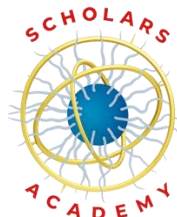
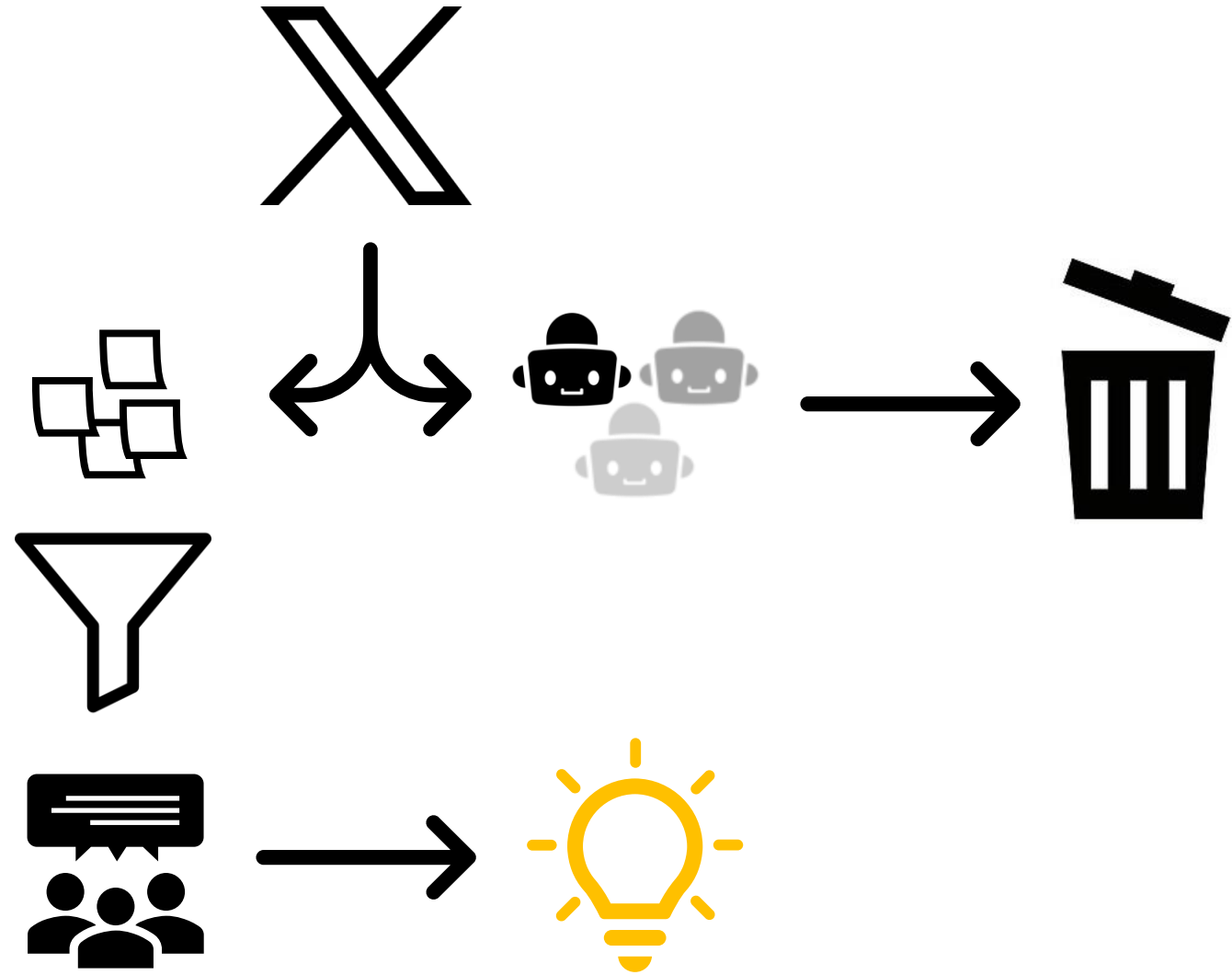


Designing a reproducible NLP workflow for social media sentiment and topic modeling

