NLP!!!

April 7, 2020

Data Science CSCI 1951A

Brown University

Instructor: Ellie Pavlick

HTAs: Josh Levin, Diane Mutako, Sol Zitter

Announcements

- S/NC Option
- “Special Topics”
- Questions/Concerns?

Today

- “1990s NLP”…i.e. counting words :)
  - Bags-of-words, Preprocessing
  - “Tools for working with text”
- No Machine Learning today
  - More on Thursday…

Resources

- Tokenization, Tagging, Parsing, all sorts of fancy things
- NLTK: https://www.nltk.org/
- Spacy: https://spacy.io/

Ways you might use NLP

Ways you might use NLP

- You want to use text as a feature for some prediction task
    - Classify sentiment in twitter, predict popularity of posts, track spread of articles/ideas across the country

Ways you might use NLP

- You want to use text as a feature for some prediction task
    - Classify sentiment in twitter, predict popularity of posts, track spread of articles/ideas across the country
- You want to make predictions/test hypotheses about language itself
    - Model changes in word use over time/across locations, find words that cause articles to be shared

Ways you might use NLP

- You want to use text as a feature for some prediction task
    - Classify sentiment in twitter, predict popularity of posts, track spread of articles/ideas across the country
- You want to make predictions/test hypotheses about language itself
    - Model changes in word use over time/across locations, find words that cause articles to be shared
- Clustering of text data
    - In either of the above use cases
    - Are these words similar, is this document similar to this query, are these documents similar to each other, etc…

Unit of analysis

- Characters (“s” “w” “i” “m” “m” “i” “n” “g” “l” “y”)
- Morphemes (“swim” “ing” “ly")
- Words (“swimmingly”)
- Sentences (“remote instruction is going swimmingly”)
- Documents (“Remote instruction is going swimmingly. Yesterday, for example, a student said…”)

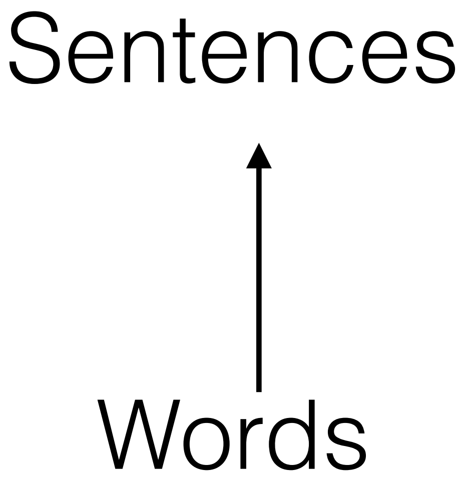
Compositionality

"meaning of the whole is a function of a meaning of the parts and the way in which they are combined"

Compositionality

Words

Compositionality



Sentences

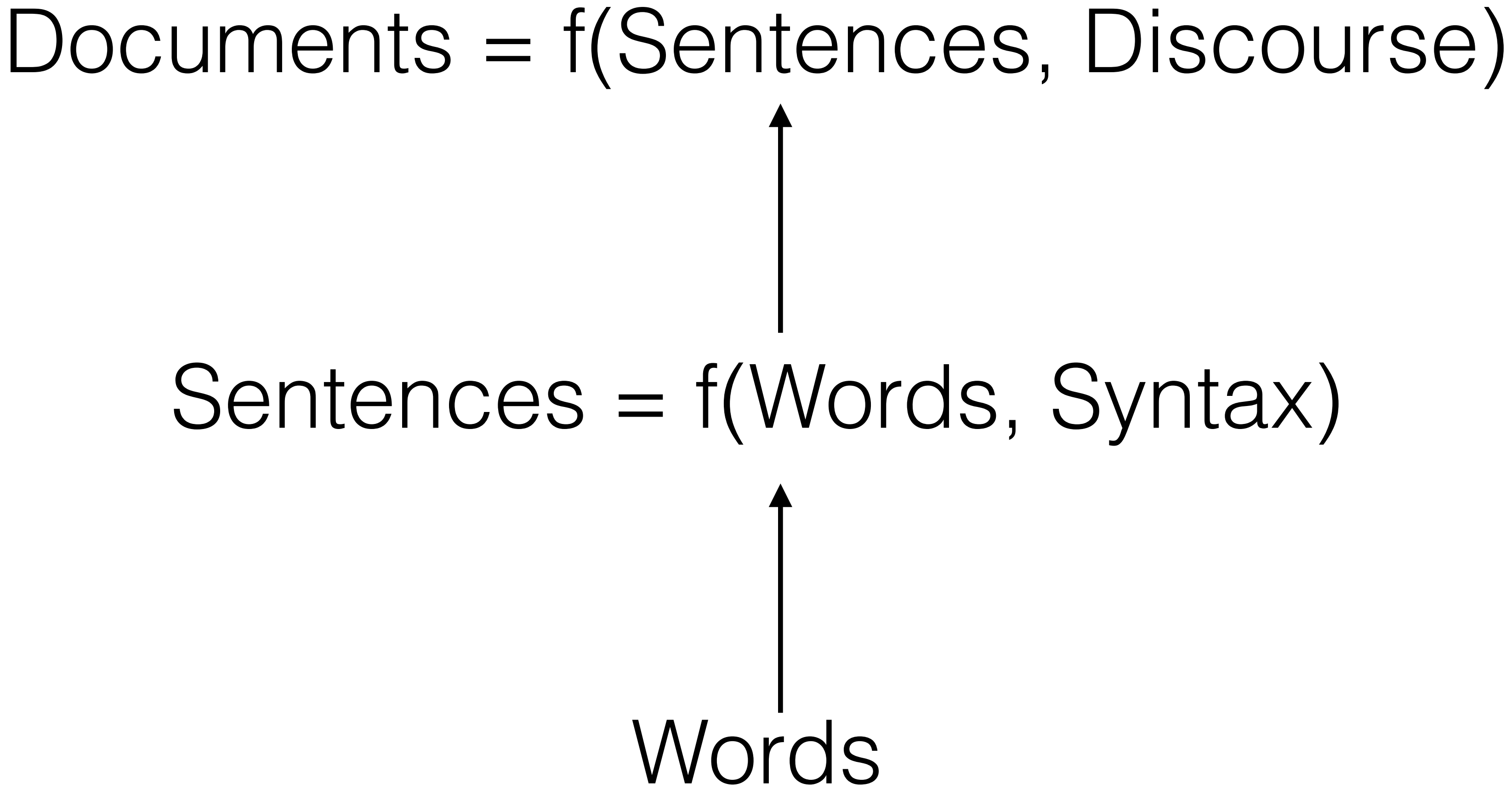
Words

Compositionality

Sentences = f(Words, Syntax)

Words

Compositionality



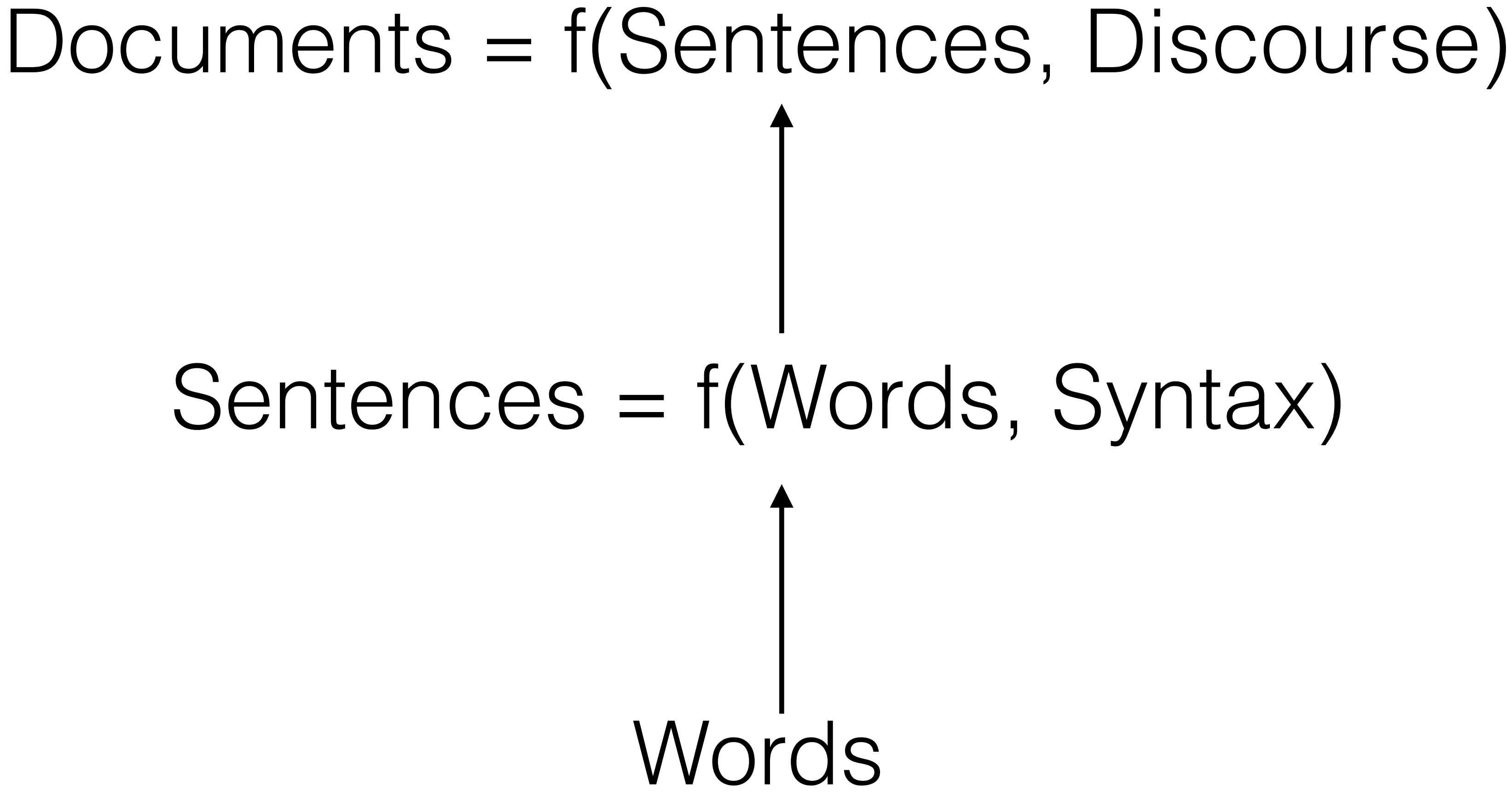
Documents = f(Sentences, Discourse)

