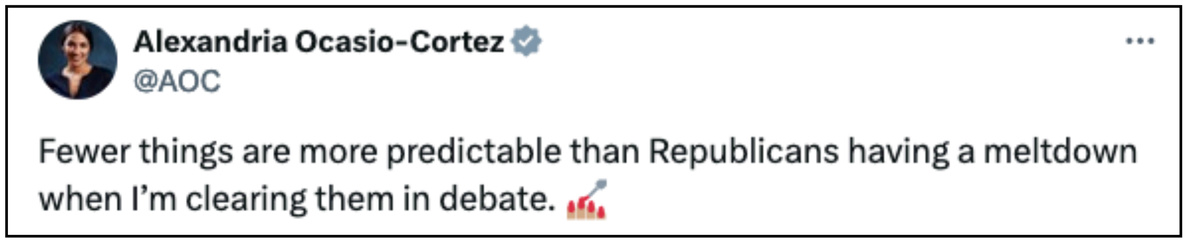
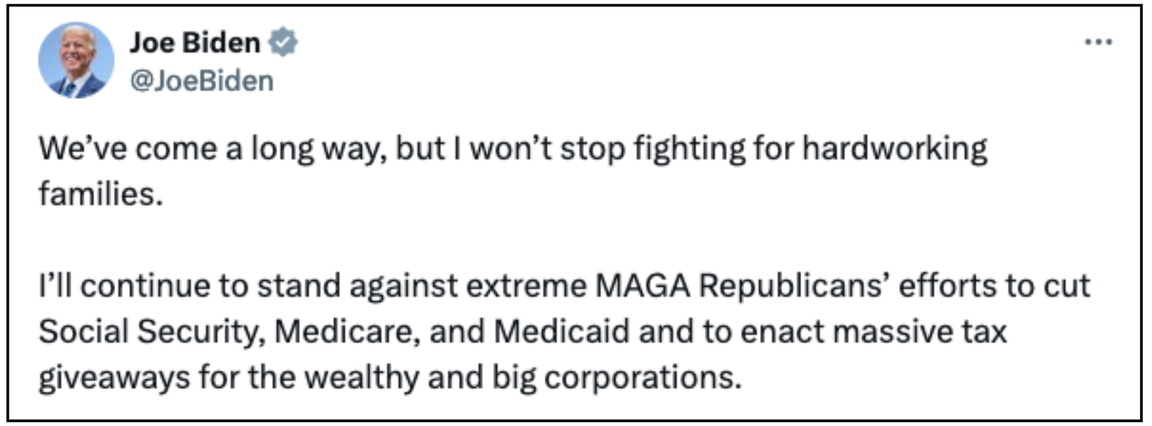
PROFESSOR BENJAMIN NOBLE (UCSD)

PULLING WORDS OUT OF A BAG

SO YOU GOT SOME TEXT…NOW WHAT?





PULLING WORDS OUT OF A BAG

```
lm(donation ~                                , df)
```

Marjorie Taylor Greene
@mtgreenee

We need President Trump back in office to unleash American energy dominance and end Joe Biden and the Democrats quest to plunge America into darkness.

Alexandria Ocasio-Cortez
@AOC

Fewer things are more predictable than Republicans having a meltdown when I'm clearing them in debate.

