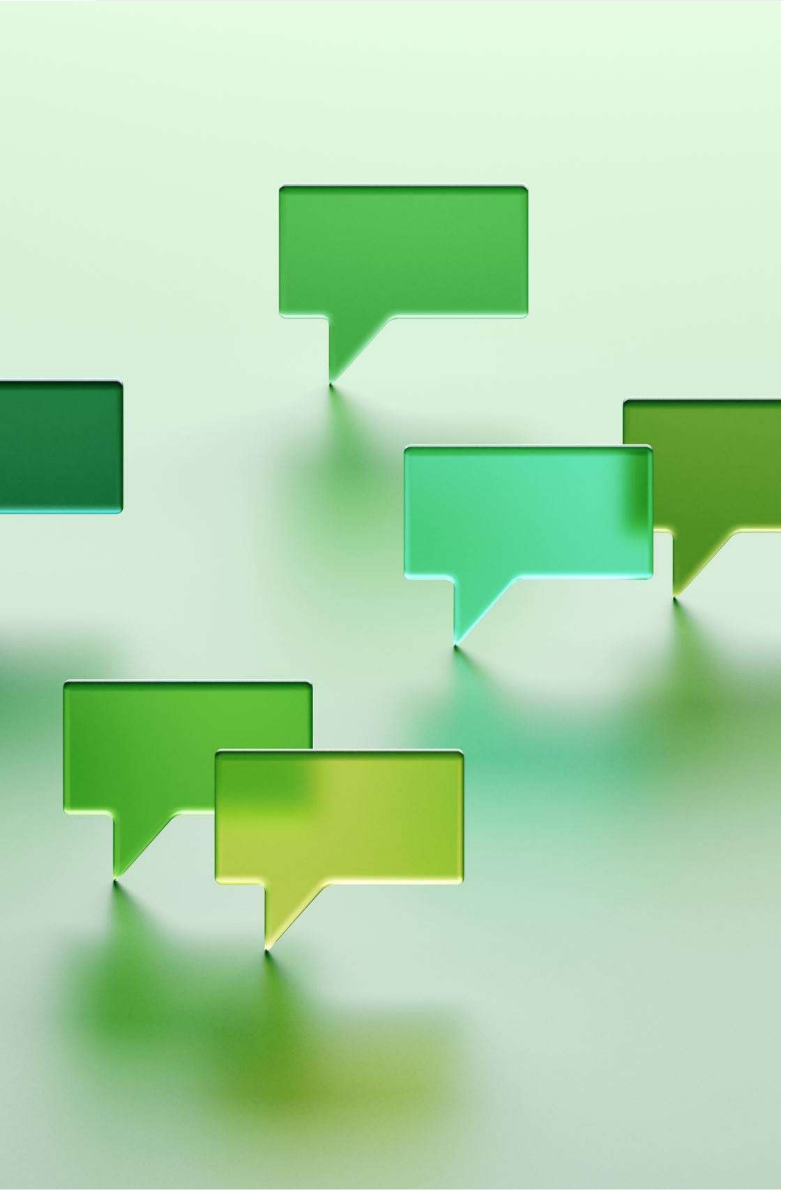
Trending Topic Analysis of Twitter Dataset using Latent Dirichlet Allocation (LDA) and ChatGPT

Christoffer Tan & Janis Joplin

(Group 7)