Sentences = f(Words, Syntax)

Words

positionality

Very difficult…
(impossible?)
…to achieve



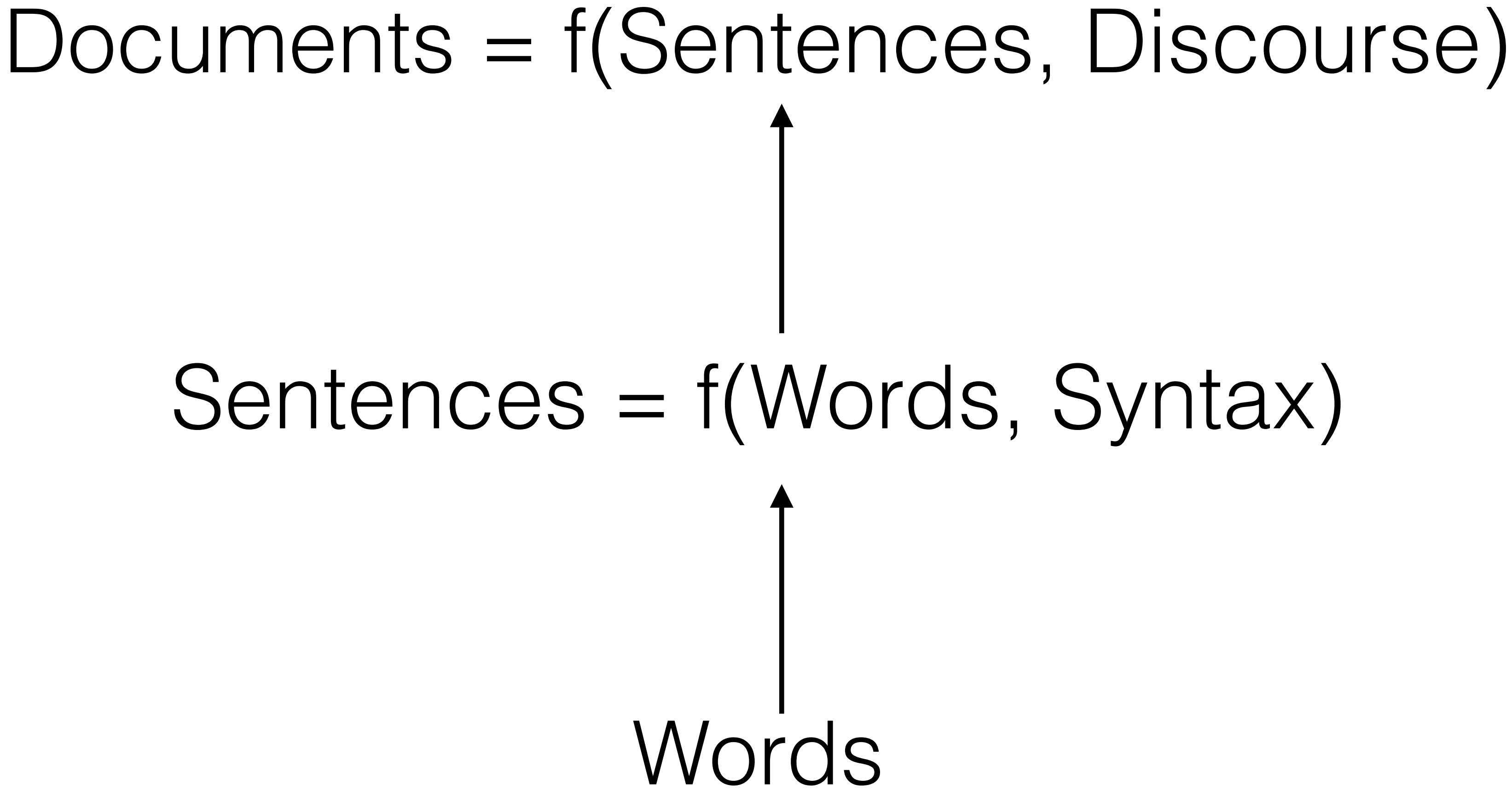
Documents = f(Sentences, Discourse)

Sentences = f(Words, Syntax)

Words

Very difficult...
(impossible?)
...to achieve

positionality



Documents = f(Sentences, Discourse)

Sentences = f(Words, Syntax)

Words

horse shoes ≈ alligator shoes?

Unit of analysis

- Characters
- Morphemes
- Words
- Sentences
- Documents

Today

- Characters
- Morphemes
- Words
- Sentences
- Documents

Today

- Characters
- Morphemes
- Words
- Sentences
- Documents

(We often treat sentences just like short documents, though)

“Bag of Words” (BOW)

“Bag of Words” (BOW)

- Represent sentences/documents as just an unordered set of words

“Bag of Words” (BOW)

- Represent sentences/documents as just an unordered set of words
- Basis of most of modern NLP
  - Information Retrieval/Search
  - Clustering/Recommendation
  - As input to most ML models

“Bag of Words” (BOW)

- Represent sentences/documents as just an unordered set of words
- Basis of most of modern NLP
  - Information Retrieval/Search
  - Clustering/Recommendation
  - As input to most ML models
- Changing a bit for sentences, but not for documents (yet)

“Bag of Words” (BOW)

Is it ok to copy and paste the data into javascript, or is there a filereader that can open a local file?

Changes I make to the nations.js file do not affect any of the html in after I load the nations.html file

When I try to display dots from part 2 on my mac (tried chrome, firefox, and safari), nothing is displayed (and the elements do not appear in the html).

“Bag of Words” (BOW)

> Is it ok to copy and paste the data into javascript, or is there a filereader that can open a local file?

1	1	1	1	1	…	0	0	1	0
is	it	a	and	copy	…	markets	below	paste	remorse

“Bag of Words” (BOW)

> Is it ok to copy and paste the data into javascript, or is there a filereader that can open a local file?

“one hot”

1	1	1	1	1	…	0	0	1	0
is	it	a	and	copy	…	markets	below	paste	remorse

“Bag of Words” (BOW)

> Is it ok to copy and paste the data into javascript, or is there a filereader that can open a local file?

counts/frequencies

2	1	2	1	1	…	0	0	1	0
is	it	a	and	copy	…	markets	below	paste	remorse

“Bag of Words” (BOW)

	is	it	a	and	copy	…	markets	below	paste	remorse
doc 1	1	1	2	1	0	…	2	1	0	0
doc 2	3	1	4	0	0	…	1	2	0	1
doc 3	2	1	2	1	1	…	0	0	1	0

“Bag of Words” (BOW)

	is	it	a	and	copy	…	markets	below	paste	remorse
doc 1	1	1	2	1	0	…	2	1	0	0
doc 2	3	1	4	0	0	…	1	2	0	1
doc 3	2	1	2	1	1	…	0	0	1	0

“Term Document Matrix”

"Bag of Words" (BOW)

	is	it	a	and	copy	…	markets	below	paste	remorse
doc 1	1	1	2	1	0	…	2	1	0	0
doc 2	3	1	4	0	0	…	1	2	0	1
doc 3	2	1	2	1	1	…	0	0	1	0

How similar are document 1 and document 2?

Similarity Metrics

Similarity Metrics

- Edit Distance: Minimal number of edits (inserts, deletes, substitutions) needed to transform string 1 into string 2.

Similarity Metrics

- Edit Distance: Minimal number of edits (inserts, deletes, substitutions) needed to transform string 1 into string 2.

Thoughts?

Similarity Metrics

- Edit Distance: Minimal number of edits (inserts, deletes, substitutions) needed to transform string 1 into string 2.
- Jaccard Similarity: words in common / total words

Clicker Question!

Clicker Question!

Query: html does not work

doc 1: When I try to display dots from part 2 the elements do not appear in the html.

doc 2: Changes I make do not affect any of the html in after I load the nations html file

Which document is more relevant to the query, according to Jaccard?

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

a) The first one

b) The second one

c) Yes

Clicker Question!

Query: html does not work

doc 1: When I try to display dots from part 2 the elements do not appear in the html.

doc 2: Changes I make do not affect any of the html in after I load the nations html file

Which document is more relevant to the query, according to Jaccard?

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

assume one-hot
(frequency
doesn't matter),
ignore case/
punctuation

a) The first one
b) The second one
c) Yes

Clicker Question!

Query: html does not work

doc 1: When I try to display dots from part 2 the elements do not appear in the html.

doc 2: Changes I make do not affect any of the html in after I load the nations html file

Which document is more relevant to the query, according to Jaccard?

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

a) The first one
b) The second one
c) Yes

Clicker Question!

Query: html does not work

doc 1: When I try to display dots from part 2 the elements do not appear in the html.

doc 2: Changes I make do not affect any of the html in after I load the nations html file

2/(4 + 17) = 0.095 ent is more relevant to the query,
cording to Jaccard?

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

a) The first one
b) The second one
c) Yes

Clicker Question!

Query: html does not work

doc 1: When I try to display dots from part 2 the elements do not appear in the html.

doc 2: Changes I make do not affect any of the html in after I load the nations html file

$2/(4 + 17) = 0.095$ ent is more relevant to the query,

$2/(4+18) = 0.091$ cording to Jaccard?

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

a) The first one
b) The second one
c) Yes

Clicker Question!

Query: html does not work

doc 1: When I try to display dots from part 2 the elements do not appear in the html.

doc 2: Changes I make do not affect any of the html in after I load the nations html file

$2/(4 + 17) = 0.095$ ent is more relevant to the query,

$2/(4+18) = 0.091$ cording to Jaccard?

