
<td>swimming</td>
<td>fitness</td>
</tr>
<tr>
<td>workout</td>
<td>yoga</td>
<td>swim</td>
<td>amp</td>
</tr>
<tr>
<td>new</td>
<td>job</td>
<td>day</td>
<td>wellness</td>
</tr>
<tr>
<td>go</td>
<td>every_woman</td>
<td>much</td>
<td>health</td>
</tr>
<tr>
<td>day</td>
<td>yoga</td>
<td>support</td>
<td>time</td>
</tr>
<tr>
<td>beyonce</td>
<td>every_woman</td>
<td>really</td>
<td>great</td>
</tr>
<tr>
<td>get</td>
<td>cooks_goe</td>
<td>try</td>
<td>look</td>
</tr>
<tr>
<td>today</td>
<td>therapy_remember</td>
<td>always</td>
<td>hiking</td>
</tr>
<tr>
<td>bitch</td>
<td>life_juggle</td>
<td>relationship</td>
<td>make</td>
</tr>
<tr>
<td>gym</td>
<td>everyone_birthDay</td>
<td>pool</td>
<td>love</td>
</tr>
<tr>
<td></td>
<td>eat</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>boyfriend</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Topic 1</strong></td>
<td><strong>Topic 2</strong></td>
<td><strong>Topic 3</strong></td>
<td><strong>Topic 4</strong></td>
</tr>
<tr>
<td>-------------------</td>
<td>-------------------</td>
<td>-------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>diet</td>
<td>vegan</td>
<td>swimming</td>
<td>fitness</td>
</tr>
<tr>
<td>workout</td>
<td>yoga</td>
<td>swim</td>
<td>amp</td>
</tr>
<tr>
<td>new</td>
<td>job</td>
<td>day</td>
<td>wellness</td>
</tr>
<tr>
<td>go</td>
<td>every_woman</td>
<td>much</td>
<td>health</td>
</tr>
<tr>
<td>day</td>
<td>cooks_goe</td>
<td>support</td>
<td>time</td>
</tr>
<tr>
<td>beyonce</td>
<td>therapy_remember</td>
<td>really</td>
<td>great</td>
</tr>
<tr>
<td>get</td>
<td>life_juggle</td>
<td>try</td>
<td>look</td>
</tr>
<tr>
<td>today</td>
<td>everyone_birthday</td>
<td>always</td>
<td>hiking</td>
</tr>
<tr>
<td>bitch</td>
<td>eat</td>
<td>relationship</td>
<td>make</td>
</tr>
<tr>
<td>gym</td>
<td>boyfriend</td>
<td>pool</td>
<td>love</td>
</tr>
</tbody>
</table>

- *coherent topics*
Visualization of Topics - pyLDAVIS

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Salient Terms

Online link: [https://tunazislam.github.io/files/LDA_Visualization_t4.html](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, 2\textsuperscript{nd} dominant topic and its percentage of contribution in each Tweet.

Example:

Veruka Salt @LesegoMasithela · Apr 18
Revoking my vegetarian status till further notice. There's something I wanna do and I can't afford the supplements that come with being veggie

<table>
<thead>
<tr>
<th>Dominant Topic</th>
<th>2\textsuperscript{nd} Dominant Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 2</td>
<td>Topic 1</td>
</tr>
<tr>
<td>vegan</td>
<td>diet</td>
</tr>
<tr>
<td>yoga</td>
<td>workout</td>
</tr>
<tr>
<td>job</td>
<td>new</td>
</tr>
<tr>
<td>every_woman</td>
<td>new</td>
</tr>
<tr>
<td>cooks_goe</td>
<td>new</td>
</tr>
<tr>
<td>therapy_remember</td>
<td>new</td>
</tr>
<tr>
<td>life_juggle</td>
<td>new</td>
</tr>
<tr>
<td>everyone_birthday</td>
<td>new</td>
</tr>
<tr>
<td>eat</td>
<td>new</td>
</tr>
<tr>
<td>boyfriend</td>
<td>new</td>
</tr>
<tr>
<td>61%</td>
<td>18%</td>
</tr>
</tbody>
</table>
Topic Inference on **New** Tweets (Test data)

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

**Example:**

Larry D. Williamson
@W nigroup

I would like to take time to wish "ALL" a very happy #EarthDay! #yoga #meditation

12:32 PM - 22 Apr 2019

Dominant Topic: Topic 2
- vegan
- yoga
- job
- every_woman
- cooks_goe
- therapy_remember
- life_juggle
- everyone_birthday
- eat
- boyfriend
- 33%

2nd Dominant Topic: Topic 4
- fitness
- amp
- wellness
- health
- time
- great
- look
- hiking
- make
- love
- 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.*
  • 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.*

  • Tweet 2: *Why You Should #LiftWeights to Lose #BellyFat #Fitness #core #abs #diet #gym #bodybuilding #workout #yoga*
Train/Test Accuracy with Ground Truth

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

Train/Test Size vs Accuracy

<table>
<thead>
<tr>
<th>Size of dataset</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>39</td>
</tr>
<tr>
<td>100</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td></td>
</tr>
<tr>
<td>300</td>
<td></td>
</tr>
<tr>
<td>400</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td></td>
</tr>
<tr>
<td>600</td>
<td></td>
</tr>
</tbody>
</table>
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...

Dominant Topic

<table>
<thead>
<tr>
<th>Topic 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>vegan</td>
</tr>
<tr>
<td>yoga</td>
</tr>
<tr>
<td>job</td>
</tr>
<tr>
<td>every_woman</td>
</tr>
<tr>
<td>cooks goe</td>
</tr>
<tr>
<td>therapy_remember</td>
</tr>
<tr>
<td>life_juggle</td>
</tr>
<tr>
<td>everyone_birthday</td>
</tr>
<tr>
<td>eat</td>
</tr>
<tr>
<td>boyfriend</td>
</tr>
</tbody>
</table>

2nd Dominant Topic

<table>
<thead>
<tr>
<th>Topic 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>swimming</td>
</tr>
<tr>
<td>swim</td>
</tr>
<tr>
<td>day</td>
</tr>
<tr>
<td>much</td>
</tr>
<tr>
<td>support</td>
</tr>
<tr>
<td>really</td>
</tr>
<tr>
<td>try</td>
</tr>
<tr>
<td>always</td>
</tr>
<tr>
<td>relationship</td>
</tr>
<tr>
<td>pool</td>
</tr>
</tbody>
</table>
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

Dominant Topic

- vegan
- yoga
- job
- every_woman
- cooks_goe
- therapy_remember
- life_juggle
- everyone_birthday
- eat
- boyfriend

43%

2nd Dominant Topic

- swimming
- swim
- day
- much
- support
- really
- try
- always
- relationship
- pool

23%

Misleading topic
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
 Observation 2

Dominant Topic

Topic 3
- swimming
- swim
- day
- much
- support
- really
- try
- always
- relationship
- pool

37%

2nd Dominant Topic

Topic 2
- vegan
- yoga
- job
- every_woman
- cooks_goe
- therapy_remember
- life_juggle
- everyone_birthday
- eat
- boyfriend

33%

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.
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.
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?

Future Work

- Analyze the Model interpretability

**LIME**: Local Interpretable model-agnostic Explanation

Summary

- 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 Modeling

LSA
NMF
LDA

Topic Inference and Correlation Mining
Tweets

Topic Modeling

- 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

LSA
NMF
LDA

Topic Inference and Correlation Mining
THANK YOU

Tunazzina Islam
Ph.D. Student
Department of Computer Science
Purdue University, West Lafayette

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