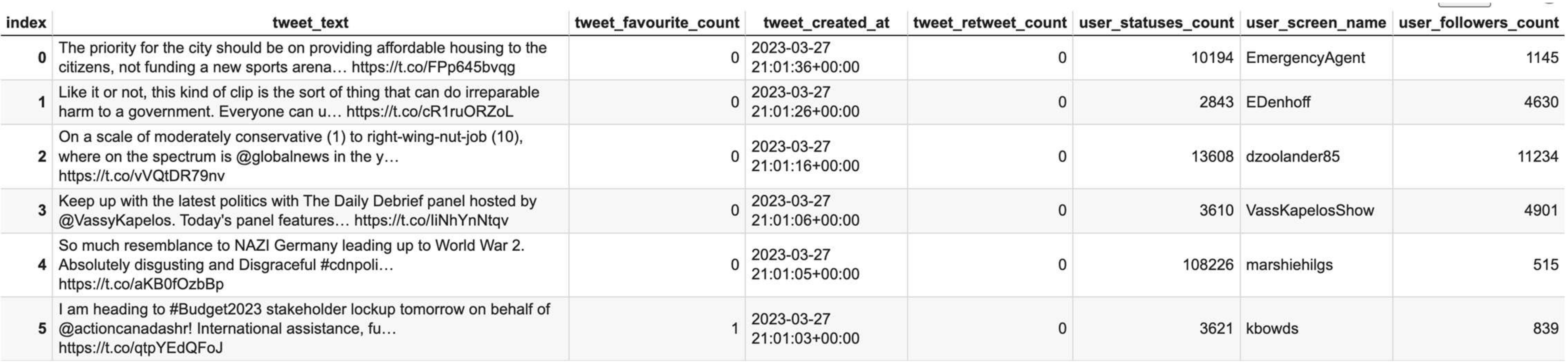


Problem: Identify Trending Topics

- **Goal:** Identify trending topics in social media (Twitter)
- **Usage:** Provide insights into public sentiment and guides marketing strategies.
- **Challenges:**
    - o dynamic nature of trends
    - o noise and bias within the discussions.

Overview of Data (Tweets)

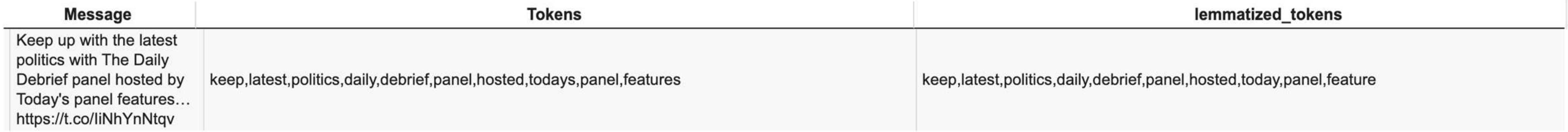
- Sample of 10,000 tweets that are created on 23-27 March 2023.
- No missing values
- Used feature: ***tweet text*** *(truncated)*



index	tweet_text	tweet_favourite_count	tweet_created_at	tweet_retweet_count	user_statuses_count	user_screen_name	user_followers_count
0	The priority for the city should be on providing affordable housing to the citizens, not funding a new sports arena… https://t.co/FPp645bvqg	0	2023-03-27 21:01:36+00:00	0	10194	EmergencyAgent	1145
1	Like it or not, this kind of clip is the sort of thing that can do irreparable harm to a government. Everyone can u… https://t.co/cR1ruORZoL	0	2023-03-27 21:01:26+00:00	0	2843	EDenhoff	4630
2	On a scale of moderately conservative (1) to right-wing-nut-job (10), where on the spectrum is @globalnews in the y… https://t.co/vVQtDR79nv	0	2023-03-27 21:01:16+00:00	0	13608	dzoolander85	11234
3	Keep up with the latest politics with The Daily Debrief panel hosted by @VassyKapelos. Today's panel features… https://t.co/IiNhYnNtqv	0	2023-03-27 21:01:06+00:00	0	3610	VassKapelosShow	4901
4	So much resemblance to NAZI Germany leading up to World War 2. Absolutely disgusting and Disgraceful #cdnpoli… https://t.co/aKB0fOzbBp	0	2023-03-27 21:01:05+00:00	0	108226	marshiehilgs	515
5	I am heading to #Budget2023 stakeholder lockup tomorrow on behalf of @actioncanadashr! International assistance, fu… https://t.co/qtpYEdQFoJ	1	2023-03-27 21:01:03+00:00	0	3621	kbowds	839

