Big Data Analytics Project

Ashmeet Saluja, Kamna Surjani, Matthew Woodfill,
Nikita Singh, Nishkarsh Gupta

Introduction

Our business concept revolves around crafting Airbnb properties themed after popular movies, TV shows, and cultural trends.This strategy provides guests with unique experiences and gives hosts a competitive advantage by using data analytics to ensure relevance and profitability.

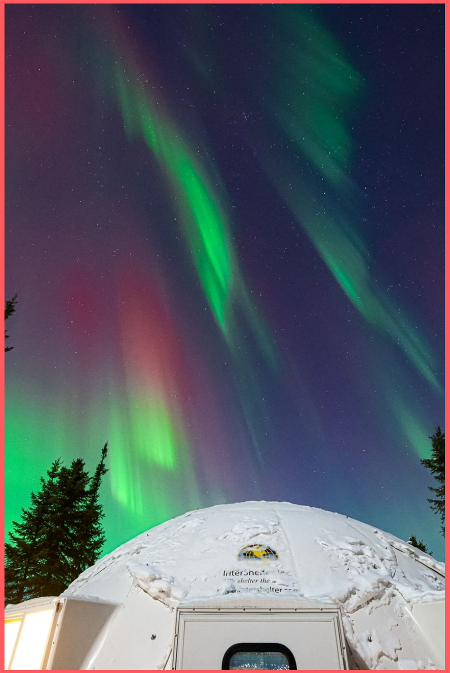
Market Opportunity