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

[✓ a) The first one]

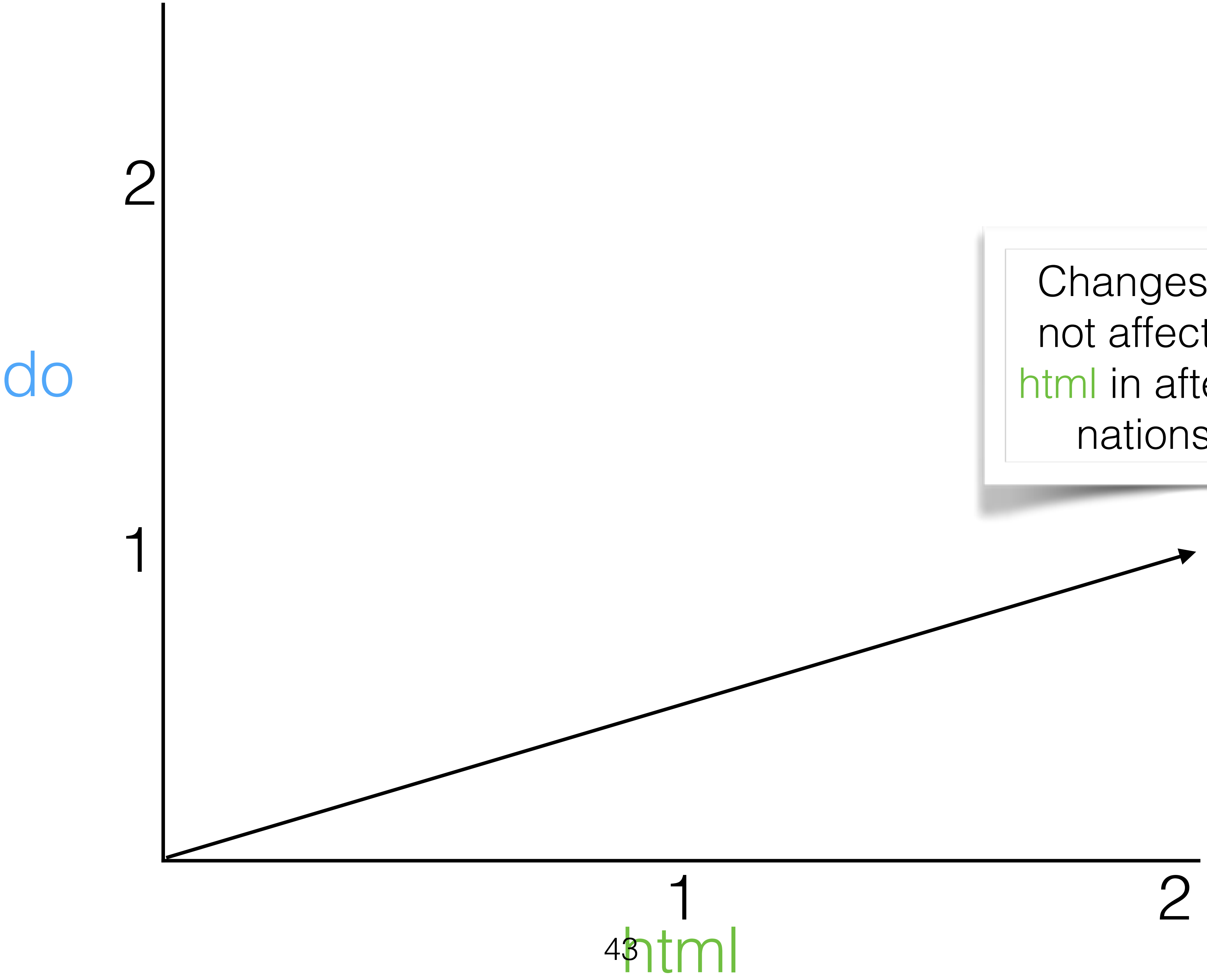
b) The second one

c) Yes

Similarity Metrics

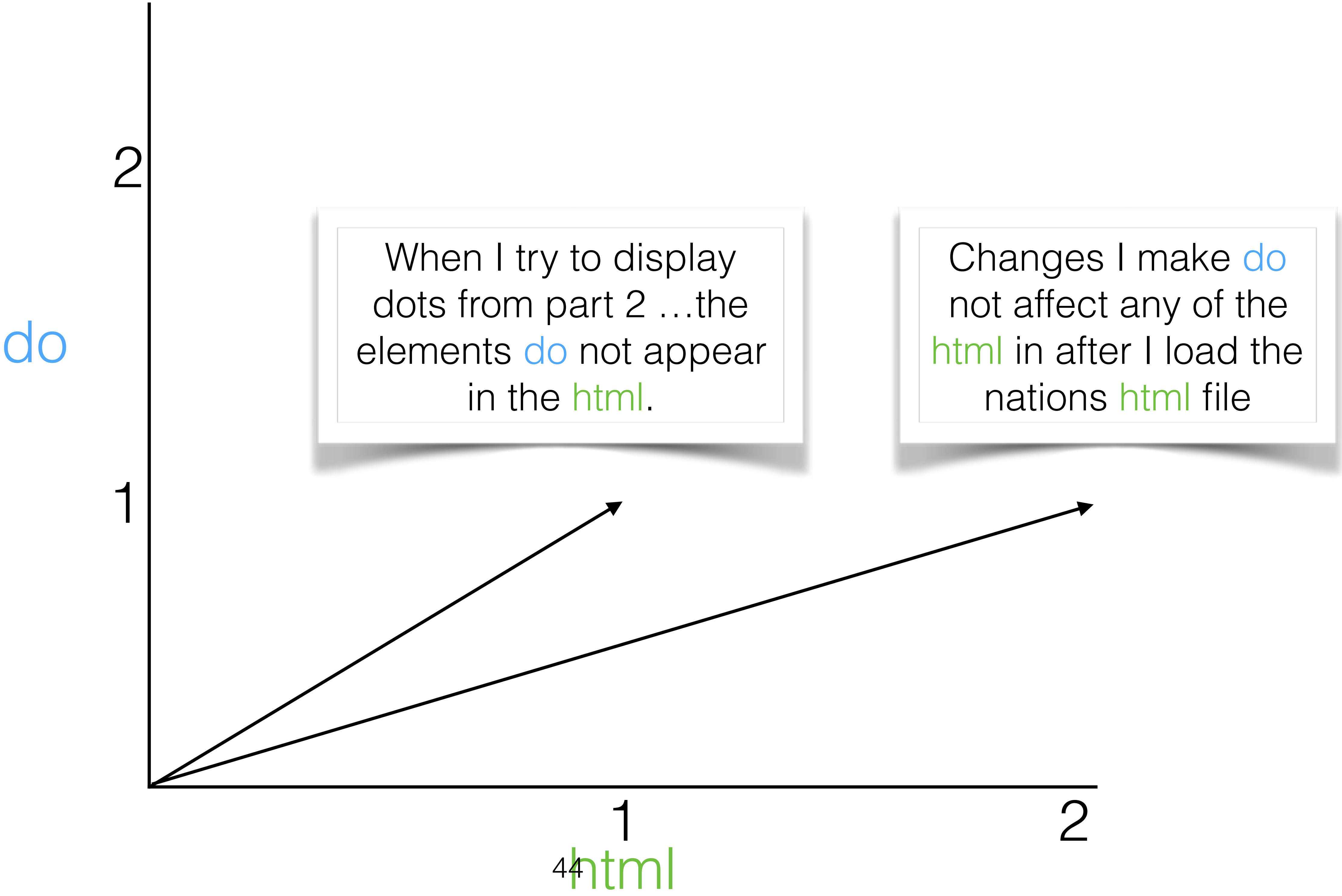
- Edit Distance: Minimal number of edits (inserts, deletes, substitutions) needed to transform string 1 into string 2.
- Jaccard Similarity: words in common / total words
- Cosine Similarity: by far the most popular metric

Cosine Similarity



Changes I make do not affect any of the html in after I load the nations html file

Cosine Similarity

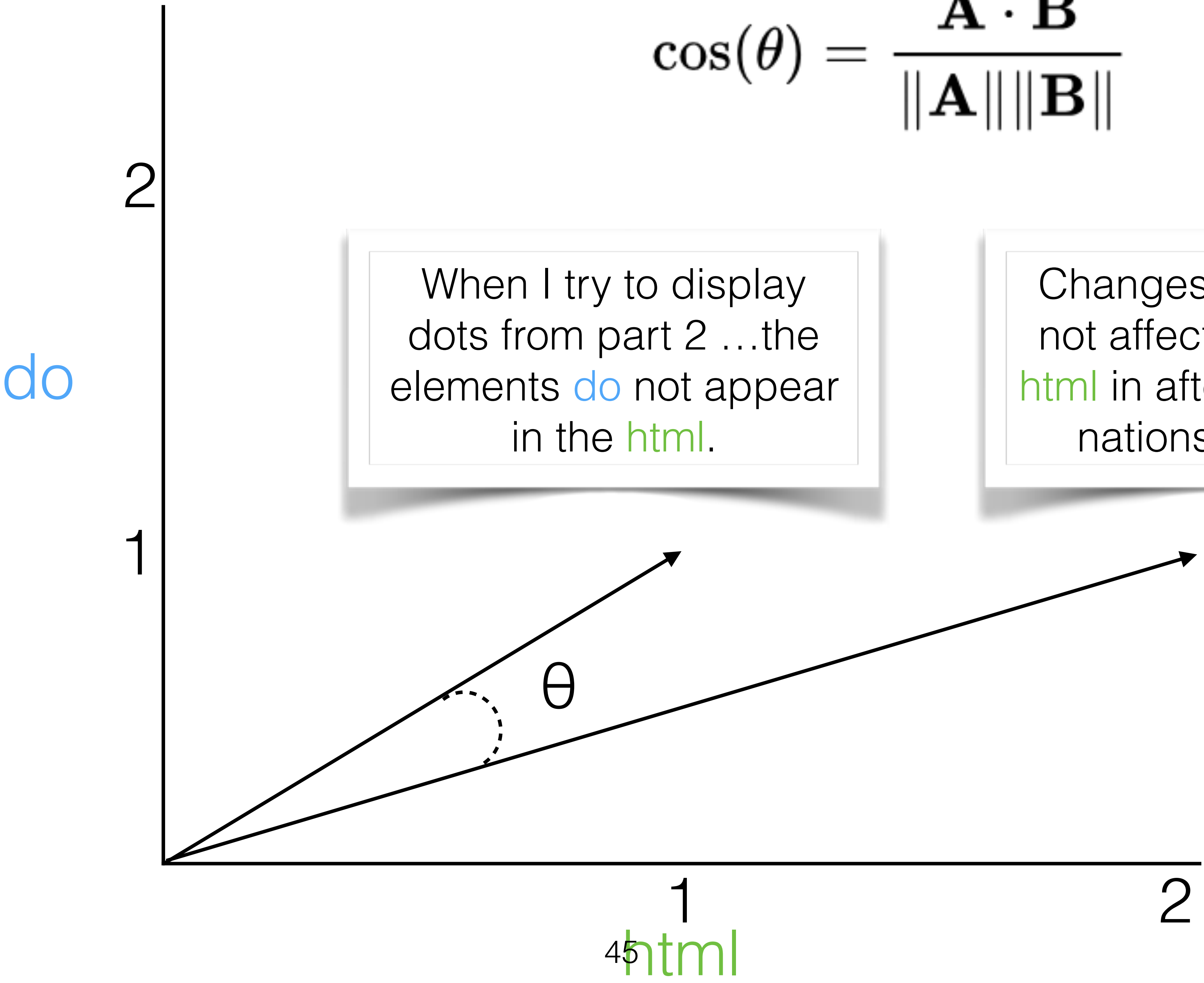


When I try to display dots from part 2 …the elements do not appear in the html.