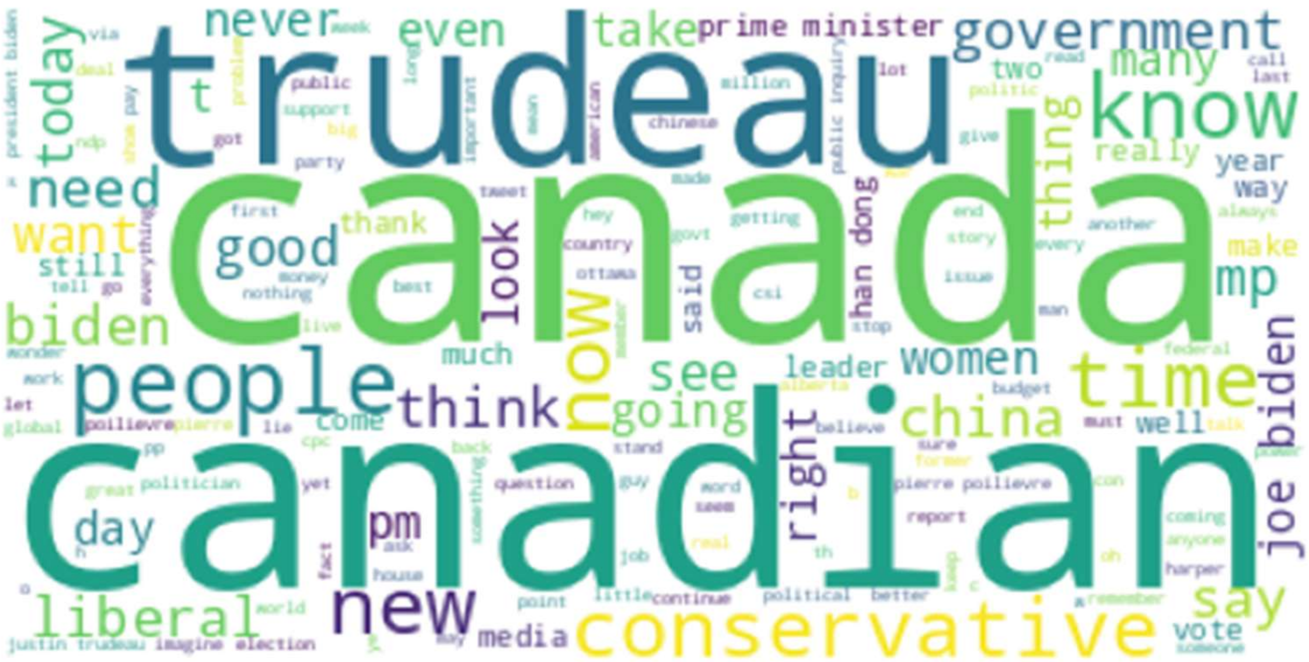
Tweet Pre-Processing

- Expand **contractions**
    - e.g. don't (do not)
- Remove **mentions** (@) and **tags** (#)
- Tokenize the tweet (**Tweet Tokenizer**)
- Remove **URLs**, **punctuation** (non-alphanumeric and whitespaces), and **stopwords** (e.g. 'a', 'is', 'the')
- Convert tokens into **lowercase**
- **Lemmatize** tokens into their base forms:
    - preserve the semantic meaning of the words