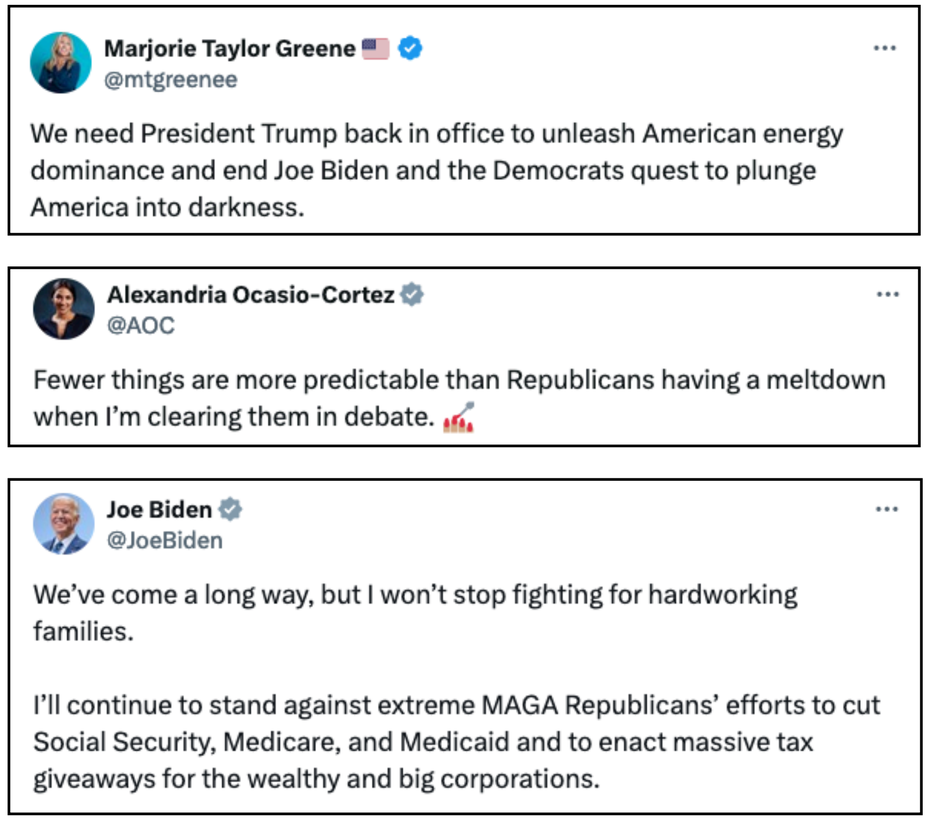
Joe Biden
@JoeBiden

We've come a long way, but I won't stop fighting for hardworking families.

I'll continue to stand against extreme MAGA Republicans' efforts to cut Social Security, Medicare, and Medicaid and to enact massive tax giveaways for the wealthy and big corporations.

THE BAG OF WORDS MODEL

- Each **document** in a **corpus** is represented by a count of the **features** that appear within it.
  - What is a corpus?
  - What is a document?
  - What is a feature?
- Create a **document feature matrix.**



PULLING WORDS OUT OF A BAG

document feature matrix

doc_id	we	need	president	trump	back	in	office	to	unleash	american	energy
1 text1	1	1	1	1	1	1	1	2	1	1	1
2 text2	0	0	0	0	0	1	0	0	0	0	0
3 text3	0	0	0	0	0	0	0	3	0	0	0

	dominance	and	end	joe	biden	the	democrats	quest	plunge	america	into
1	1	2	1	1	1	1	1	1	1	1	1
2	0	0	0	0	0	0	0	0	0	0	0
3	0	3	0	0	0	1	0	0	0	0	0

	darkness	.	fewer	things	are	more	predictable	than	republicans	having	a
1	1	1	0	0	0	0	0	0	0	0	0
2	0	1	1	1	1	1	1	1	1	1	1
3	0	2	0	0	0	0	0	0	1	0	1

	meltdown	when	i'm	clearing	them	debate	we've	come	long	way	,	but	i	won't
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	1	1	1	1	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	1	1	1	1	3	1	1	1

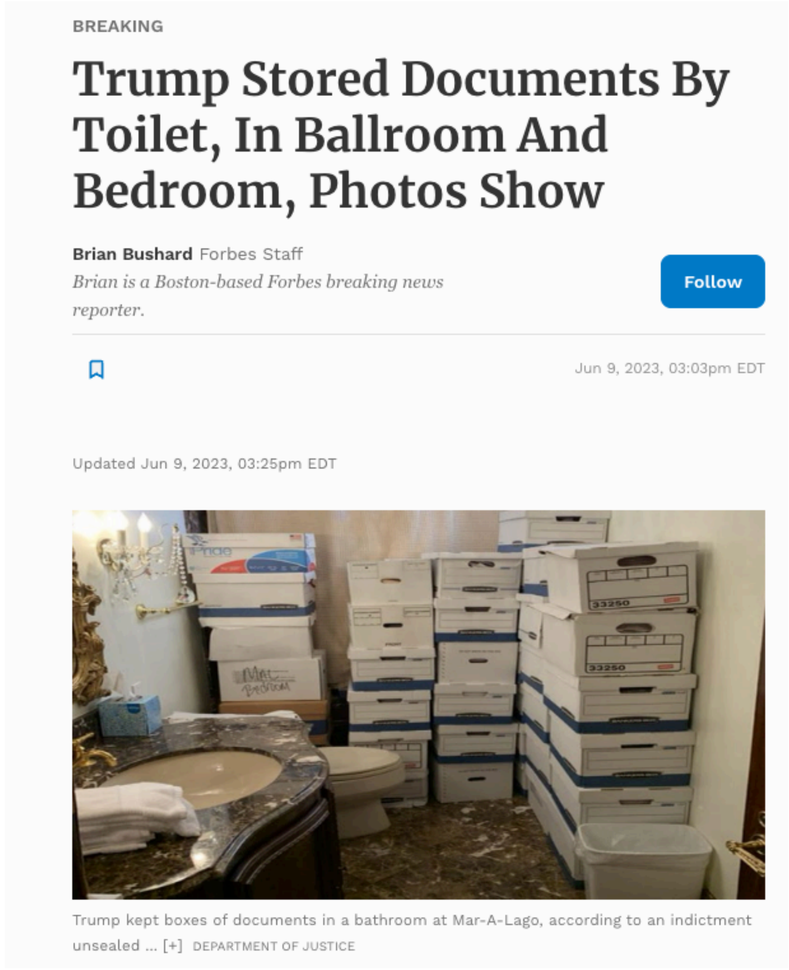
	stop	fighting	for	hardworking	families	i'll	continue	stand	against
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	1	1	2	1	1	1	1	1	1