Treehouse Villa in Atlanta

Tiny House in Austin

Lighthouse in Marine

Glamping Tent in Sedona

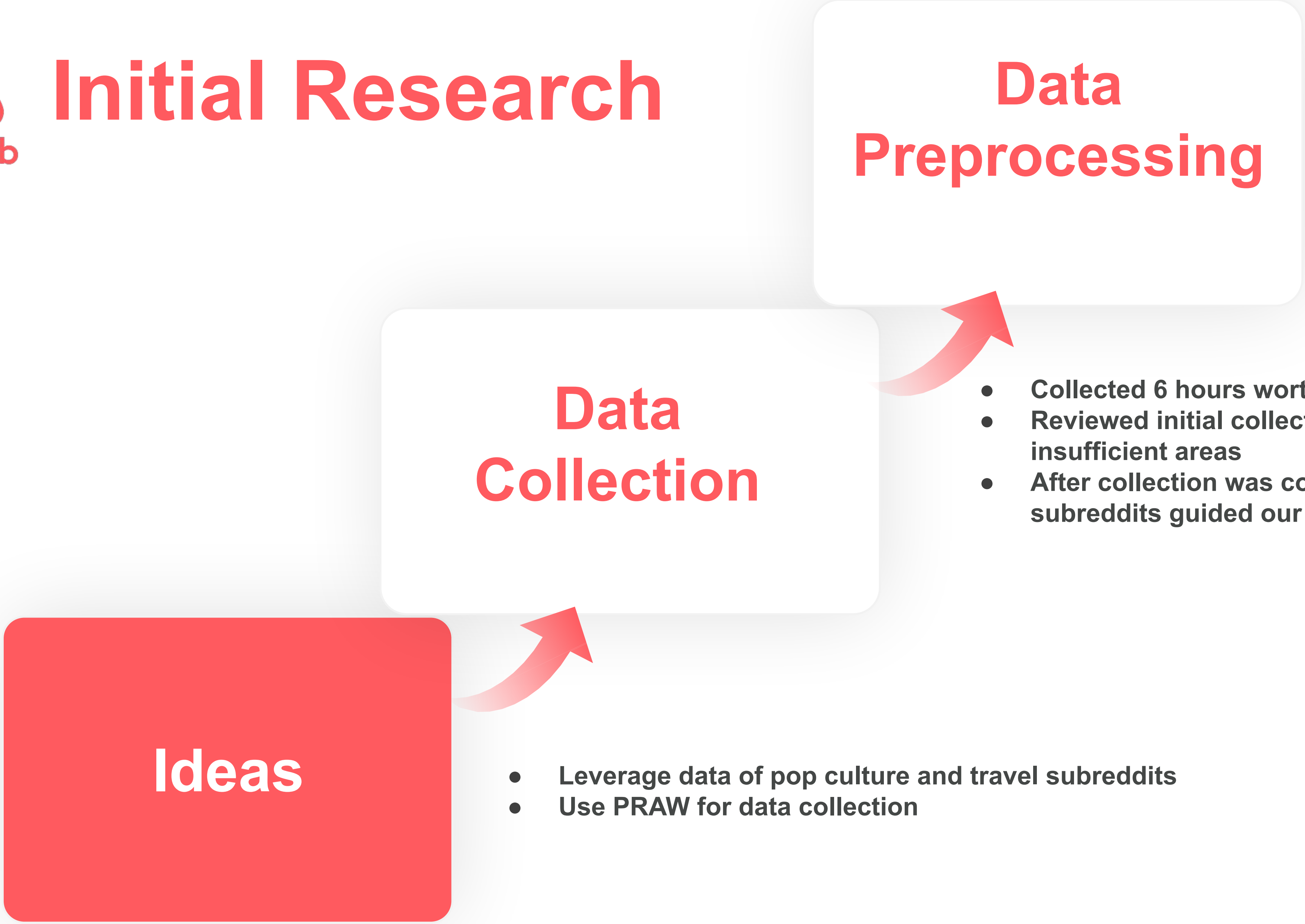


Igloo in Alaska

CUSTOMER SEGMENTS

- Pop Culture Enthusiasts
- Young Adults and Millennials
- Event Attendees
- Families
- Nostalgia Seekers
- Luxury Travelers
- Influencers and Content Creators
- Solo Travellers

Initial Research



Ideas

- Leverage data of pop culture and travel subreddits
- Use PRAW for data collection