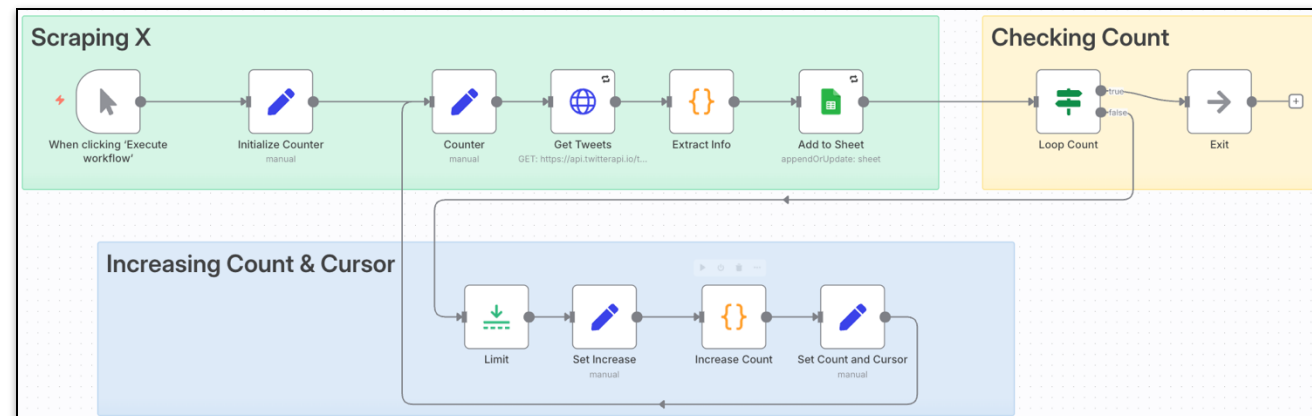
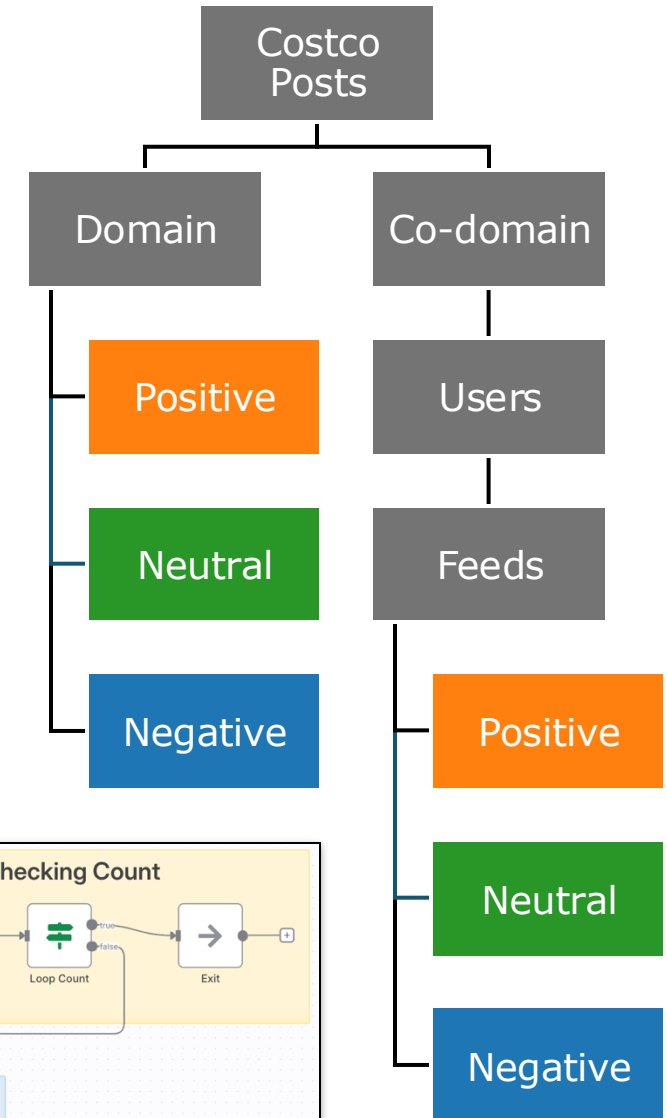
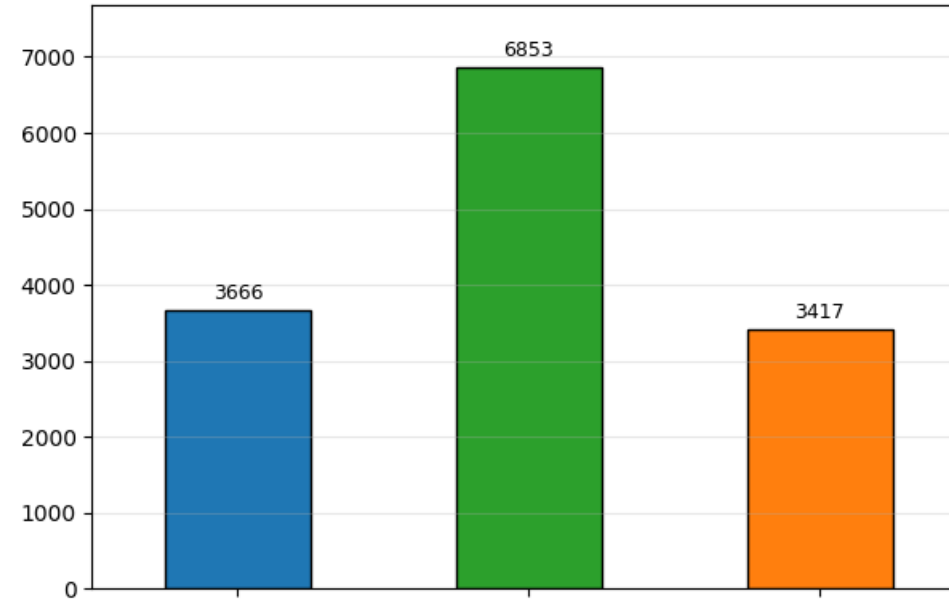
Thomas Linden, Katherine Shoemaker, PhD