	extreme	maga	'	efforts	cut	social	security	medicare	medicaid	enact
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0
3	1	1	1	1	1	1	1	1	1	1

	massive	tax	giveaways	wealthy	big	corporations
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	1	1	1	1	1	1

SO...WHAT IS A DOCUMENT?

- A document in your minds eye: a newspaper article, a presidential speech, a tweet.
- But documents can be anything...
  - A headline, all of the tweets sent by a person in a single day, an entire book, a paragraph of a news story, a news headline.



BREAKING

Trump Stored Documents By Toilet, In Ballroom And Bedroom, Photos Show

Brian Bushard Forbes Staff

Brian is a Boston-based Forbes breaking news reporter.

Follow

Jun 9, 2023, 03:03pm EDT

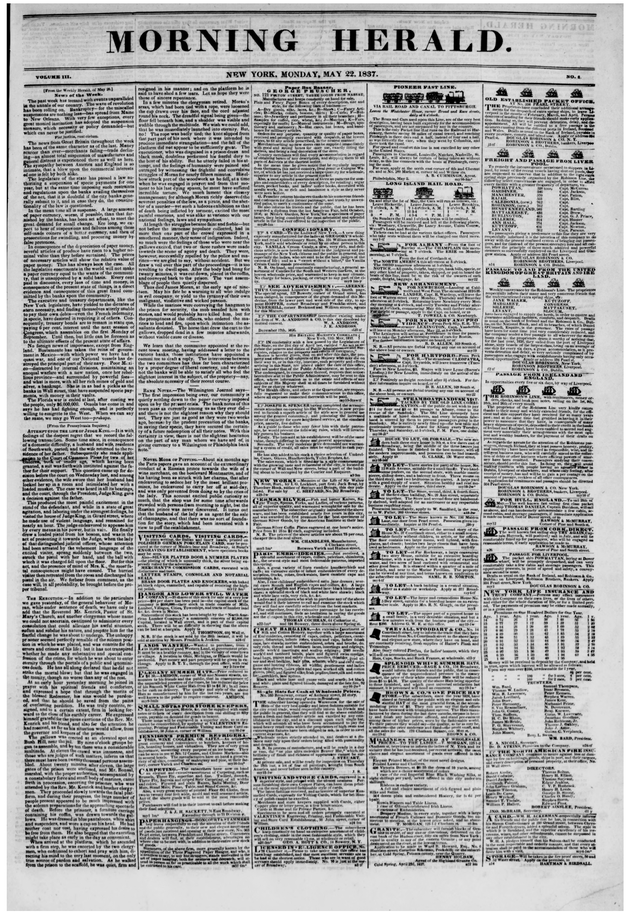
Updated Jun 9, 2023, 03:25pm EDT

Trump kept boxes of documents in a bathroom at Mar-A-Lago, according to an indictment unsealed ... [+] DEPARTMENT OF JUSTICE