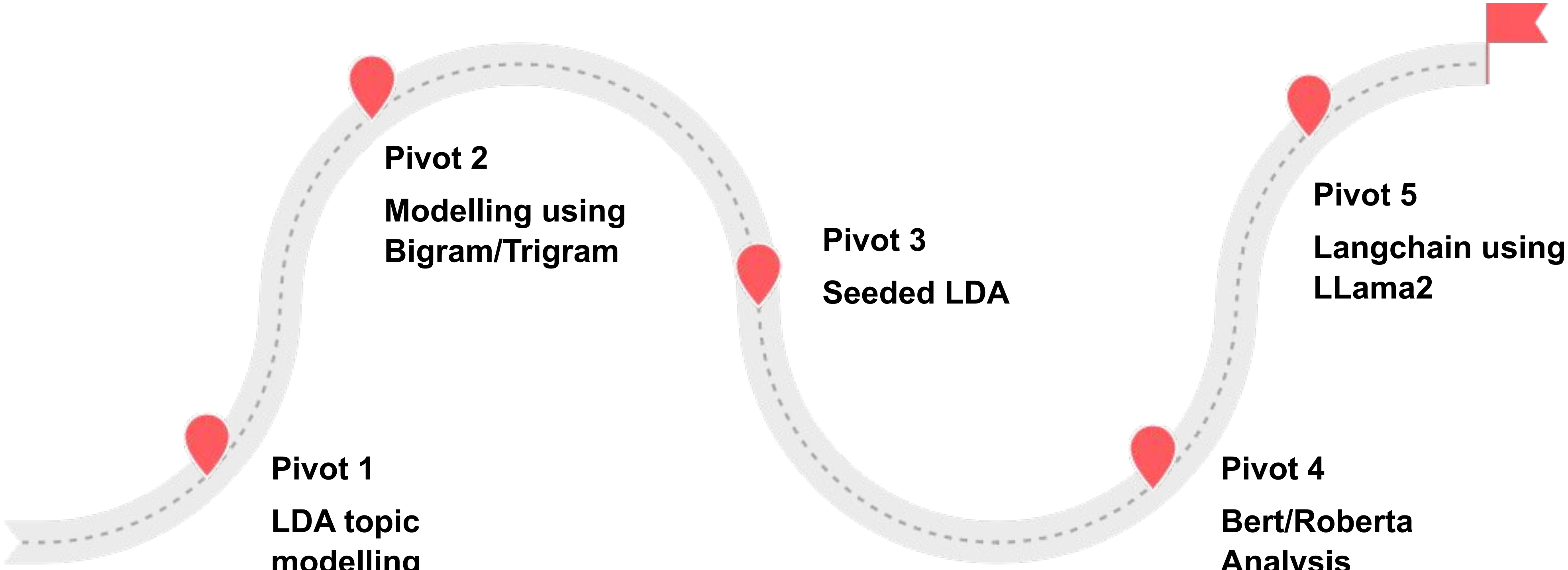
Data Collection

- Collected 6 hours worth of data from each subreddit
- Reviewed initial collection, additional data collected for insufficient areas
- After collection was completed, limited data among subreddits guided our focus to pop culture themes

Data Preprocessing

- Approach 1: Given script
- Approach 2: Iterative stop word removal
- Approach 3: ThreadID-Based Comment Grouping

Project Timeline and Major Pivots



Initial Research

Pivot 1

LDA topic modelling

Pivot 2

Modelling using Bigram/Trigram

Pivot 3

Seeded LDA

Pivot 4

Bert/Roberta Analysis

Pivot 5

Langchain using LLama2

1

LDA Topic Modelling

Why tried this: Popular method for extracting topics and can help in identifying most latent topics

Reason to pivot:

- Struggled with the granularity and specificity of topics
- The actionable topics were few, and lacked clear actionable words for application