- Social media \approx the world's biggest focus group.
- But the signal is buried.
- How can we extract this information?
- Our answer: a reproducible NLP workflow.
- Question: does co-domain context help or dilute insights?



- ~60K posts via twitterv2.io + n8n (automated pulls).
- Three entities: NFL, UK PM (Keir Starmer), Costco.
- 5-day collection windows; **domain** vs **co-domain** sampling.
- Code, figures, and outputs → GitHub.



- Standard cleaning; RoBERTa sentiment (pos/neu/neg)
- Key terms via two TF-IDF views
 - Aggregate = sum across documents
 - Sentiment-weighted = coverage x rarity

$$TFIDF_{t,d} = tf_{t,d} \times \log \left(\frac{N}{|\{d \in D : t \in d\}|} \right)$$

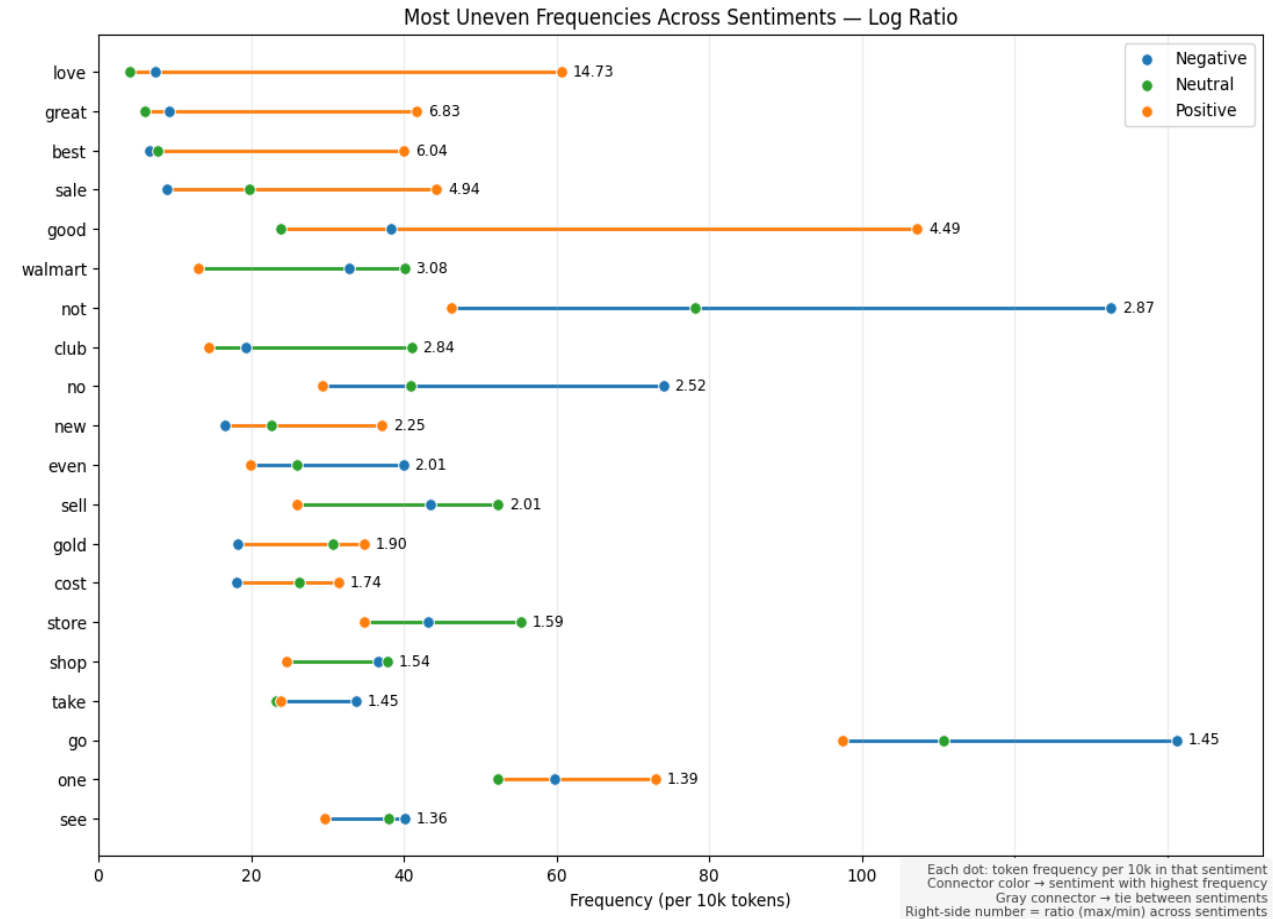
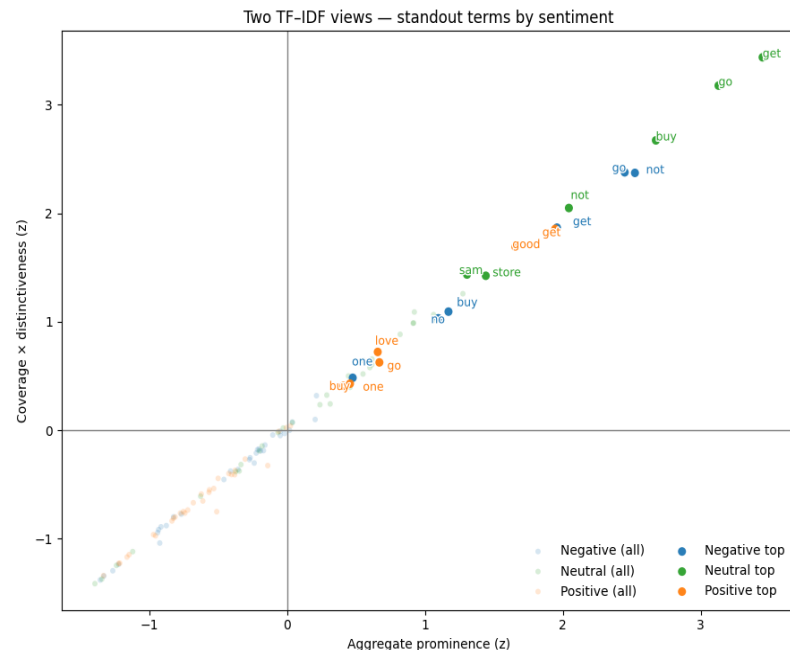
t = term