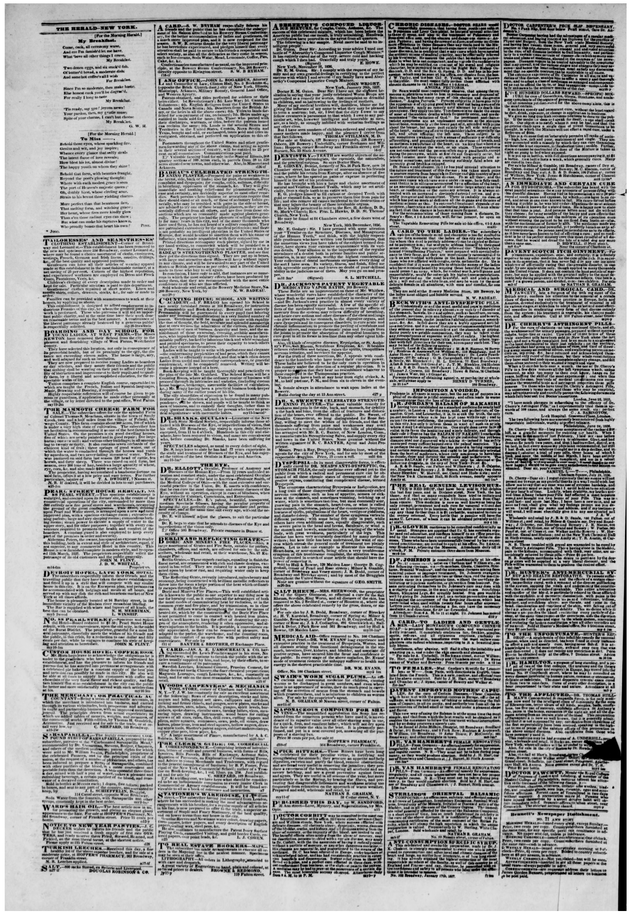
via Forbes

HOW TO CHOOSE YOUR UNIT OF ANALYSIS (WHAT CONSTITUTES A DOCUMENT)

- Logistics.


MORNING HERALD.

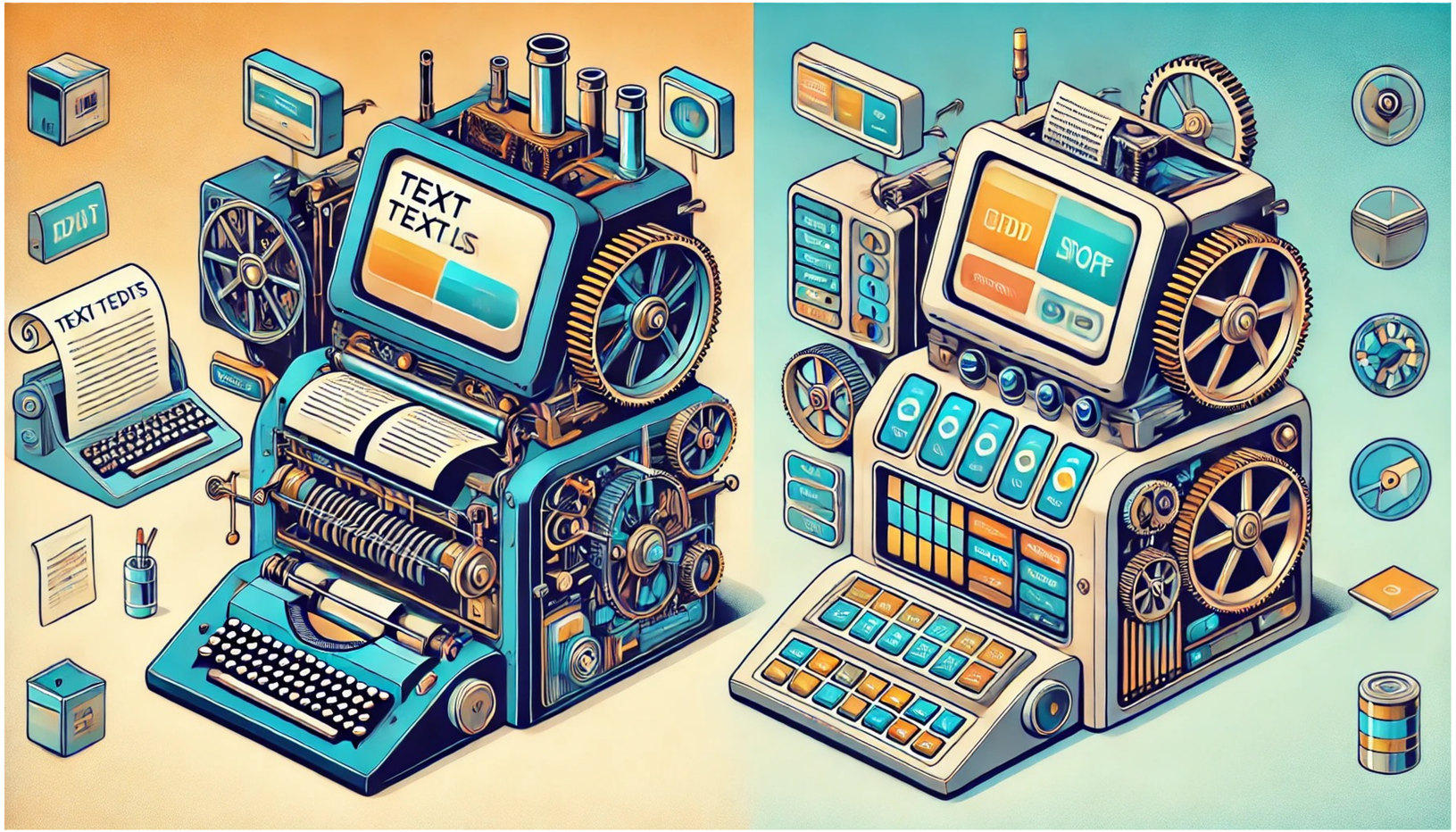
NEW YORK, MONDAY, MAY 22, 1837.