- It had limited useful insights

2

Topic Modelling in Bigram/Trigram

Why tried this: Aimed to capture more specific phrases and context that single words might miss

Reason to pivot:

- Resulted in a limited number of actionable insights
- Too much noise from less relevant phrase combinations
- Did not improve the robustness of our output and actually did not repeat results we had already achieved.

3

Seeded LDA

Why tried this: To further focus on the topics by introducing seed words

Reason to pivot:

- Underperformed because the seed topics might not have been well represented in the dataset
- Chosen seeds were too narrow, limiting the model's ability to generalize from the data

4

BERT & Roberta Sentiment Analysis

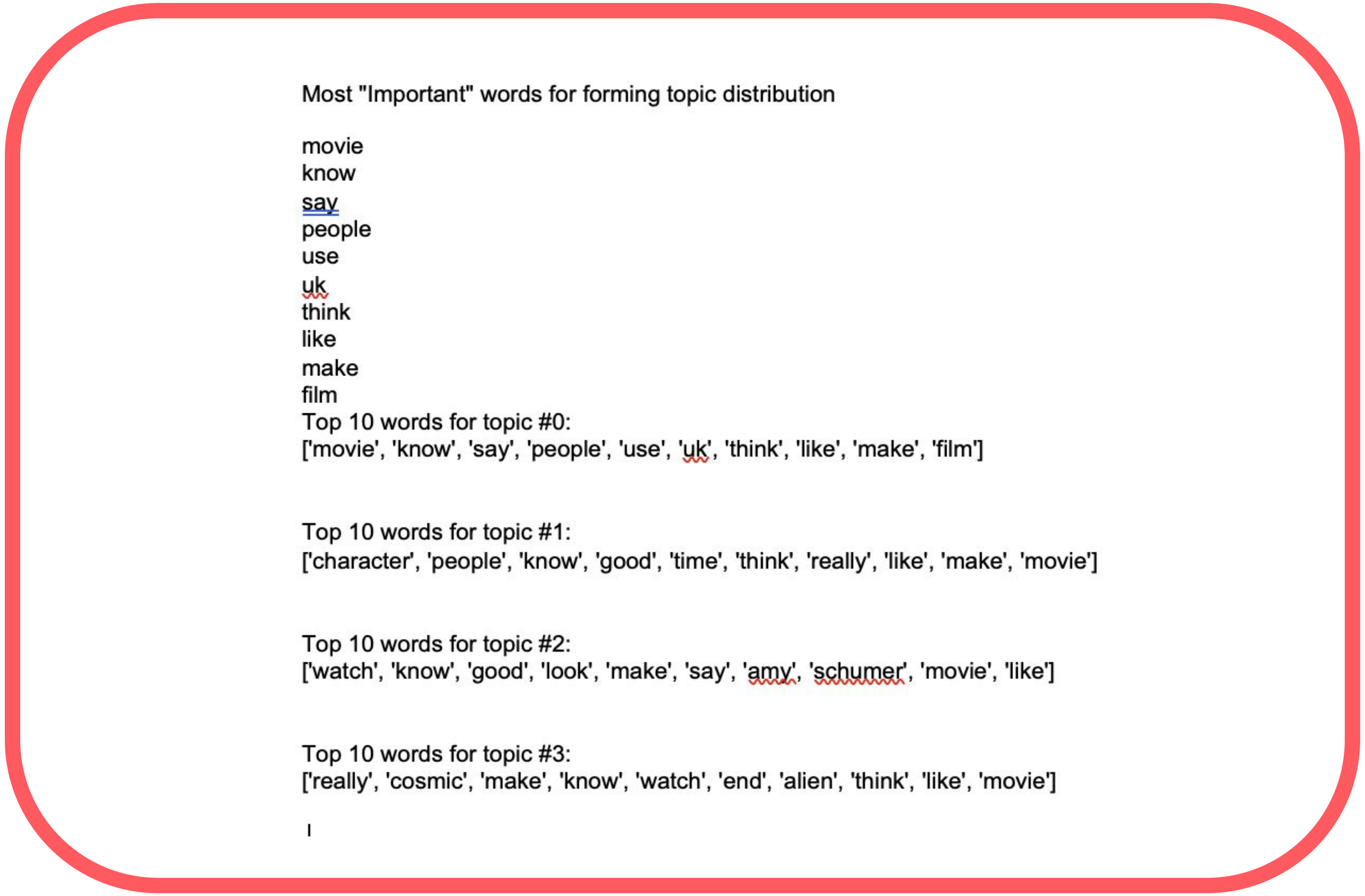
Why tried this: It is to provide robust results because of the bidirectional approach it reviews words.