d = document

tf = frequency of term t in document d

N = total number of documents

D = number of documents d containing term t



Content	@Ronxyz00 Case of coconut water was 10lastweekatCostco. Todaywas12.79
Content_clean	case of coconut water was 10 last week at costco today was 12 79
text_for_sent	@USER Case of coconut water was 10lastweekatCostco. Todaywas12.79
tokens	case of coconut water was last week at costco today was
lemmas	case of coconut water be last week at costco today be
no_stop	case coconut water last week costco

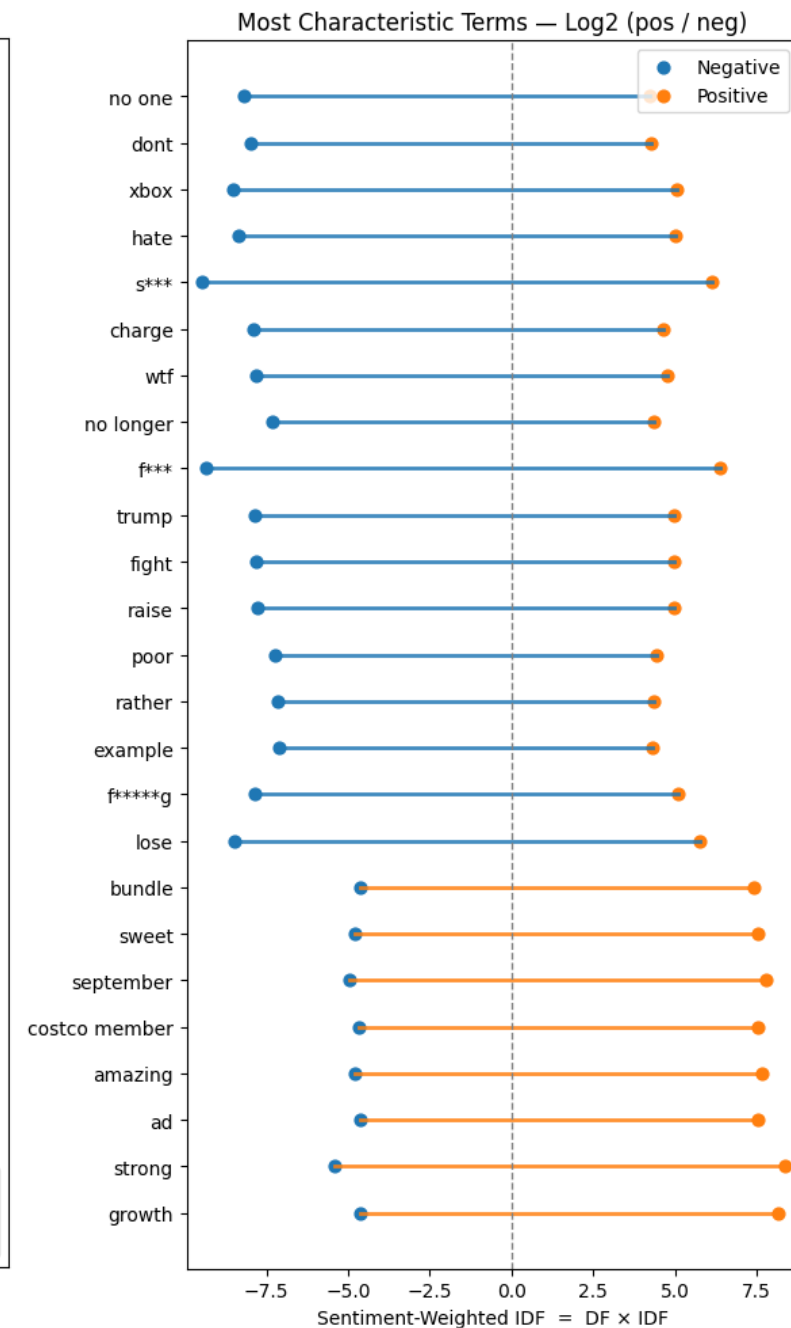
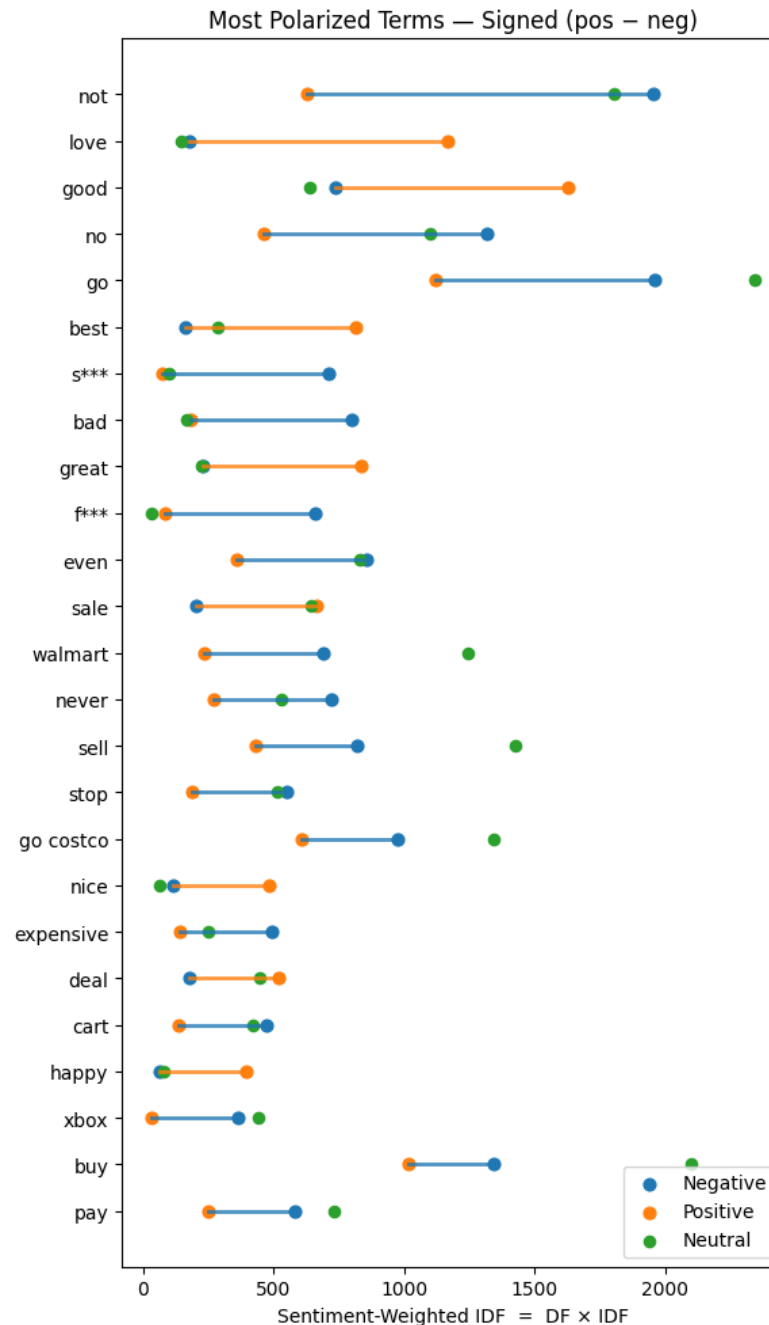
- Additional insights when adjusting n-grams parameter.