via Library of Congress, Chronicling America

HOW TO CHOOSE YOUR UNIT OF ANALYSIS (WHAT CONSTITUTES A DOCUMENT)

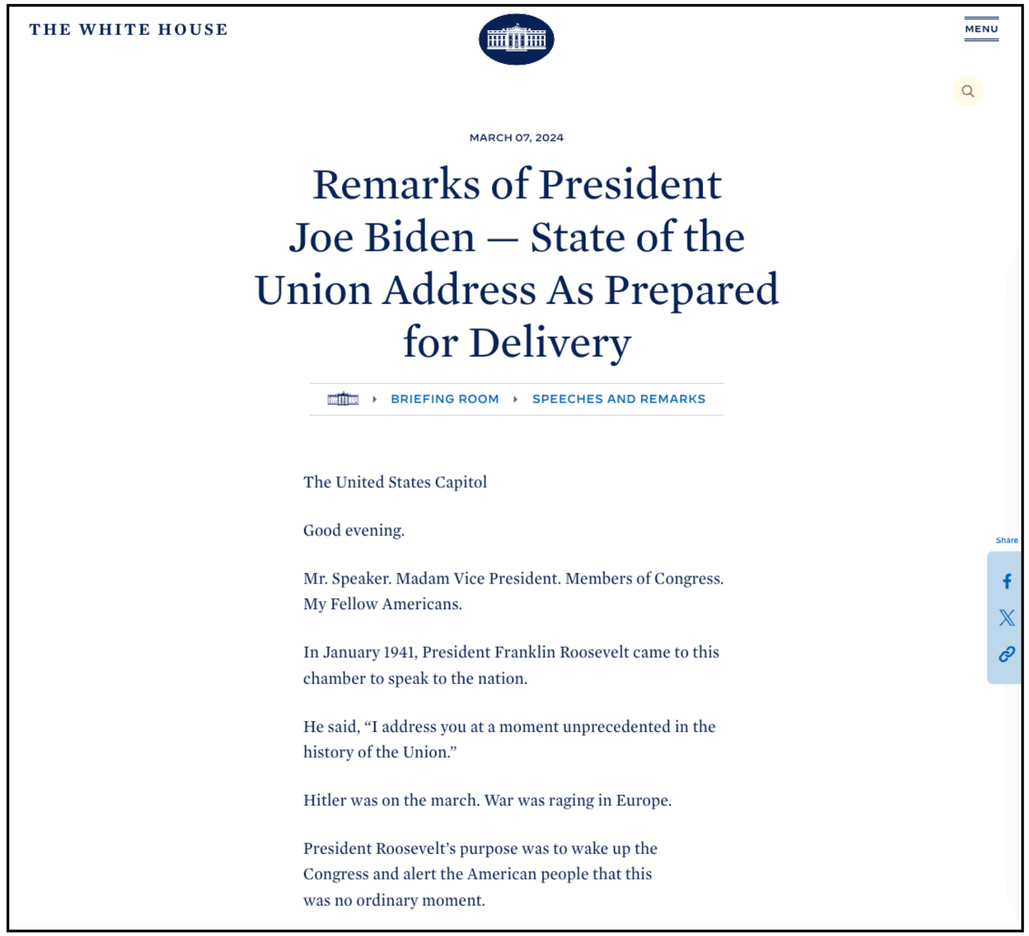
- Logistics.
- Model.



via DALL-E

HOW TO CHOOSE YOUR UNIT OF ANALYSIS (WHAT CONSTITUTES A DOCUMENT)

- Logistics.
- Model.
- Research question.


THE WHITE HOUSE

MENU

MARCH 07, 2024

Remarks of President Joe Biden — State of the Union Address As Prepared for Delivery

BRIEFING ROOM › SPEECHES AND REMARKS

The United States Capitol

Good evening.

Mr. Speaker. Madam Vice President. Members of Congress. My Fellow Americans.

In January 1941, President Franklin Roosevelt came to this chamber to speak to the nation.

He said, “I address you at a moment unprecedented in the history of the Union.”

Hitler was on the march. War was raging in Europe.