Reason to pivot:

- We were not receiving robust output from clustering and it was unclear how to use the results from the sentiment analysis
- This model also was time intensive to run and could only work on a couple of our teams computers.

<BACK

LDA Topic Modelling



Most "Important" words for forming topic distribution

movie
know
say
people
use
uk
think
like
make
film
Top 10 words for topic #0:
['movie', 'know', 'say', 'people', 'use', 'uk', 'think', 'like', 'make', 'film']

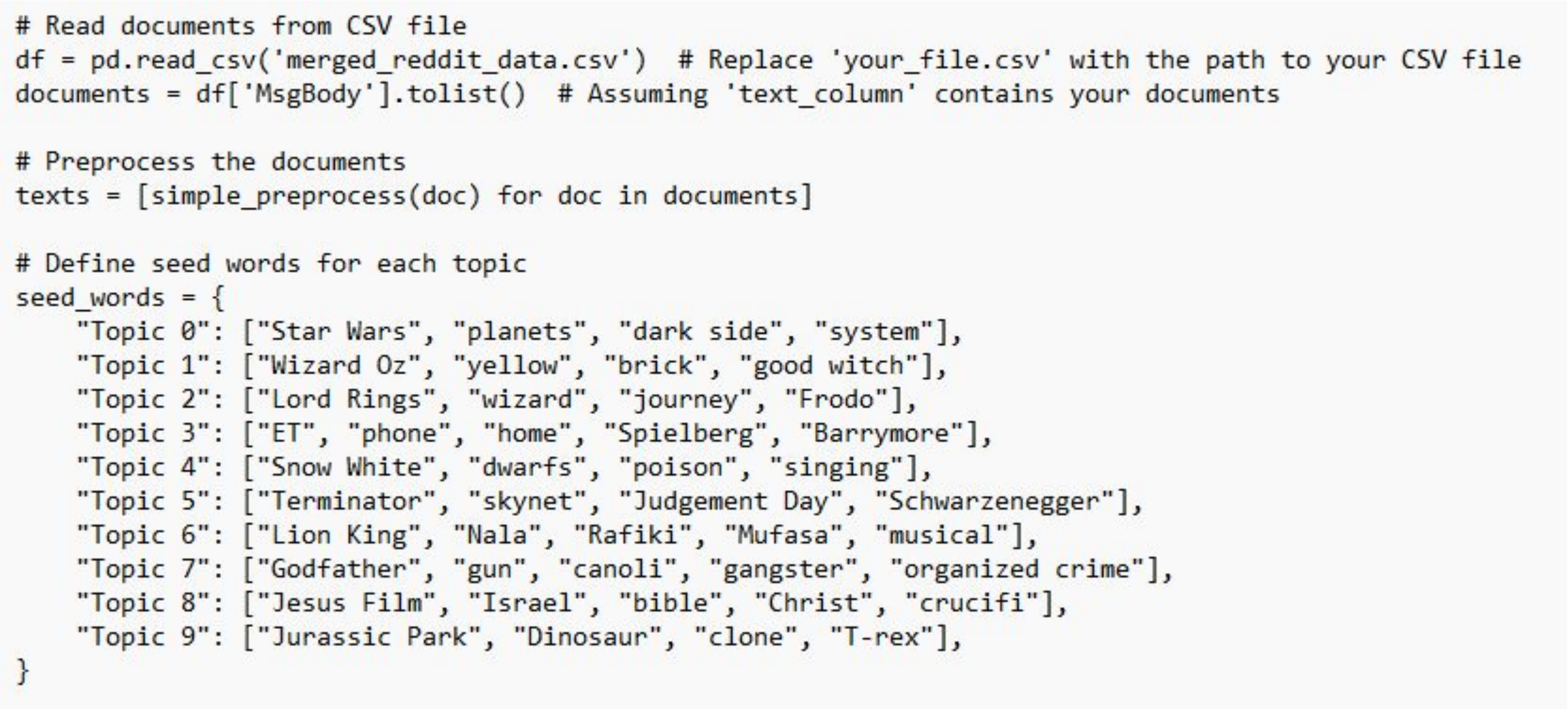
Top 10 words for topic #1:
['character', 'people', 'know', 'good', 'time', 'think', 'really', 'like', 'make', 'movie']

Top 10 words for topic #2:
['watch', 'know', 'good', 'look', 'make', 'say', 'amy', 'schumer', 'movie', 'like']

Top 10 words for topic #3:
['really', 'cosmic', 'make', 'know', 'watch', 'end', 'alien', 'think', 'like', 'movie']