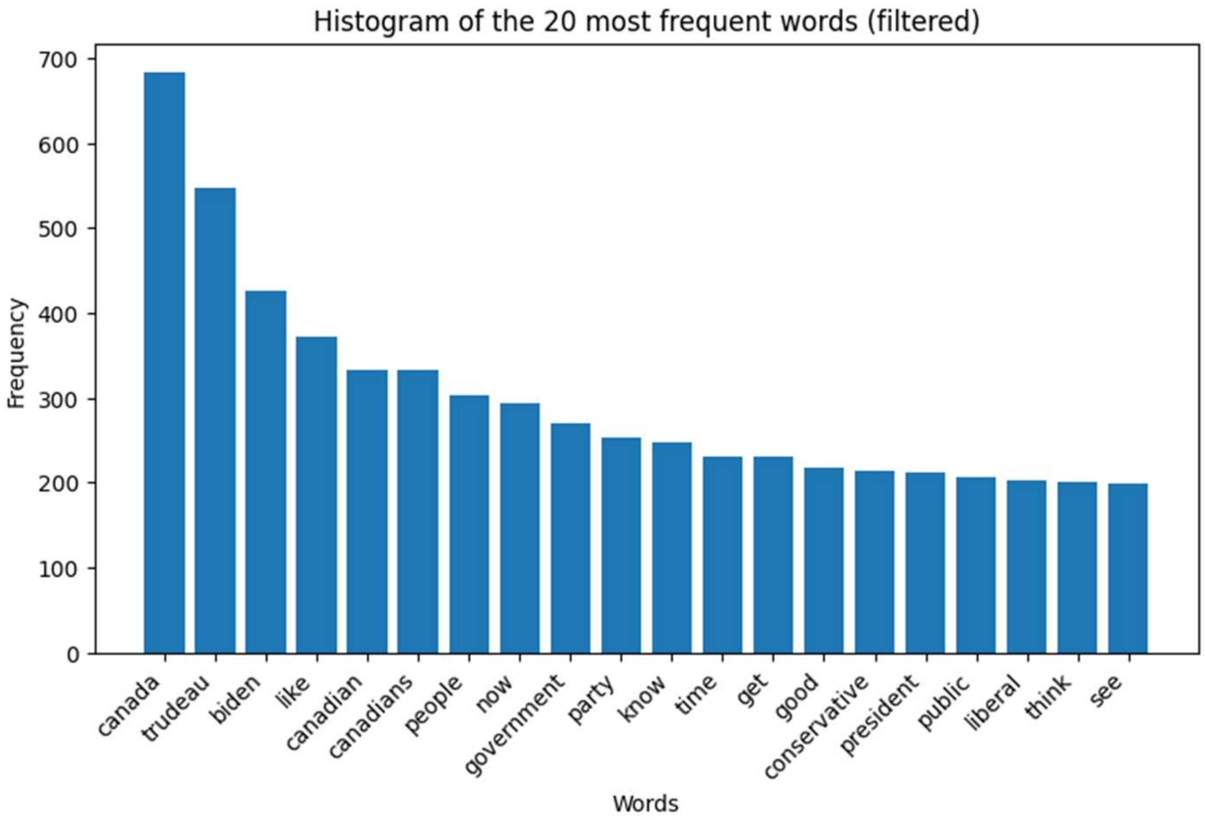


Message	Tokens	lemmatized_tokens
Keep up with the latest politics with The Daily Debrief panel hosted by Today's panel features… https://t.co/liNhYnNtqv	keep,latest,politics,daily,debrief,panel,hosted,todays,panel,features	keep,latest,politics,daily,debrief,panel,hosted,today,panel,feature

Exploratory Data Analysis (WordCloud)

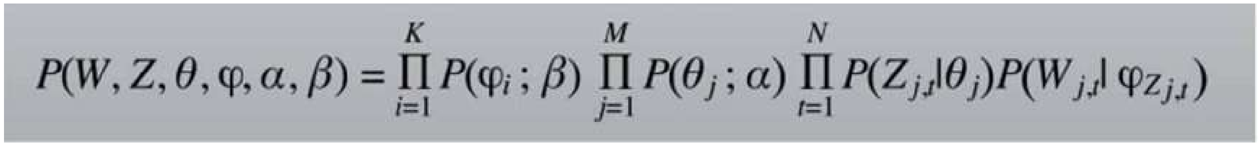


Exploratory Data Analysis (Word Frequency)

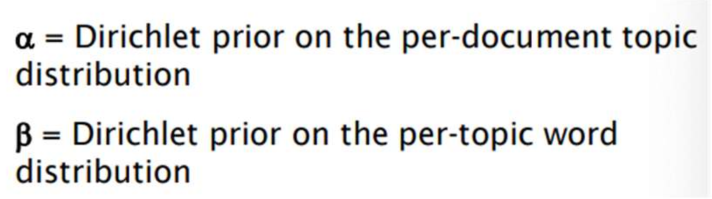


What is Latent Dirichlet Allocation (LDA)?