President Roosevelt’s purpose was to wake up the Congress and alert the American people that this was no ordinary moment.

Share

via The White House

FEATURES AND TOKENS

- A feature in your minds eye: a word.
  - But where are the words…?
  - Why words?

新小説豫告
燃ゆる星座
尾崎士郎氏近來の快作
來る七日から連載

暴動事件
追加起訴二件
並に警察記錄

燃ゆる星」梗概

女優

運命

現在流行

ホノルル座、
人妻椿大會
昨夜又滿員
今明晩限り

via MDPI

PULLING WORDS OUT OF A BAG

bag of...letters?

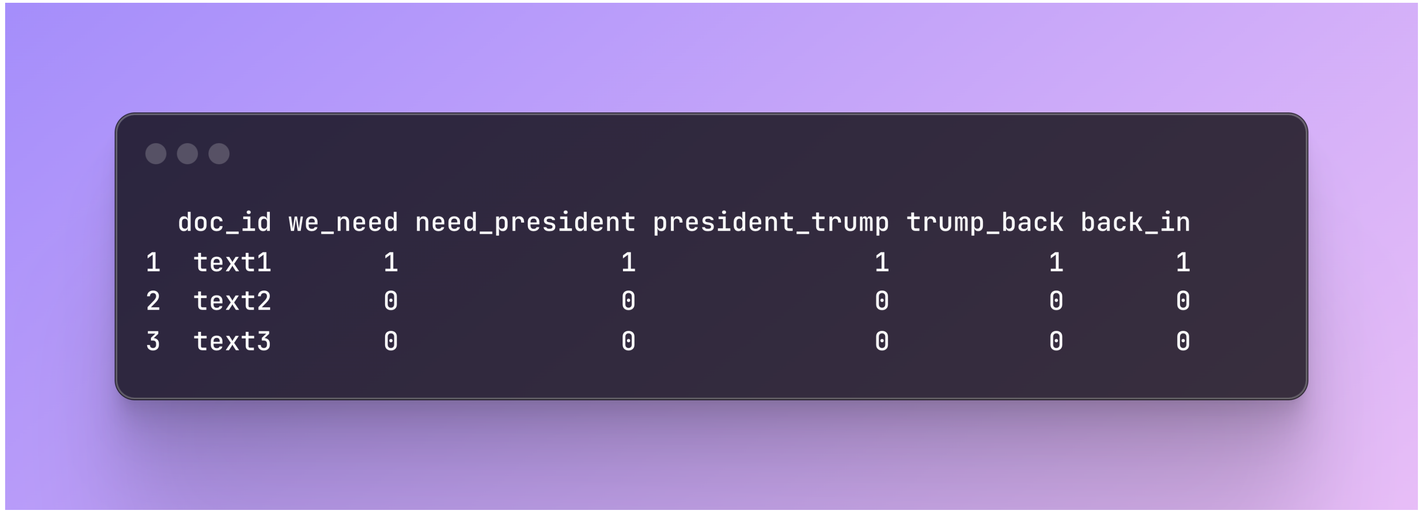
	doc_id	w	e	n	d	p	r	s	i	t	u	m	b	a	c	k	o	f	l	h	g	y	j	q	.	v	'	,	x
1	text1	1	20	15	9	3	7	6	8	8	4	5	2	11	6	2	7	2	2	2	2	1	1	1	1	0	0	0	0
2	text2	3	13	8	3	2	6	2	7	6	1	4	3	8	3	0	2	1	4	5	3	0	0	0	1	1	1	0	0
3	text3	6	19	14	8	3	11	12	19	19	5	7	3	23	10	1	16	6	7	4	8	4	0	0	2	3	4	3	2

bag of...sentences?

```
  doc_id
1  text1
2  text2
3  text3
  we need president trump back in office to unleash american energy dominance and end joe
biden and the democrats quest to plunge america into darkness.
1
1
2
0
3
0
  fewer things are more predictable than republicans having a meltdown when i'm clearing them
in debate.
1
0
2
1
3
0
  we've come a long way, but i won't stop fighting for hardworking families.
1                                                                             0
2                                                                             0
3                                                                             1
  i'll continue to stand against extreme maga republicans' efforts to cut social security,
medicare, and medicaid and to enact massive tax giveaways for the wealthy and big
corporations.
1
0
2
0
3
1
```

WHAT'S WRONG WITH WORDS?