Changes I make do not affect any of the html in after I load the nations html file

Cosine Similarity

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$$



When I try to display dots from part 2 …the elements do not appear in the html.

Changes I make do not affect any of the html in after I load the nations html file

Clicker Question!

Clicker Question!

	html	does	not	work	at	all	webdev	is	awesome
query	1	1	1	1	1	1	0	0	0
doc 1	1	1	0	0	0	1	1	1	1
doc 2	1	1	0	1	0	0	1	0	0

Which document is more relevant to the query, according to cosine?

a) doc1
b) doc2
c) Yes

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Clicker Question!

	html	does	not	work	at	all	webdev	is	awesome
query	1	1	1	1	1	1	0	0	0
doc 1	1	1	0	0	0	1	1	1	1
doc 2	1	1	0	1	0	0	1	0	0

$3/(\sqrt{6}\sqrt{6}) = 0.5$

Which document is more releva
according to cosin

a) doc1
b) doc2
c) Yes

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Clicker Question!

	html	does	not	work	at	all	webdev	is	awesome
query	1	1	1	1	1	1	0	0	0
doc 1	1	1	0	0	0	1	1	1	1
doc 2	1	1	0	1	0	0	1	0	0

$3/(\sqrt{6}\sqrt{6}) = 0.5$

$3/(\sqrt{6}\sqrt{4}) = 0.6$

Which document is more releva
according to cosin

a) doc1
b) doc2
c) Yes

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Clicker Question!

	html	does	not	work	at	all	webdev	is	awesome
query	1	1	1	1	1	1	0	0	0
doc 1	1	1	0	0	0	1	1	1	1
doc 2	1	1	0	1	0	0	1	0	0

$3/(\sqrt{6}\sqrt{6}) = 0.5$

$3/(\sqrt{6}\sqrt{4}) = 0.6$

Which document is more releva
according to cosin

a) doc1
[✓ b) doc2]
c) Yes

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Linguistic Preprocessing

Linguistic Preprocessing

Language is ambiguous

Linguistic Preprocessing

Language is ambiguous

They freaked out when they found
the bug in their apartment.

Linguistic Preprocessing

Language is ambiguous

They freaked out when they found
the **bug** in their apartment.

Linguistic Preprocessing

Language is ambiguous

They freaked out when they found
the **bug** in their apartment.

They've always been terrified of anything crawly.

Linguistic Preprocessing

Language is ambiguous

They freaked out when they found
the **bug** in their apartment.

They ran back the CIT right away to tell
everyone they'd finally figured it out.

Linguistic Preprocessing

Language is ambiguous
but also redundant

They freaked out when they found
the **problem** in their apartment.

They ran back the CIT right away to tell
everyone they'd finally figured it out.

Linguistic Preprocessing

Constant Tradeoff

Linguistic Preprocessing

Constant Tradeoff

Collapse!
Try to treat
more words as
though they are
the same

Linguistic Preprocessing

Constant Tradeoff

Collapse!
Try to treat more words as though they are the same

Differentiate!
Try to preserve as much differences/ nuance as possible

Linguistic Preprocessing

Constant Tradeoff

Collapse!
Try to treat more words as though they are the same

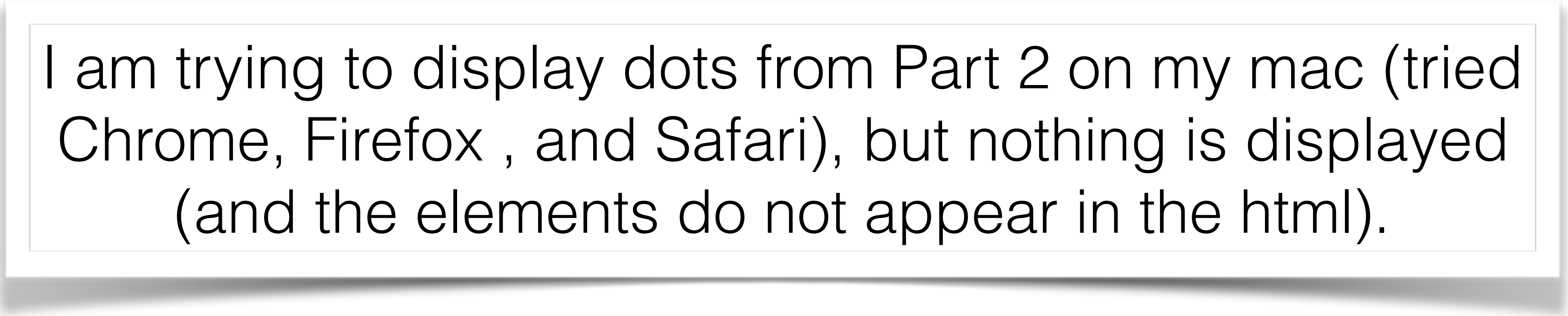
Differentiate!
Try to preserve as much differences/ nuance as possible

normalization, stemming

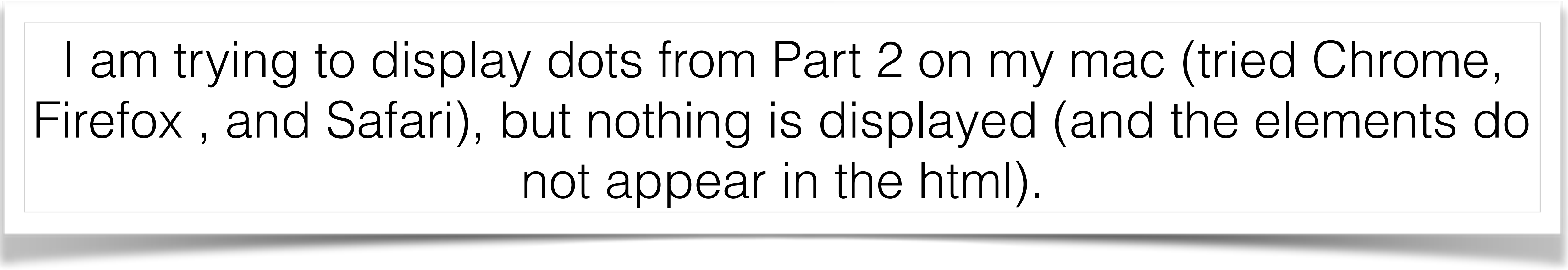
tagging, collocations

Linguistic Preprocessing

Linguistic Preprocessing

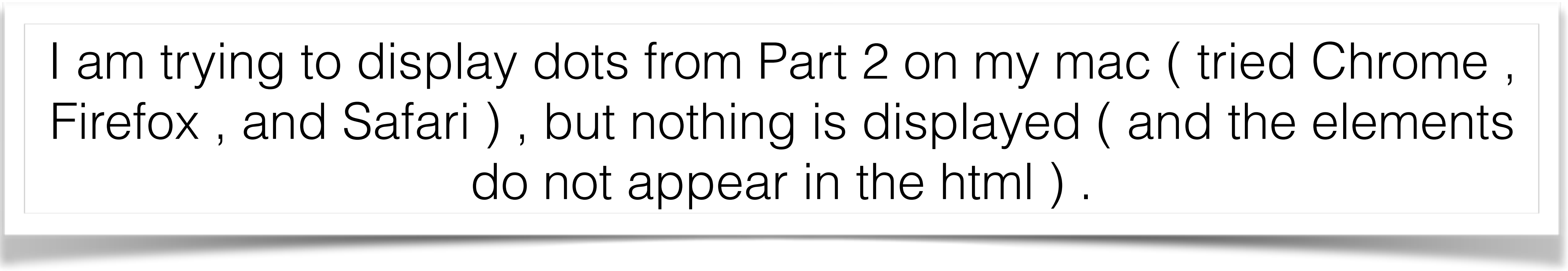


I am trying to display dots from Part 2 on my mac (tried Chrome, Firefox , and Safari), but nothing is displayed (and the elements do not appear in the html).



I am trying to display dots from Part 2 on my mac (tried Chrome, Firefox , and Safari), but nothing is displayed (and the elements do not appear in the html).

- Tokenization (Phrasal Collocations/Morphological Analysis?)



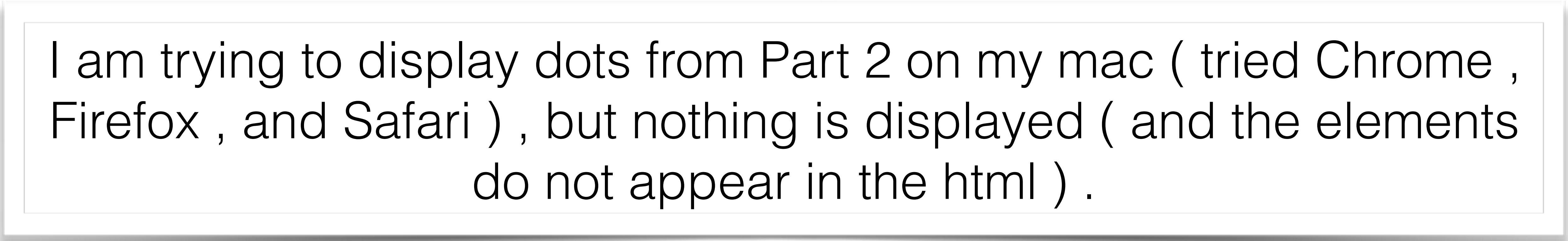
I am trying to display dots from Part 2 on my mac (tried Chrome , Firefox , and Safari) , but nothing is displayed (and the elements do not appear in the html) .

- Tokenization (Phrasal Collocations/Morphological Analysis?)

I am trying to display dots from Part 2 on my mac (tried Chrome , Firefox , and Safari) , but nothing is displayed (and the elements do not appear in the html) .