- An unsupervised clustering model to uncover latent topics within a collection of documents (*tweets*)
- **Assumption**: "each document is made up of various words, and each topic also has various words belonging to it"
- **How does LDA works?**
  - o The prior distribution of the words that belong to a tweet
  - o The probability of words belonging to a topic



$$P(W, Z, \theta, \varphi, \alpha, \beta) = \prod_{i=1}^{K} P(\varphi_i ; \beta) \prod_{j=1}^{M} P(\theta_j ; \alpha) \prod_{t=1}^{N} P(Z_{j,t} | \theta_j) P(W_{j,t} | \varphi_{Z_{j,t}})$$

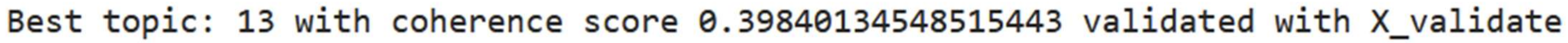


α = Dirichlet prior on the per-document topic distribution

β = Dirichlet prior on the per-topic word distribution

Building Model: Tune Hyperparameters

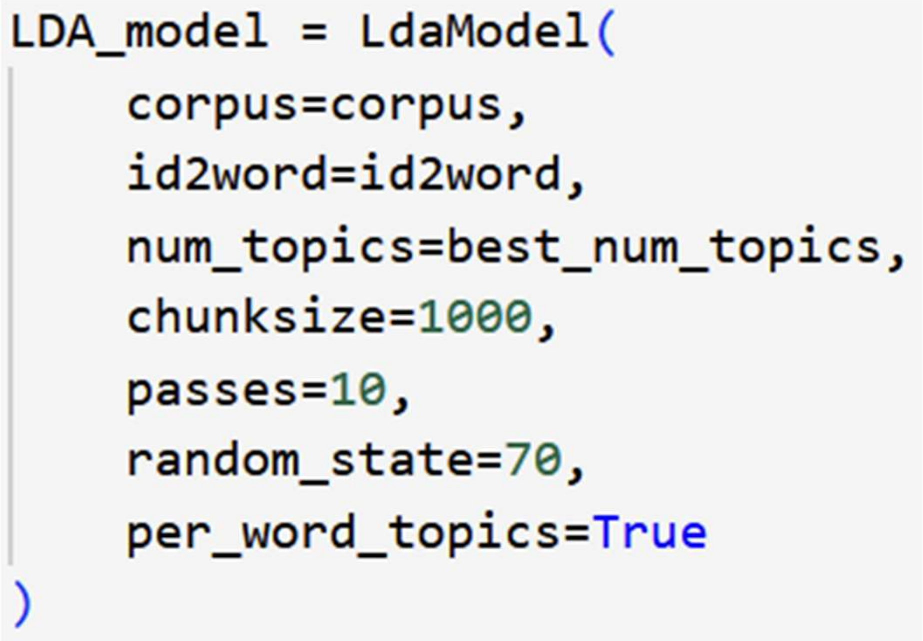
- Hyperparameters: α, β, num_topics
- Split **training** (60%), **validate** (20%), and **test** dataset (20%)
- For loop (validate dataset) to choose `num_topics`
- Aim for high coherence score



```
Best topic: 13 with coherence score 0.39840134548515443 validated with X_validate
```