- Some concepts bridge across words: *social security*, *White House*, *Supreme Court Justice*.
  - A solution: *n-grams*.
    - Unigrams (words), bigrams (two words), trigrams (three words), oh my...



	doc_id	we_need	need_president	president_trump	trump_back	back_in
1	text1	1	1	1	1	1
2	text2	0	0	0	0	0
3	text3	0	0	0	0	0

REDUCE COMPLEXITY

```
We need President_Trump
back in office to unleash
American energy dominance
and end Joe_Biden and the
Democrats quest to plunge
America into darkness.
```

REDUCE COMPLEXITY

- Lowercase (e.g., “We” == “we”).

```
we need president_trump
back in office to unleash
American energy dominance
and end joe_biden and the
democrats quest to plunge
america into darkness.
```

REDUCE COMPLEXITY

- Lowercase (e.g., “We” == “we”).
- Remove punctuation.

```
we need president_trump
back in office to unleash
American energy dominance
and end joe_biden and the
democrats quest to plunge
america into darkness
```

REDUCE COMPLEXITY

- Lowercase (e.g., “We” == “we”).
- Remove punctuation.
- Remove “stop words” (e.g., and, it, in).

```
need president_trump back
office unleash american
energy dominance end
joe_biden democrats quest
plunge america darkness
```

REDUCE COMPLEXITY

- Lowercase (e.g., “We” == “we”).
- Remove punctuation.
- Remove “stop words” (e.g., and, it, in).
- Create equivalence classes (e.g., stemming).

```
need president_trump back
offic unleash american
energi domin end
joe_biden democrat quest
plung america darkness
```

REDUCE COMPLEXITY

- Lowercase (e.g., “We” == “we”).
- Remove punctuation.
- Remove “stop words” (e.g., and, it, in).
- Create equivalence classes (e.g., stemming).
- Filter rare words.

```
need president_trump back
offic unleash american
energi domin end
joe_biden democrat quest
plung america darkness
```

```
Document-feature matrix of: 241 documents, 6,358 features (81.42% sparse) and 6 docvars.
                  features
docs               fellow-citizen senat hous repres embrac great satisfact opportun now
present
  Washington-1790               1     2    3      3      1     4         2        1   1
  Washington-1790b              1     2    3      3      0     4         1        0   2
  Washington-1791               1     3    3      3      1     0         3        1   0
  Washington-1792               1     2    3      4      1     0         3        0   2
  Washington-1793               1     2    3      3      1     0         0        1   2
  Washington-1794               1     2    5      4      1     1         0        1   3
[ reached max_ndoc ... 235 more documents, reached max_nfeat ... 6,348 more features ]
```

GETTING MORE TECHNICAL

- A document feature matrix **W**...
  - Has **N** documents (or rows) indexed by i.
  - Has **J** features in the vocabulary (or columns) indexed by j.
  - Thus, **W** is an **N x J** matrix where W_ij is a cell which contains a count of that feature in that document.
  - So the State of the Union Address corpus (**W**) is a 241 (**N**) x 6,358 (**J**) matrix. W_3,4 is 3 (it contains the word “repres” which appears 3 times in speech number 3.

```
lm(donation ~ 
```

```
Document-feature matrix of: 241 documents, 6,358 features (81.42% sparse) and 6 docvars.
                 features
docs              fellow-citizen senat hous repres embrac great satisfact opportun now
present
  Washington-1790              1     2    3      3      1     4         2        1   1
  Washington-1790b             1     2    3      3      0     4         1        0   2
  Washington-1791              1     3    3      3      1     0         3        1   0
  Washington-1792              1     2    3      4      1     0         3        0   2
  Washington-1793              1     2    3      3      1     0         0        1   2
  Washington-1794              1     2    5      4      1     1         0        1   3
[ reached max_ndoc ... 235 more documents, reached max_nfeat ... 6,348 more features ]
```

```
, df)
```