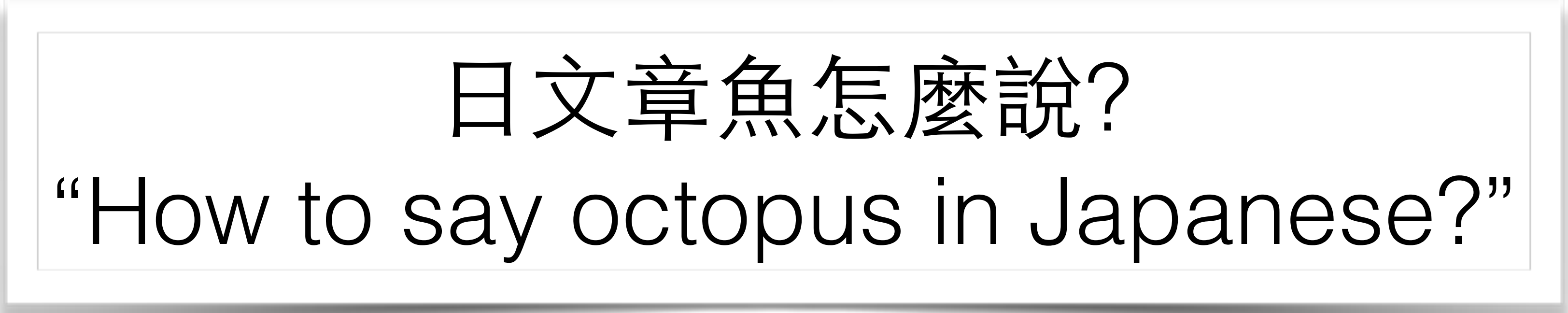
- Tokenization (Phrasal Collocations/Morphological Analysis?)

日文章魚怎麼說?
“How to say octopus in Japanese?”

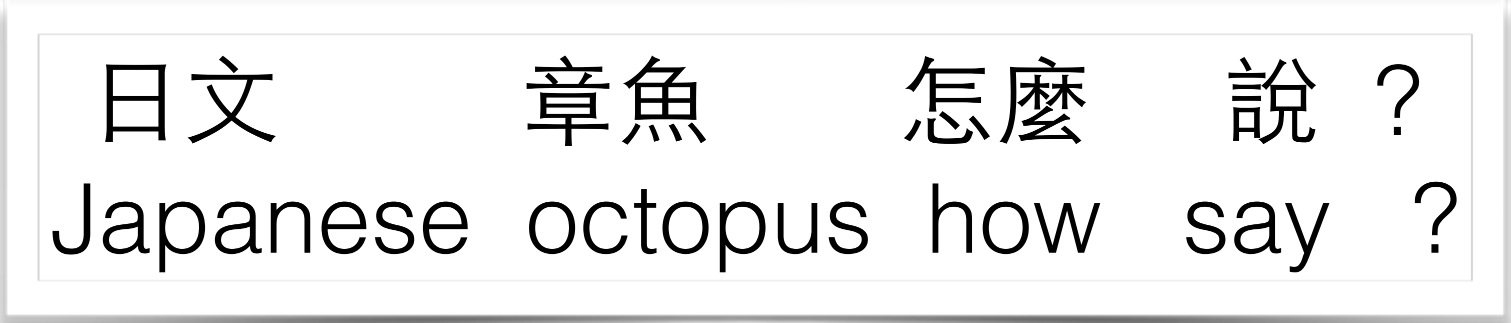


I am trying to display dots from Part 2 on my mac (tried Chrome , Firefox , and Safari) , but nothing is displayed (and the elements do not appear in the html) .

- Tokenization (Phrasal Collocations/Morphological Analysis?)



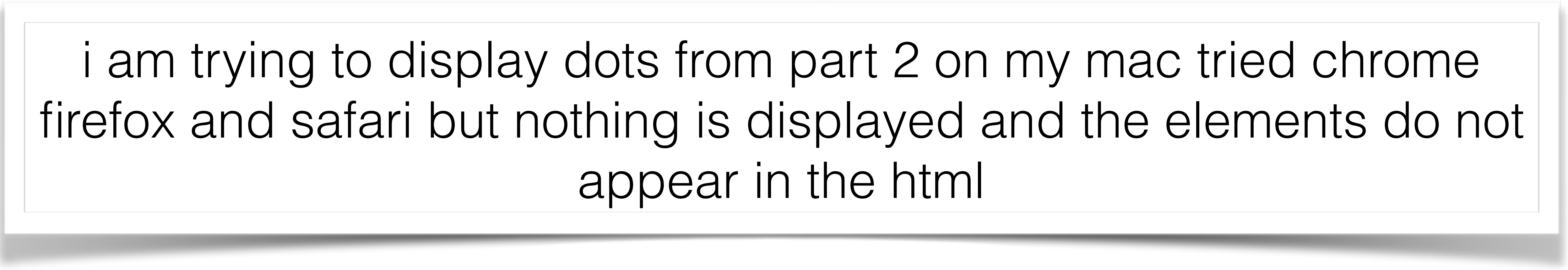
日文章魚怎麼說?
“How to say octopus in Japanese?”



日文	章魚	怎麼	說	?
Japanese	octopus	how	say	?

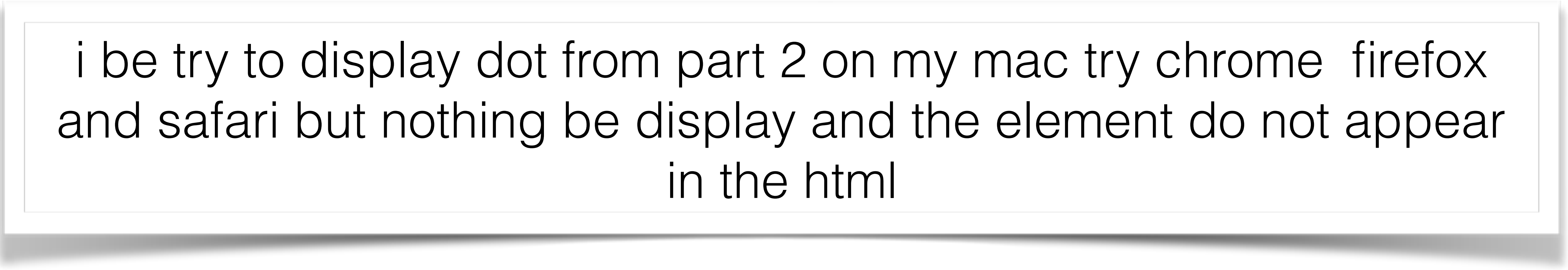
> I am trying to display dots from Part 2 on my mac tried Chrome Firefox and Safari but nothing is displayed and the elements do not appear in the html

- Tokenization (Phrasal Collocations/Morphological Analysis?)
- Punctuation — “okay…” vs. “okay!”



i am trying to display dots from part 2 on my mac tried chrome firefox and safari but nothing is displayed and the elements do not appear in the html

- Tokenization (Phrasal Collocations/Morphological Analysis?)
- Punctuation — “okay…” vs. “okay!”
- Normalization — “Trump” vs. “trump”



i be try to display dot from part 2 on my mac try chrome firefox and safari but nothing be display and the element do not appear in the html

- Tokenization (Phrasal Collocations/Morphological Analysis?)

- Punctuation — “okay…” vs. “okay!”

- Normalization — “Trump” vs. “trump”

> i be try to display dot from part <NUM> on my mac try chrome firefox and safari but nothing be display and the element do not appear in the html

- Tokenization (Phrasal Collocations/Morphological Analysis?)
- Punctuation — “okay…” vs. “okay!”
- Normalization — “Trump” vs. “trump”

try display dot part <NUM> mac try chrome firefox safari nothing display element not appear html

- Tokenization (Phrasal Collocations/Morphological Analysis?)

- Punctuation — “okay…” vs. “okay!”

- Normalization — “Trump” vs. “trump”

- Stop words — “pb and jelly” vs. “pb or jelly”

try_VB display_VB dot_NN part_NN <NUM>_NUM mac_NNP
try_VB chrome_NNP firefox_NNP safari_NNP nothing_DT
display_VB element_NNP not_RB appear_VB html_NN

- Tokenization (Phrasal Collocations/Morphological Analysis?)
- Punctuation — “okay…” vs. “okay!”
- Normalization — “Trump” vs. “trump”
- Stop words — “pb and jelly” vs. “pb or jelly”
- Tagging — “fish fish fish fish fish”

```
try_VB display_VB dot_NN part_NN <NUM>_NUM mac_NNP
try_VB chrome_NNP <OOV> <OOV> nothing_DT display_VB
element_NNP not_RB appear_VB html_NN
```

- Tokenization (Phrasal Collocations/Morphological Analysis?)

- Punctuation — “okay…” vs. “okay!”

- Normalization — “Trump” vs. “trump”

- Stop words — “pb and jelly” vs. “pb or jelly”

- Tagging — “fish fish fish fish fish”

- Remove out-of-vocabulary (OOV)

try_VB display_VB dot_NN part_NN <NUM>_NUM mac_NNP
try_VB chrome_NNP <OOV> <OOV> nothing_DT display_VB
element_NNP not_RB appear_VB html_NN

- Tokenization (Phrasal Collocations/Morphological Analysis?)
- Punctuation — “okay…” vs. “okay!”
- Normalization — “Trump” vs. “trump”
- Stop words — “pb and jelly” vs. “pb or jelly”
- Tagging — “fish fish fish fish fish”
- Remove out-of-vocabulary (OOV)

Choosing a vocabulary

(what goes on the columns)