- Positive**


- fsd v14
- costco deal
- caramel brownie sundae
- costco pumpkin pie
- costco gold
- costco gold bar

- Negative**

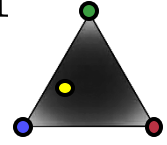
- park lot
- line costco gas
- costco sam
- costco walmart
- costco silver
- costco silver bar




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



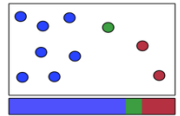
Document



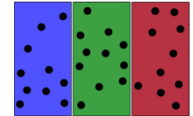
Document



Word



Topic



Word

Topics - Dirichlet Distributions

Sampling - Multinomial Distributions

\mathbf{W} Identity of word w in document d

\mathbf{Z} Identity of topic of word w in document d

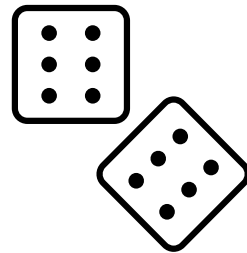
$\boldsymbol{\theta}$ Probability of topic k occurring in document d

$\boldsymbol{\varphi}$ Probability of word w occurring in topic k

α Prior weight of topic k in a document

β Prior weight of word w in a topic

Latent Dirichlet Allocation (LDA)



Shakespeare's *Othello*

RODERIGO: Tush, never tell me, I take it much **unkindly** ←

That thou, Iago, who hast had my **purse** ←

As if the strings were thine, should know of this.

IAGO: 'Sblood, but you'll not hear me. If ever I did dream

Of such a matter, **abhor** me. ←

RODERIGO: Thou told'st me

Thou didst hold him in thy **hate**. ←

IAGO: **Despise** me ←

If I do not. Three great ones of the city,

In personal suit to make me his **lieutenant**, ←

Off-capped to him; and, by the faith of man,

I know my **price**; I am **worth** no worse a place. ←

But he, as loving his own pride and purposes,

Evades them with a bombast circumstance,

Horribly stuffed with epithets of **war**, ←

And in conclusion,

Nonsuits my mediators. For "Certes," says he,

"I have already chose my **officer**." ←

And what was he?

Forsooth, a great arithmetician,

One Michael Cassio, a Florentine,

A fellow almost **damned** in a fair wife,

That never set a **squadron** in the field, ←

Nor the division of a **battle** knows ←

Topics

negative	0.04	money	0.04	military	0.04
hate	0.02	worth	0.02	conflict	0.02
bad	0.01	purse	0.01	officer	0.02

- Hot dog price/changes as shorthand for “is Costco still on our side?” and overall value perception.



Costco hot dog as value symbol

- Repeated focus on price vs. quality/safety of staples (rotisserie chicken, meat, etc.).
- Reliability and transparency here act as trust levers.



Price and quality concerns (esp. chicken/meat)

- Cross-shopping Costco, Trader Joe's, BJ's, etc.
- Parents weighing bulk value, convenience, and “healthy enough” options for kids.



Store choice & family/bulk tradeoffs

- Interest in halal/healthier options, protein sources, specialty goods.
- Signals whether Costco aligns with specific dietary, cultural, or lifestyle needs.



Niche products, ethics, and lifestyle

- Joining/renewing justified by savings on batteries, tools, pantry staples, etc.
- Small frictions in staple items can tip sentiment on overall membership value.



Membership ROI, everyday essentials

- Direct comparisons on stock, electronics, and service.
- Perception is shaped by concrete outcomes (what's on shelves, issue resolution).



Costco vs. Sam's performance

What worked

- Reproducible workflow
- Domain-first sampling
- LLM summarization

Use with care

- Co-domain corpus metrics
- n-gram interpretation
- Alternative document sources (call logs, etc.)

Next steps

- Extend snapshot to time series
- Explore multimodal (images)
- Compare LDA to graph-based topic methods

References

Herk, N. (2025, March 14). *How to Actually Scrape Twitter/X Data with n8n* [Video]. YouTube. <https://www.youtube.com/watch?v=IEo7IAgj0UY>

Starmer, J. (2020, March 18). *Latent Dirichlet Allocation* [Video]. YouTube. <https://youtu.be/T05t-SqKArY?si=ZxP2DZt9vDicqIsc>

Acknowledgements

Minority Science and Engineering Improvement Program (MSEIP) Grant number: P120A220015, "Boosting STEM Student Success", Dr. Bernadette Hence at U.S. Department of Education, and University of Houston-Downtown.

We thank Dr. Mary Parker, PI Project Director, MSEIP..