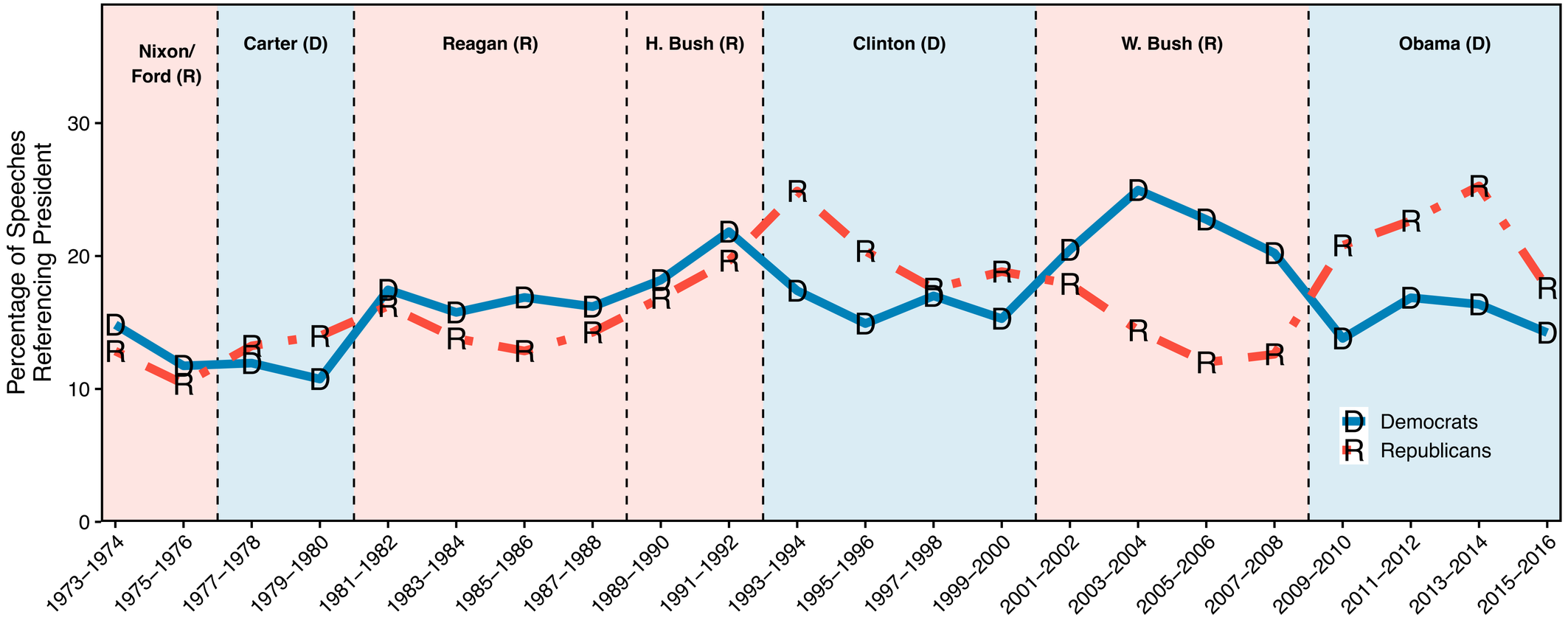
WHAT DO WE DO WITH W?

- Our goal in text analysis is to try to learn the relationship between some variable **Q** and a representation of the text **W**.
  - Which texts are negative? Some mapping function of **W** (e.g., dfm) to **Q** (0 = positive, 1 = negative).
  - Then, we can use **Q** for subsequent analysis (e.g., `donations ~ Q`).

KEYWORD COUNTING

- Ideal mapping: you read every document and apply a sophisticated coding rule.
- Example: is a document about the U.S. President?
  - Starting heuristic: does the document use the word “president?”
    - False positives: President of Syracuse.
    - False negatives: The White House.
- **Q:** is `W_i(j=president) > 0`?

PULLING WORDS OUT OF A BAG



Noble (2024)

MORE WITH DICTIONARIES

- We can measure more complicated concepts like sentiment.
- $\pi_i = \sum_{j=1}^{J} \frac{\mu_j W_{ij}}{M_i}$ where μ_j is a feature-specific weight and $M_i = \sum_{j=1}^{J} W_{ij}$.

bing dictionary

	word	sentiment
	<chr>	<chr>
1	exemplary	positive
2	suppress	negative
3	capitulate	negative
4	ridicules	negative
5	immobilized	negative
6	effusive	positive
7	entrap	negative
8	degrade	negative
9	foreboding	negative
10	barbarity	negative

MORE WITH DICTIONARIES

- We can measure more complicated concepts like sentiment.
- $\pi_i = \sum_{j=1}^{J} \frac{\mu_j W_{ij}}{M_i}$ where μ_j is a feature-specific weight and $M_i = \sum_{j=1}^{J} W_{ij}$.

afinn dictionary

	word	value
	<chr>	<dbl>
1	injury	-2
2	fail	-2
3	pity	-2
4	degraded	-2
5	improves	2
6	keen	1
7	unmatched	1
8	disrespected	-2
9	killed	-3
10	fondness	2

PULLING WORDS OUT OF A BAG

afinn-scores

```
# A tibble: 5 × 3
  doc_id word     value
   <int> <chr>    <dbl>
1      1 darkness    -1
2      3 stop        -1
3      3 cut         -1
4      3 wealthy      2
5      3 big          1
```

PULLING WORDS OUT OF A BAG

afinn-scores

	doc_id	score
	<int>	<dbl>
1	1	-0.04
2	2	0
3	3	0.0244