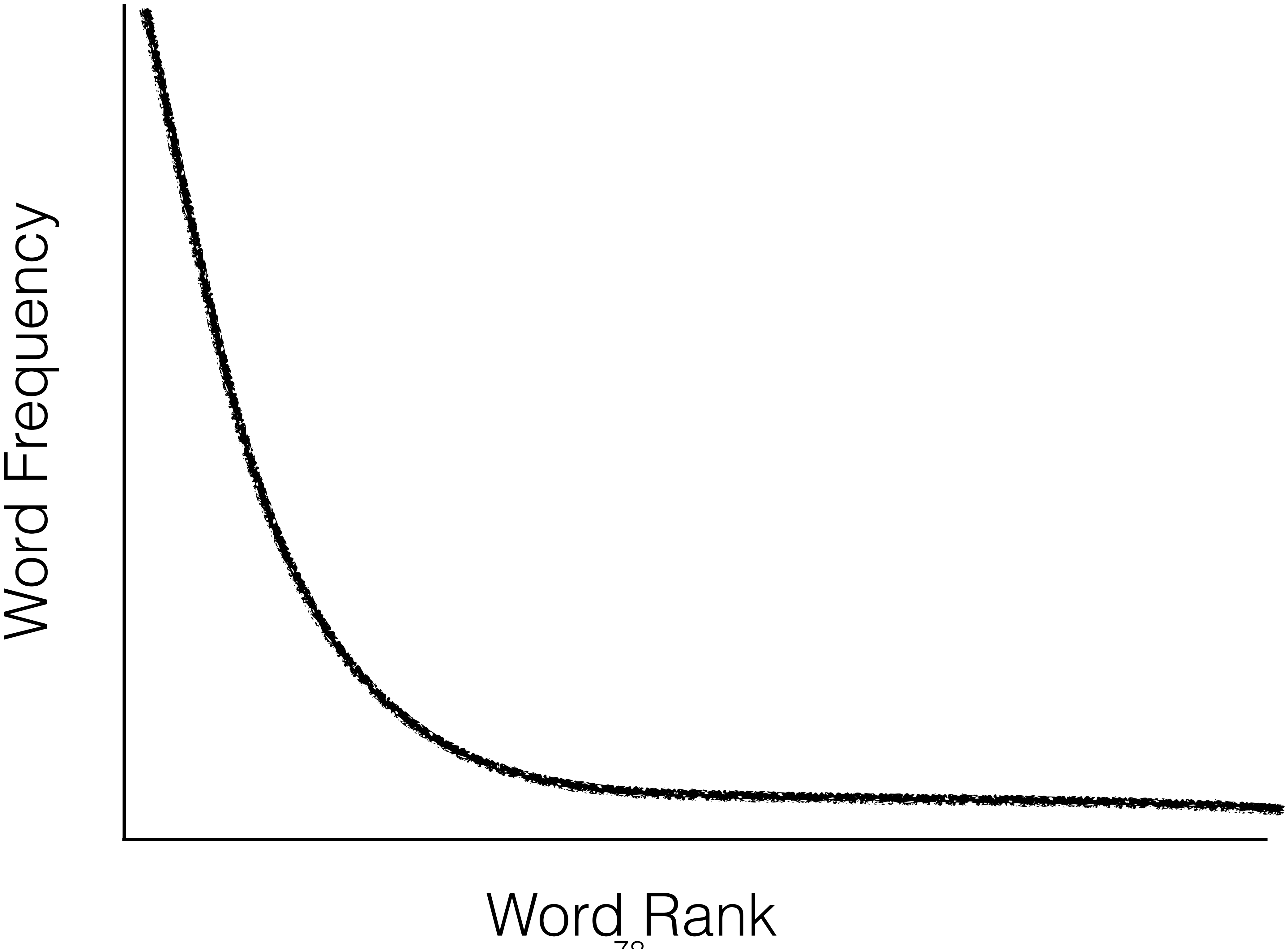
- Remove frequent words? (“stop words”)
- Remove rare words? (unlikely to appear in test)
- Remove uninteresting words? (tf-idf? pmi?)
- Try to add a little syntax? (POS tags? ngrams? pmi?)

Choosing a vocabulary

(what goes on the columns)

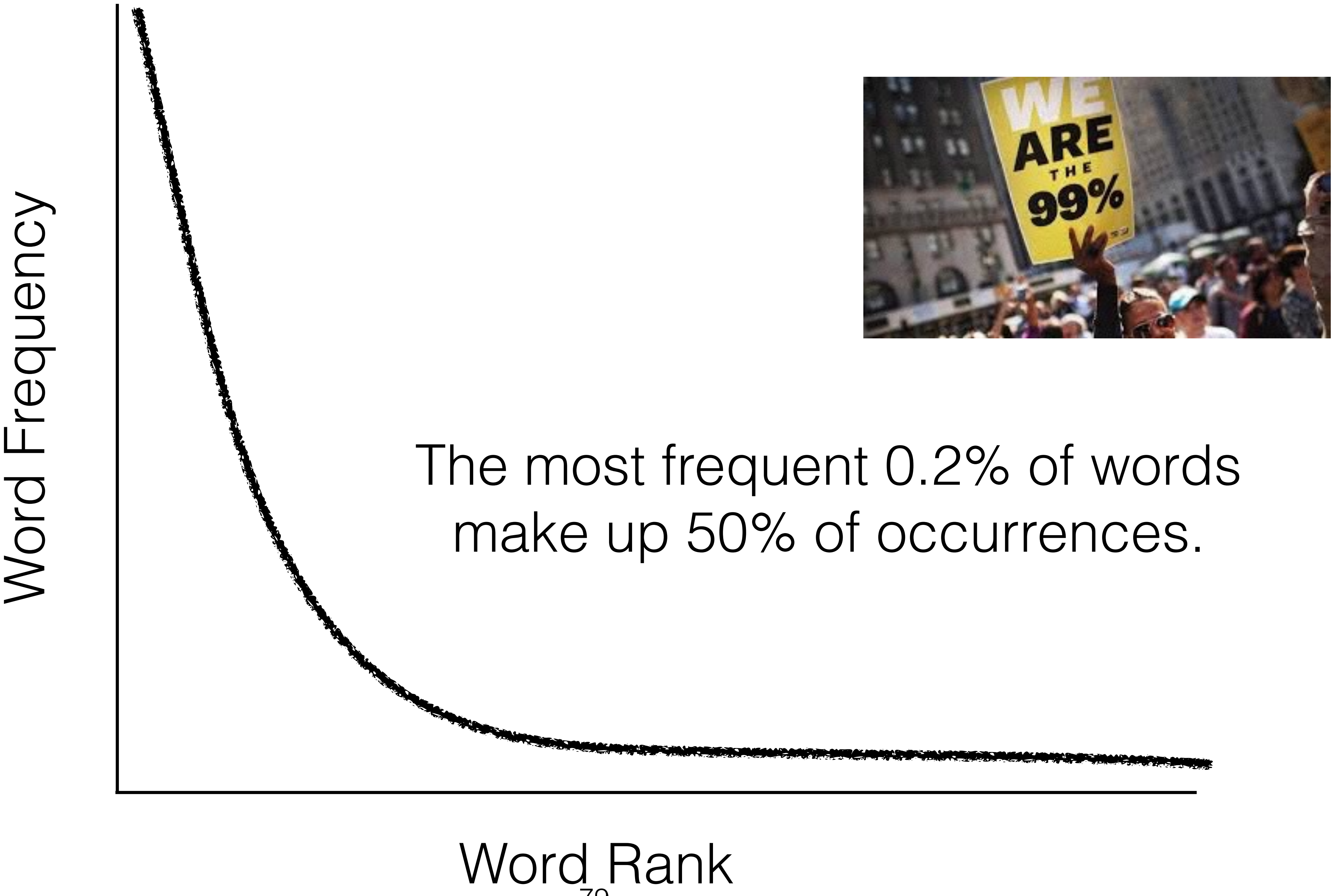
- Remove frequent words? (“stop words”)
- Remove rare words? (unlikely to appear in test)
- Remove uninteresting words? (tf-idf? pmi?)
- Try to add a little syntax? (POS tags? ngrams? pmi?)

Zipf’s Law



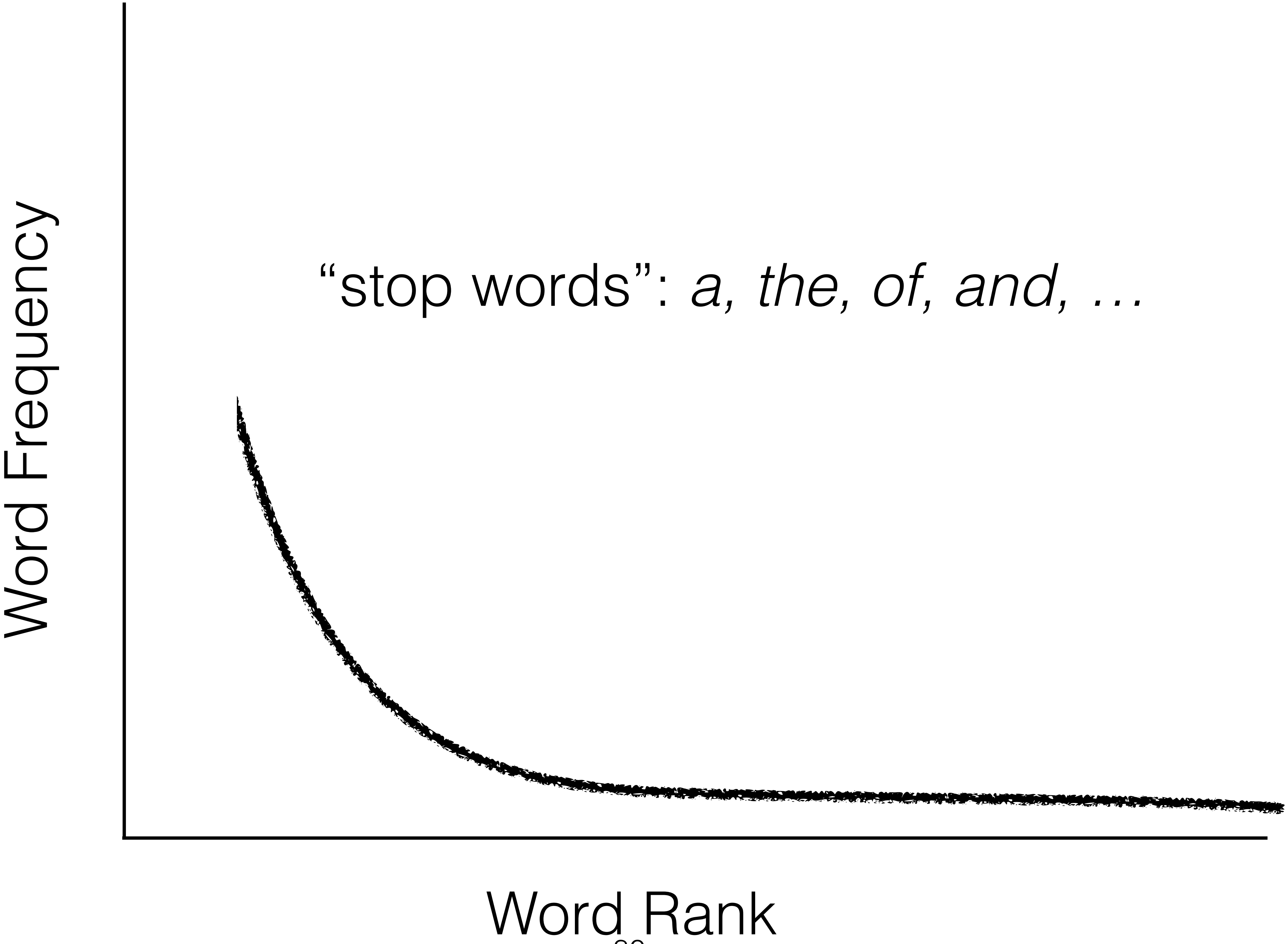
https://en.wikipedia.org/wiki/Zipf%27s_law