<BACK

Seeded LDA



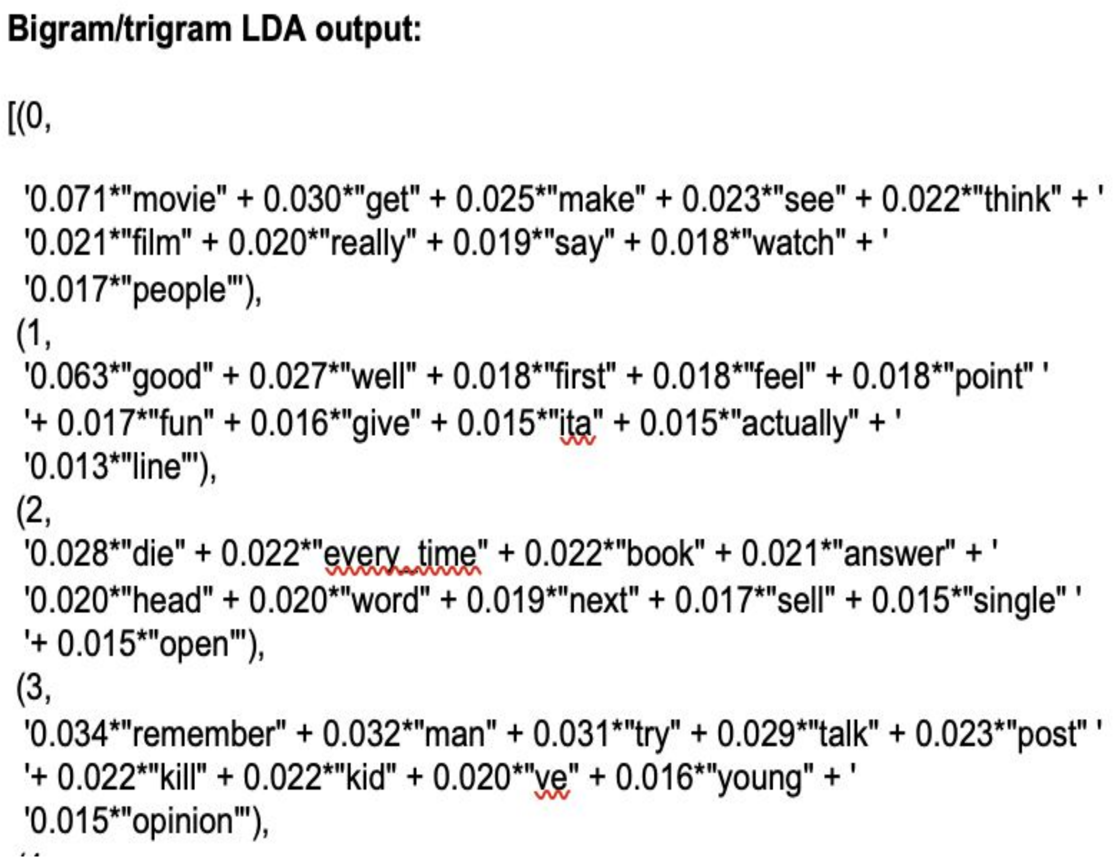
```
# Read documents from CSV file
df = pd.read_csv('merged_reddit_data.csv')  # Replace 'your_file.csv' with the path to your CSV file
documents = df['MsgBody'].tolist()  # Assuming 'text_column' contains your documents

# Preprocess the documents
texts = [simple_preprocess(doc) for doc in documents]

# Define seed words for each topic
seed_words = {
    "Topic 0": ["Star Wars", "planets", "dark side", "system"],
    "Topic 1": ["Wizard Oz", "yellow", "brick", "good witch"],
    "Topic 2": ["Lord Rings", "wizard", "journey", "Frodo"],
    "Topic 3": ["ET", "phone", "home", "Spielberg", "Barrymore"],
    "Topic 4": ["Snow White", "dwarfs", "poison", "singing"],
    "Topic 5": ["Terminator", "skynet", "Judgement Day", "Schwarzenegger"],
    "Topic 6": ["Lion King", "Nala", "Rafiki", "Mufasa", "musical"],
    "Topic 7": ["Godfather", "gun", "canoli", "gangster", "organized crime"],
    "Topic 8": ["Jesus Film", "Israel", "bible", "Christ", "crucifi"],
    "Topic 9": ["Jurassic Park", "Dinosaur", "clone", "T-rex"],
}
```

<BACK

Bigram/Trigram LDA

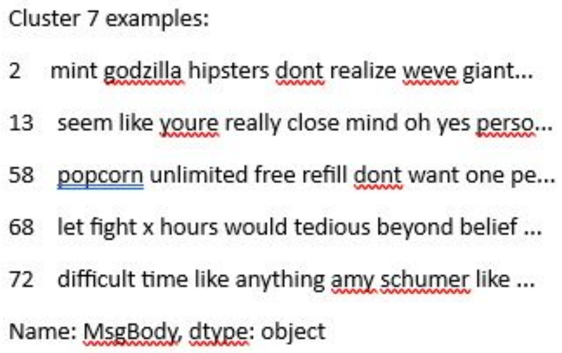


Bigram/trigram LDA output:

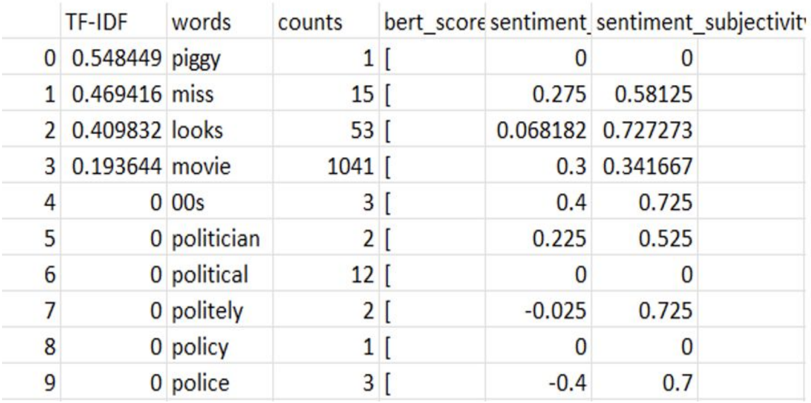
```
[(0,
  '0.071*"movie" + 0.030*"get" + 0.025*"make" + 0.023*"see" + 0.022*"think" + '
  '0.021*"film" + 0.020*"really" + 0.019*"say" + 0.018*"watch" + '
  '0.017*"people"'),
 (1,
  '0.063*"good" + 0.027*"well" + 0.018*"first" + 0.018*"feel" + 0.018*"point" '
  '+ 0.017*"fun" + 0.016*"give" + 0.015*"ita" + 0.015*"actually" + '
  '0.013*"line"'),
 (2,
  '0.028*"die" + 0.022*"every_time" + 0.022*"book" + 0.021*"answer" + '
  '0.020*"head" + 0.020*"word" + 0.019*"next" + 0.017*"sell" + 0.015*"single" '
  '+ 0.015*"open"'),
 (3,
  '0.034*"remember" + 0.032*"man" + 0.031*"try" + 0.029*"talk" + 0.023*"post" '
  '+ 0.022*"kill" + 0.022*"kid" + 0.020*"ve" + 0.016*"young" + '
  '0.015*"opinion"'),
```

airbnb

BERT

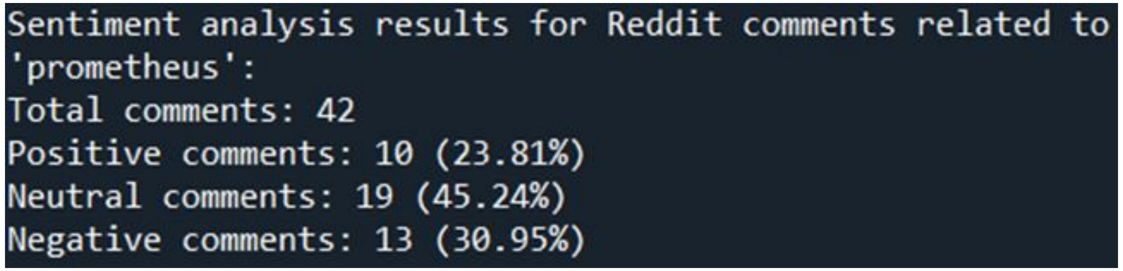


```
Cluster 7 examples:
2     mint godzilla hipsters dont realize weve giant...
13    seem like youre really close mind oh yes perso...
58    popcorn unlimited free refill dont want one pe...
68    let fight x hours would tedious beyond belief ...
72    difficult time like anything amy schumer like ...
Name: MsgBody, dtype: object
```



	TF-IDF	words	counts	bert_score	sentiment_	sentiment_subjectivit
0	0.548449	piggy	1	[	0	0
1	0.469416	miss	15	[	0.275	0.58125
2	0.409832	looks	53	[	0.068182	0.727273
3	0.193644	movie	1041	[	0.3	0.341667
4	0	00s	3	[	0.4	0.725
5	0	politician	2	[	0.225	0.525
6	0	political	12	[	0	0
7	0	politely	2	[	-0.025	0.725
8	0	policy	1	[	0	0
9	0	police	3	[	-0.4	0.7

ROBERTA



```
Sentiment analysis results for Reddit comments related to
'prometheus':
Total comments: 42
Positive comments: 10 (23.81%)
Neutral comments: 19 (45.24%)
Negative comments: 13 (30.95%)
```

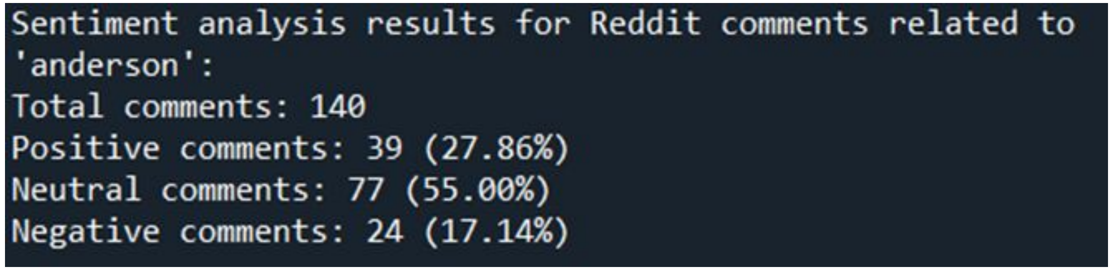
Topic Modeling Results

Throughout the different methods the repeated topics were:

- Godzilla & King Kong
- Alien, Prometheus & Predator
- Wes Anderson

Roberta sentiment analysis supports pursuing themes associated with:

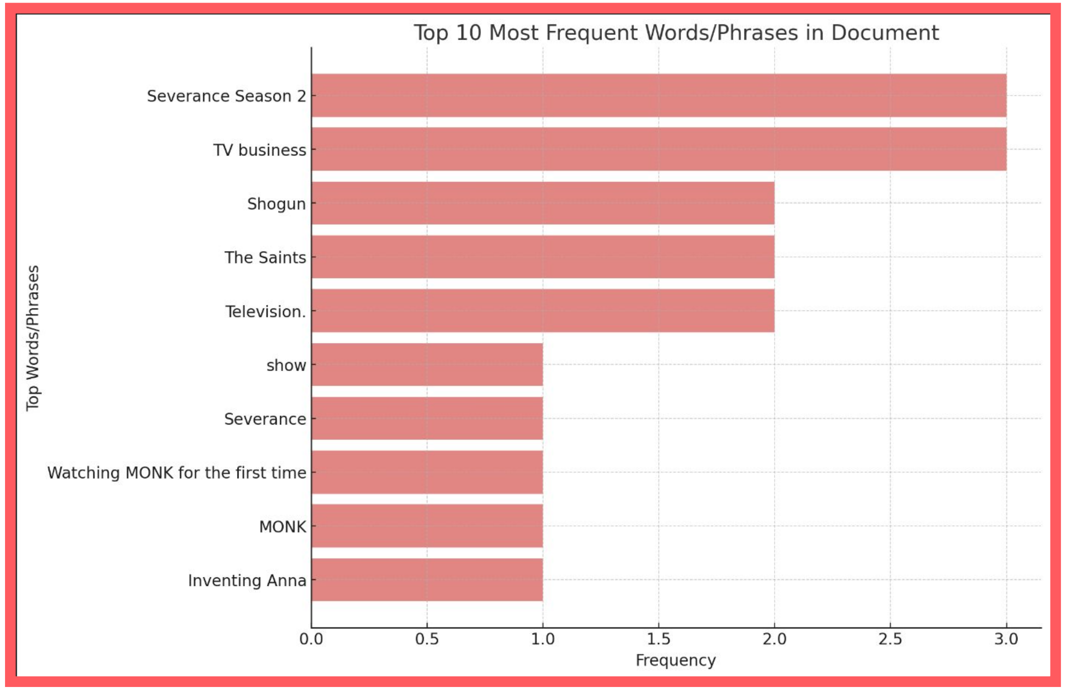
Wes Anderson



```
Sentiment analysis results for Reddit comments related to
'anderson':
Total comments: 140
Positive comments: 39 (27.86%)
Neutral comments: 77 (55.00%)
Negative comments: 24 (17.14%)
```

Langchain with LLama2

We utilized oLLaMa to locally harness the capabilities of large language models, incorporating LangChain with the LLaMA2 model to enhance our text data analysis. This approach was specifically chosen to improve text classification and comprehension. Our goal was to systematically categorize and summarize prevalent themes into distinct categories like television, movies, and books, thus offering a clear and organized snapshot of popular conversation topics.



Future Directions

- Apply sentiment analysis using RoBERTa to evaluate user opinions on outputs from the LangChain model.
- Planning to explore other models, beside LLam2 for more efficient results.
- Expand the training of LangChain to use Ollama on the full dataset; previously, it was trained on just 100 rows as a proof of concept
- Refine current preprocessing techniques and explore additional tools such as word2vec and fast text
- Experiment further with the models in the transformer architect family
- Expand data collection across additional social media platforms such as X and Instagram

airbnb

Thank You!