Building Model: Training the Model

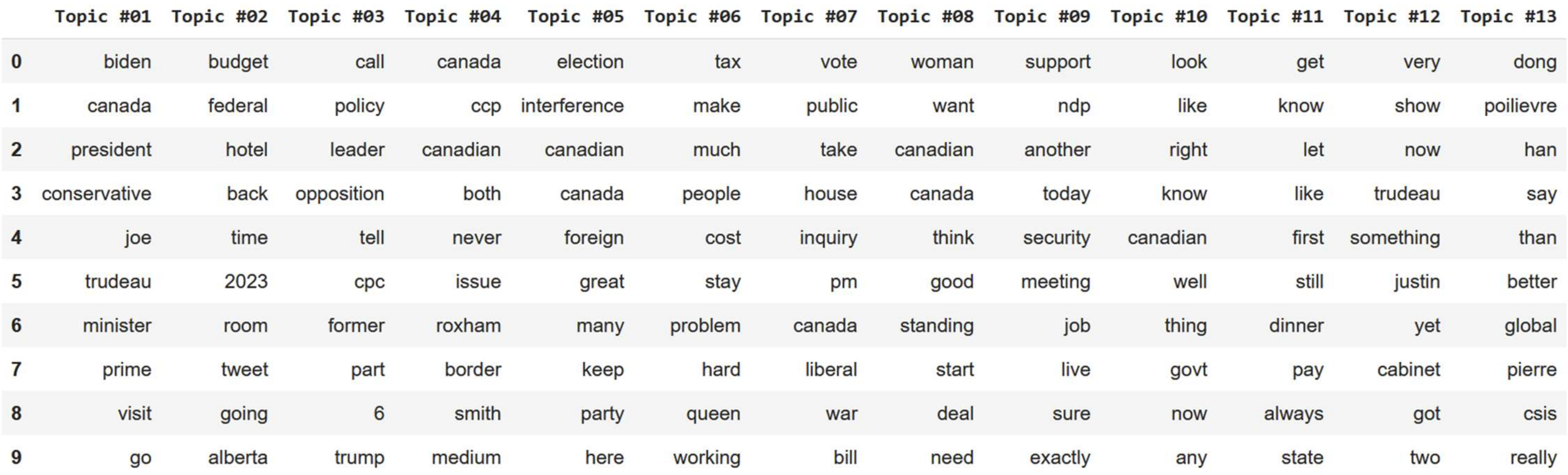
- Default values for α and β
- Train the model
  with *best_num_topics* = 13



```
LDA_model = LdaModel(
    corpus=corpus,
    id2word=id2word,
    num_topics=best_num_topics,
    chunksize=1000,
    passes=10,
    random_state=70,
    per_word_topics=True
)
```

Building Model: LDA Output

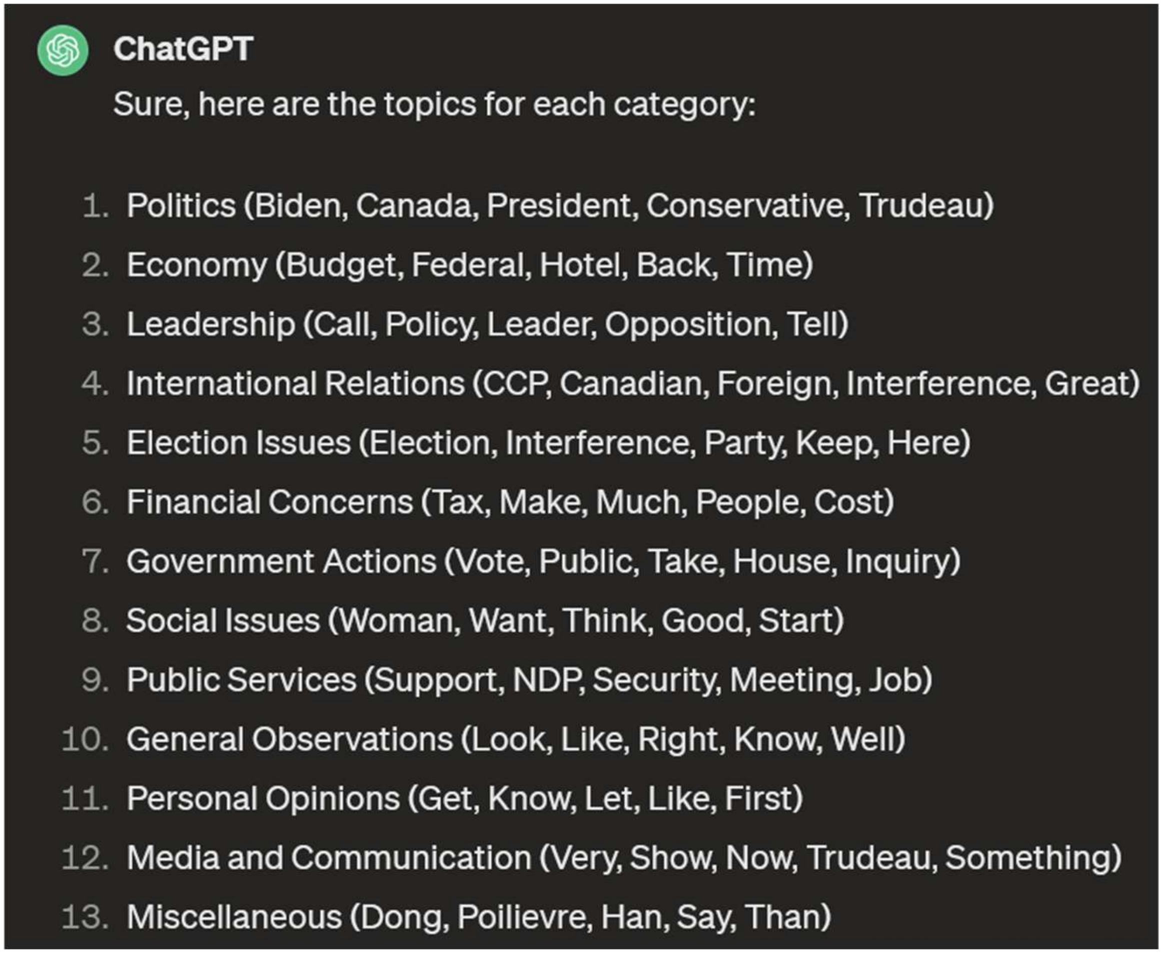
Generate 13 **topics** with 10 **words** each