Zipf’s Law



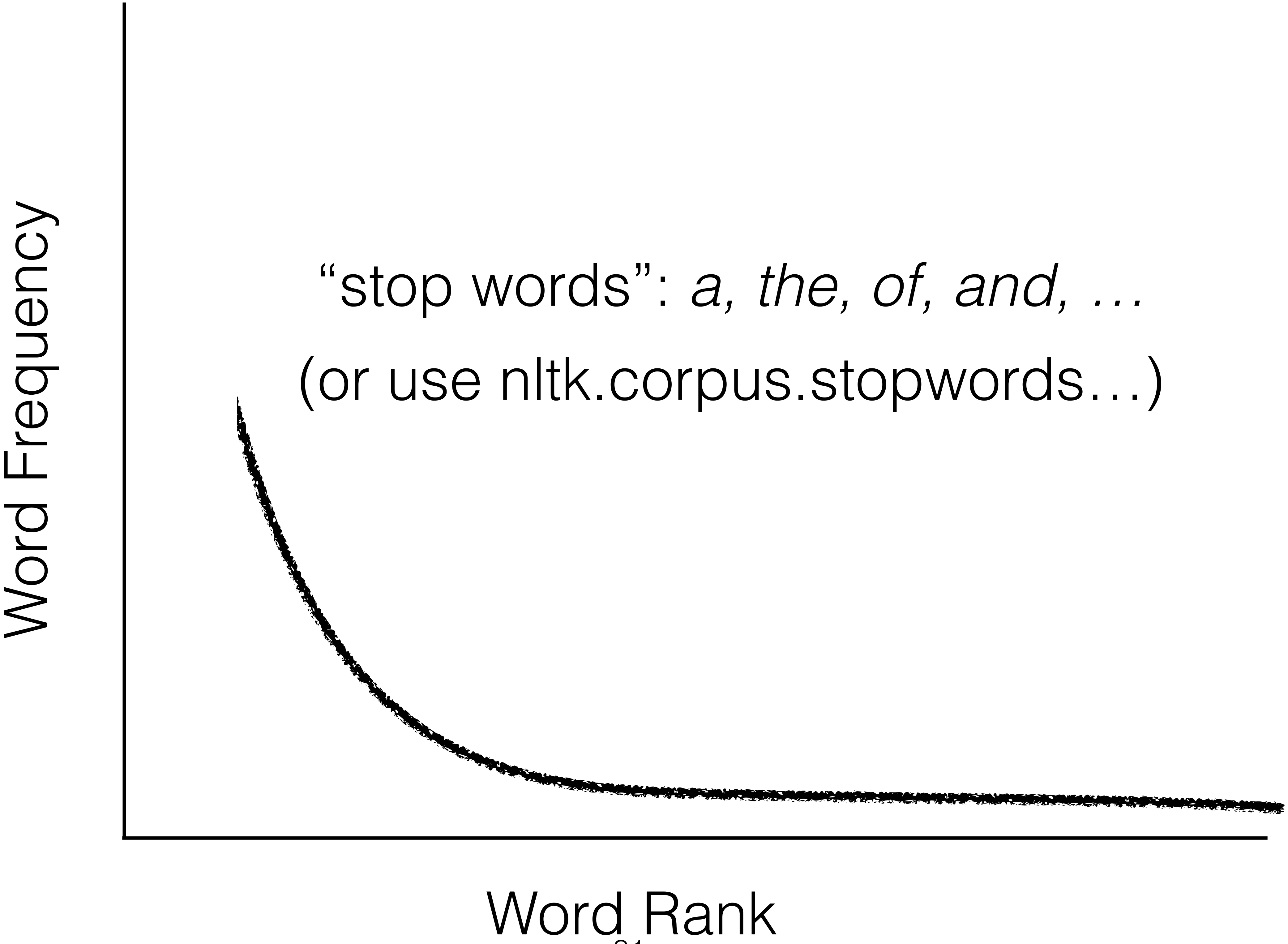
The most frequent 0.2% of words make up 50% of occurrences.

Zipf’s Law



“stop words”: *a, the, of, and, …*

Zipf’s Law



“stop words”: *a, the, of, and, …*

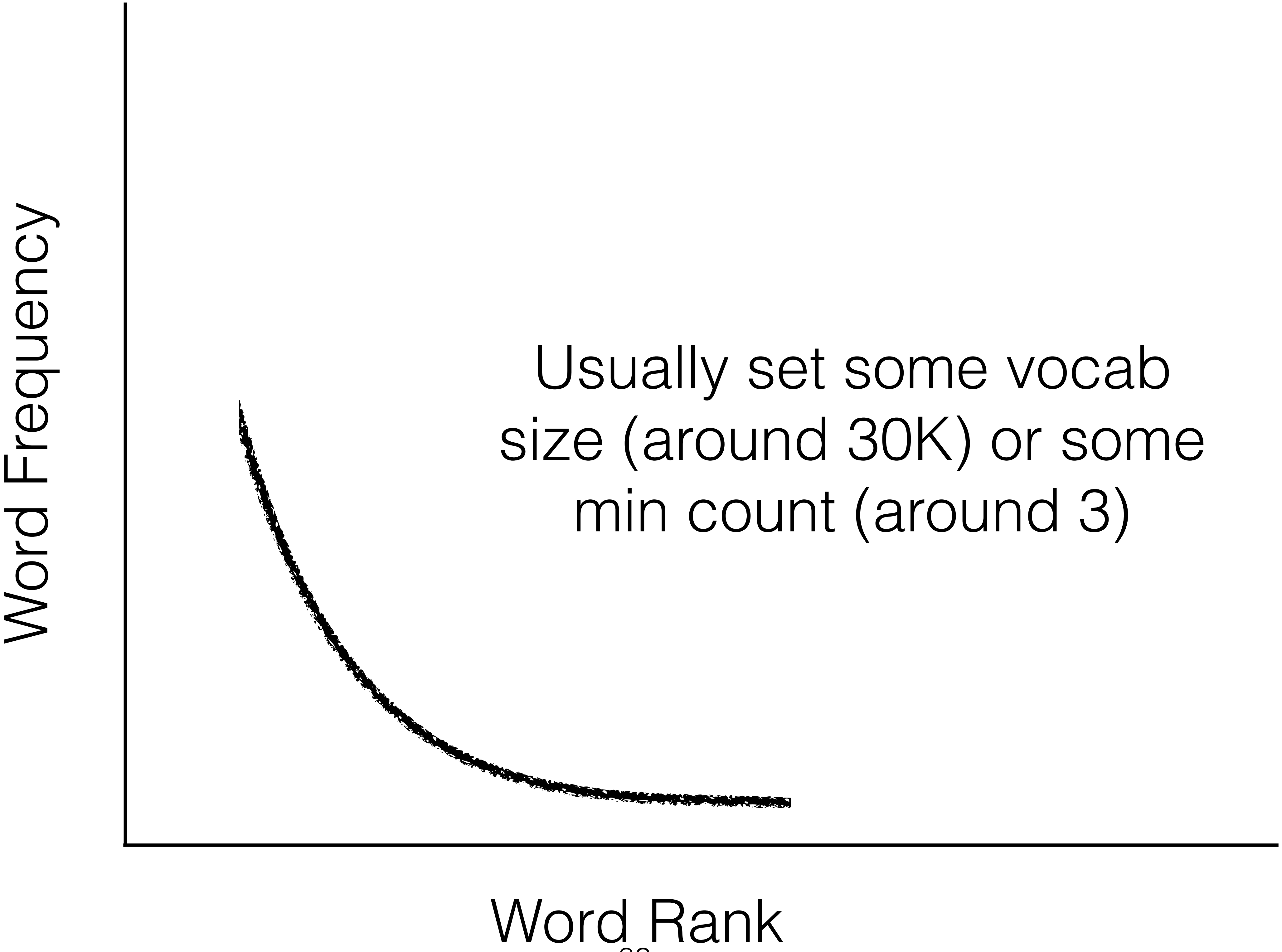
(or use nltk.corpus.stopwords…)

Choosing a vocabulary

(what goes on the columns)

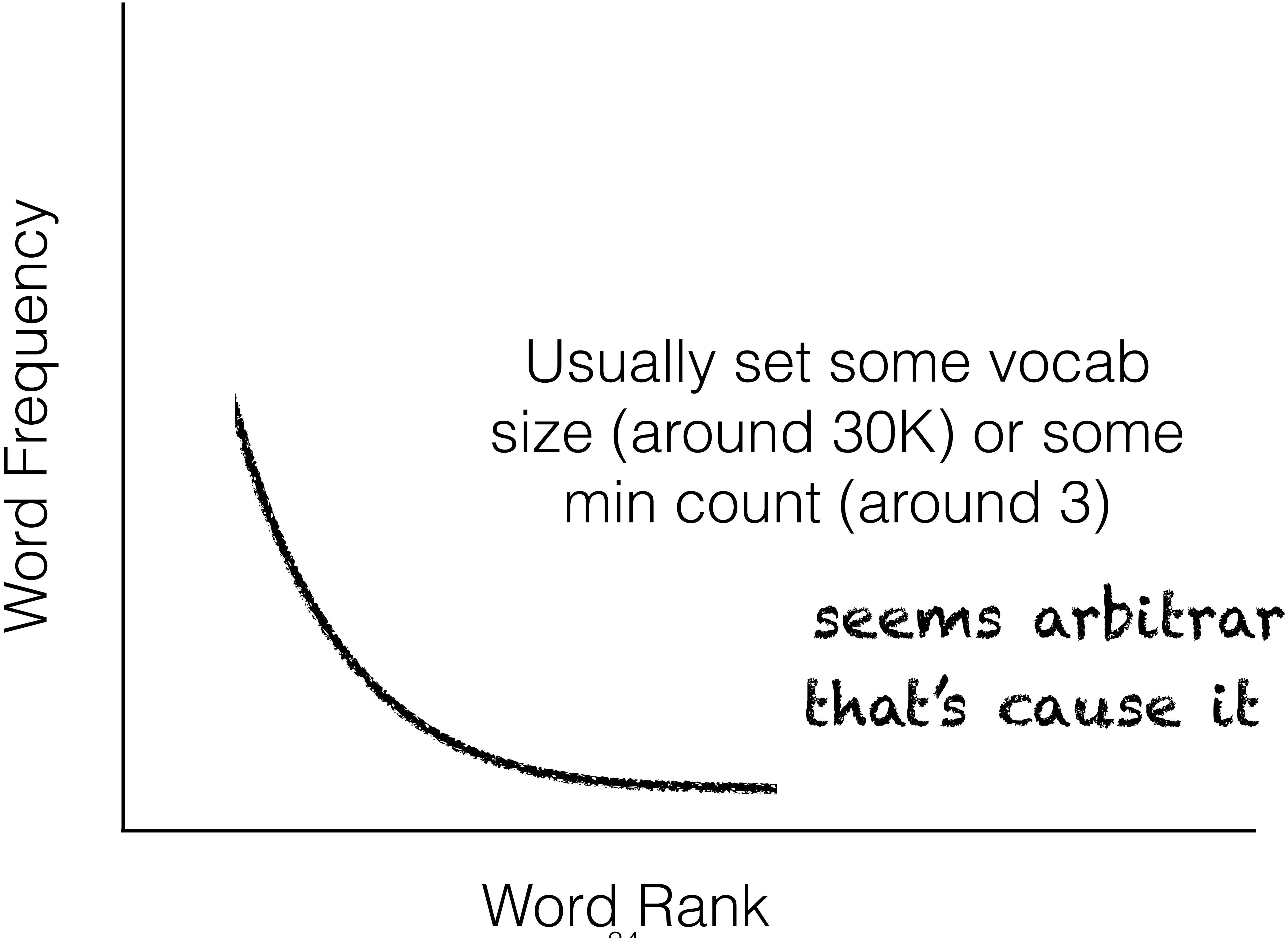
- Remove frequent words? ("stop words")
- Remove rare words? (unlikely to appear in test)
- Remove uninteresting words? (tf-idf? pmi?)
- Try to add a little syntax? (POS tags? ngrams? pmi?)