	Topic #01	Topic #02	Topic #03	Topic #04	Topic #05	Topic #06	Topic #07	Topic #08	Topic #09	Topic #10	Topic #11	Topic #12	Topic #13
0	biden	budget	call	canada	election	tax	vote	woman	support	look	get	very	dong
1	canada	federal	policy	ccp	interference	make	public	want	ndp	like	know	show	poilievre
2	president	hotel	leader	canadian	canadian	much	take	canadian	another	right	let	now	han
3	conservative	back	opposition	both	canada	people	house	canada	today	know	like	trudeau	say
4	joe	time	tell	never	foreign	cost	inquiry	think	security	canadian	first	something	than
5	trudeau	2023	cpc	issue	great	stay	pm	good	meeting	well	still	justin	better
6	minister	room	former	roxham	many	problem	canada	standing	job	thing	dinner	yet	global
7	prime	tweet	part	border	keep	hard	liberal	start	live	govt	pay	cabinet	pierre
8	visit	going	6	smith	party	queen	war	deal	sure	now	always	got	csis
9	go	alberta	trump	medium	here	working	bill	need	exactly	any	state	two	really

Building Model: Labelling

Utilize **Generative AI (ChatGPT)** to label each topic



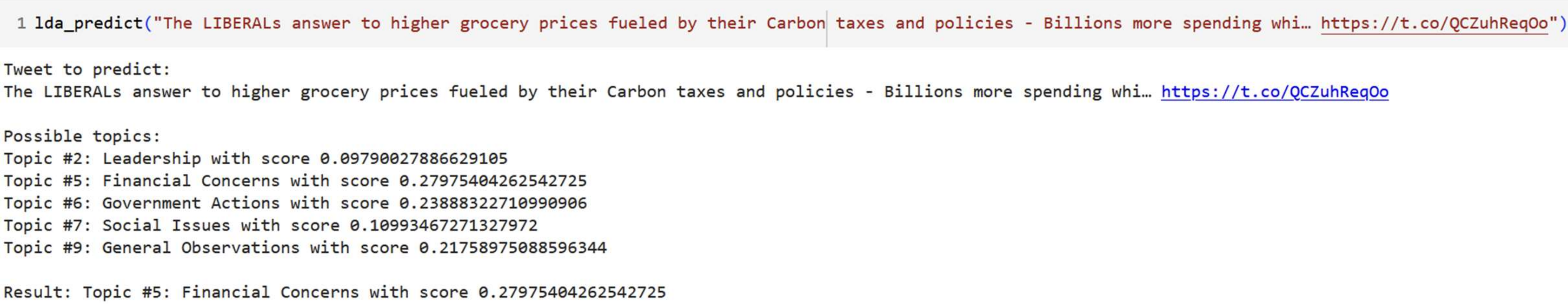
ChatGPT

Sure, here are the topics for each category:

1. Politics (Biden, Canada, President, Conservative, Trudeau)
2. Economy (Budget, Federal, Hotel, Back, Time)
3. Leadership (Call, Policy, Leader, Opposition, Tell)
4. International Relations (CCP, Canadian, Foreign, Interference, Great)
5. Election Issues (Election, Interference, Party, Keep, Here)
6. Financial Concerns (Tax, Make, Much, People, Cost)
7. Government Actions (Vote, Public, Take, House, Inquiry)
8. Social Issues (Woman, Want, Think, Good, Start)
9. Public Services (Support, NDP, Security, Meeting, Job)
10. General Observations (Look, Like, Right, Know, Well)
11. Personal Opinions (Get, Know, Let, Like, First)
12. Media and Communication (Very, Show, Now, Trudeau, Something)
13. Miscellaneous (Dong, Poilievre, Han, Say, Than)

Result: Topic Generator

Choose the topic with largest score



```
lda_predict("The LIBERALs answer to higher grocery prices fueled by their Carbon taxes and policies - Billions more spending whi… https://t.co/QCZuhReqOo")
```

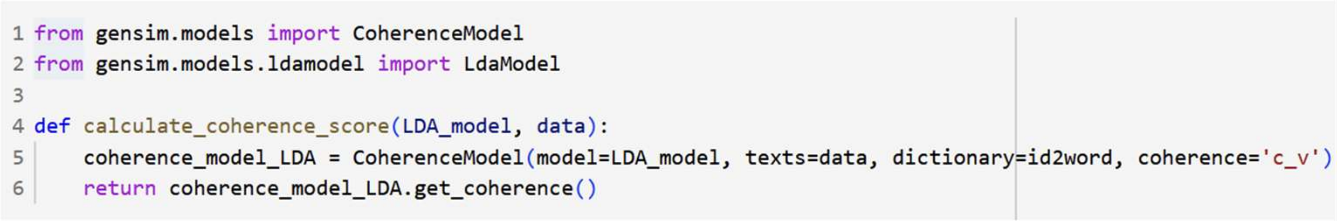
```
Tweet to predict:
The LIBERALs answer to higher grocery prices fueled by their Carbon taxes and policies - Billions more spending whi… https://t.co/QCZuhReqOo

Possible topics:
Topic #2: Leadership with score 0.09790027886629105
Topic #5: Financial Concerns with score 0.27975404262542725
Topic #6: Government Actions with score 0.23888322710990906
Topic #7: Social Issues with score 0.10993467271327972
Topic #9: General Observations with score 0.21758975088596344

Result: Topic #5: Financial Concerns with score 0.27975404262542725
```

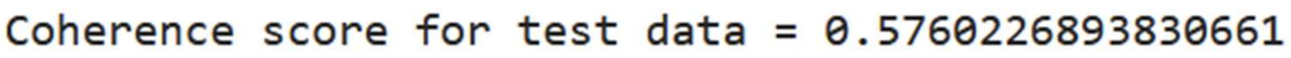
ML Model Evaluation

Coherence Test



```python
from gensim.models import CoherenceModel
from gensim.models.ldamodel import LdaModel

def calculate_coherence_score(LDA_model, data):
    coherence_model_LDA = CoherenceModel(model=LDA_model, texts=data, dictionary=id2word, coherence='c_v')
    return coherence_model_LDA.get_coherence()
```



```
Coherence score for test data = 0.5760226893830661
```

Conclusion

The model generate a topic for a tweet

Significance and Usage:

Fast generation
Labelling for supervised ML
Semantic meaning for clustering

Limitation and potential improvement:

Irrelevant topics for short tweets
Topics generated are limited to training data
Limited to unigram interpretation

THANK YOU