Zipf’s Law



Usually set some vocab size (around 30K) or some min count (around 3)

Zipf’s Law



Usually set some vocab size (around 30K) or some min count (around 3)

seems arbitrary?
that’s cause it is.

Choosing a vocabulary

(what goes on the columns)

- Remove frequent words? (“stop words”)
- Remove rare words? (unlikely to appear in test)
- Remove uninteresting words? (tf-idf? pmi?)
- Try to add a little syntax? (POS tags? ngrams? pmi?)

Tf-Idf

- Term-Frequency Inverse-Document-Frequency
- Assigns higher weights to words that differentiate this document from other documents
- tf-idf(word,doc) = (# times word appears in doc) / (# of times word appears across all documents)
- Can filter out low tf-idf words or else just reweight the term-document matrix accordingly

Clicker Question!

Clicker Question!

doc1

html does not work

doc 2

html does work. all
webdev is awesome.

doc 3

webdev: html
does work

	html	does	not	work	at	all	webdev	is	awesome
doc1	1	1	1	1	1	1	0	0	0
doc 2	1	1	0	0	0	1	1	1	1
doc 3	1	1	0	1	0	0	1	0	0

Clicker Question!

doc1

html does not work

doc 2

html does work. all webdev is awesome.

doc 3

webdev: html does work

	html	does	not	work	at	all	webdev	is	awesome
doc1	1	1	1	1	1	1	0	0	0
doc 2	1	1	0	0	0	1	1	1	1
doc 3	1	1	0	1	0	0	1	0	0

What is the tf-idf vector for doc1

a)	1/3	1/3	1	1/3	0	1/2	1	0	1
b)	1/2	1/3	1	1/3	1	1/2	0	1/2	1
c)	1/3	1/3	1	1/2	1	1/2	0	0	0

Clicker Question!

html does not work

html does work. all webdev is awesome.

webdev: html does work

	html	does	not	work	at	all	webdev	is	awesome
doc1	1	1	1	1	1	1	0	0	0
doc 2	1	1	0	0	0	1	1	1	1
doc 3	1	1	0	1	0	0	1	0	0

df
html: 3
does: 3
not: 1
work: 2
at: 1
all: 2
webdev: 2
is: 1
awesome: 1

What is the tf-idf vector for doc1

a)	1/3	1/3	1	1/3	0	1/2	1	0	1
b)	1/2	1/3	1	1/3	1	1/2	0	1/2	1
c)	1/3	1/3	1	1/2	1	1/2	0	0	0

Clicker Question!

html does not work

html does work. all webdev is awesome.

webdev: html does work

	html	does	not	work	at	all	webdev	is	awesome
doc1	1	1	1	1	1	1	0	0	0
doc 2	1	1	0	0	0	1	1	1	1
doc 3	1	1	0	1	0	0	1	0	0

df

html:	**3**
does:	3
not:	1
work:	2
at:	1
all:	2
webdev:	2
is:	1
awesome:	1

What is the tf-idf vector for doc1

a)	1/3	1/3	1	1/3	0	1/2	1	0	1
b)	1/2	1/3	1	1/3	1	1/2	0	1/2	1
[✓ c)]	1/3	1/3	1	1/2	1	1/2	0	0	0

PMI

- Pointwise Mutual Information
- Again: assigns higher weights to words that differentiate this document from other documents
- PMI(word,doc) = log P(word|doc)/P(word)
- Used more for finding word-label relationships or word-word collocations (more info in two seconds)

Choosing a vocabulary

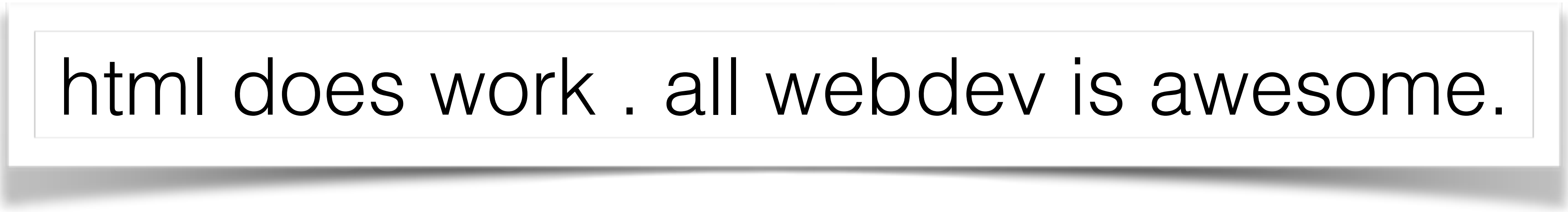
(what goes on the columns)

- Remove frequent words? ("stop words")
- Remove rare words? (unlikely to appear in test)
- Remove uninteresting words? (tf-idf? pmi?)
- Try to add a little syntax? (POS tags? ngrams? pmi?)

N-Grams

- N-length sequence of words (unigrams, bigrams, trigrams, 4-grams, …)
- Provides some context (differentiating “cute **dog”** from “hot **dog**”)
- Blows up size of vocabulary, increases sparsity

N-Grams


html does work . all webdev is awesome.

1gms: [‘html’, ‘does’, ‘work’, ‘.’, ‘all’, …]

2gms: [‘html does’, ‘does work’, ‘work .’, ‘. all’, …]

3gms: [‘html does work’, ‘does work .’, ‘work . all’, …]

skip-gms: [‘html does’, ‘html work’, ‘does html’, ‘does work’, ‘does .’, …]

Collocations

- Try to find just the interesting phrases (e.g. hot dog) by finding words that occur together above chance
- Often use PMI for this

Topic Models

Can you elaborate on exactly what the directions are in part 2 step 3, the stencil code does not quite imply what we are supposed to do…

When I try to display dots from part 2 on my mac (tried chrome, firefox, and safari), the elements do not appear in the html.

Changes I make to the nations.js file do not affect any of the html in after I load the nations.html file

Topic Models

Can you elaborate on exactly what the directions are in part 2 step 3, the stencil code does not quite imply what we are supposed to do…

When I try to display dots from part 2 on my mac (tried chrome, firefox, and safari), the elements do not appear in the html.

Changes I make to the nations.js file do not affect any of the html in after I load the nations.html file

instructions: stencil, instructions, part, step, rubric, handin…
UI: html, javascript, debug, display, elements…
systems: mac, windows, linux, chrome, firefox, os…
fillers: I, you, when, the, and, a

Topic Models

Where do documents come from?
“The generative story”

instructions: stencil, instructions, part, step, rubric, handin…
UI: html, javascript, debug, display, elements…
systems: mac, windows, linux, chrome, firefox, os…
fillers: I, you, when, the, and, a

Topic Models

Where do documents come from?
“The generative story”

instructions: stencil, instructions, part, step, rubric, handin…
UI: html, javascript, debug, display, elements…
systems: mac, windows, linux, chrome, firefox, os…
fillers: I, you, when, the, and, a

1. Sample a topic

Topic Models

Where do documents come from?
"The generative story"

instructions: stencil, instructions, part, step, rubric, handin…
UI: html, javascript, debug, display, elements…
systems: mac, windows, linux, chrome, firefox, os…
fillers: I, you, when, the, and, a

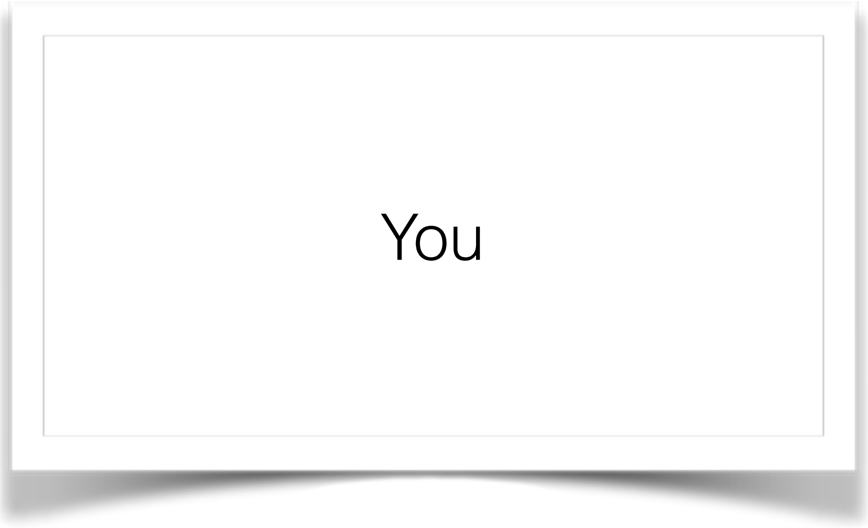
You

2. Sample a word from that topic

Topic Models

Where do documents come from?
"The generative story"

instructions: stencil, instructions, part, step, rubric, handin…
UI: html, javascript, debug, display, elements…
systems: mac, windows, linux, chrome, firefox, os…
fillers: I, you, when, the, and, a



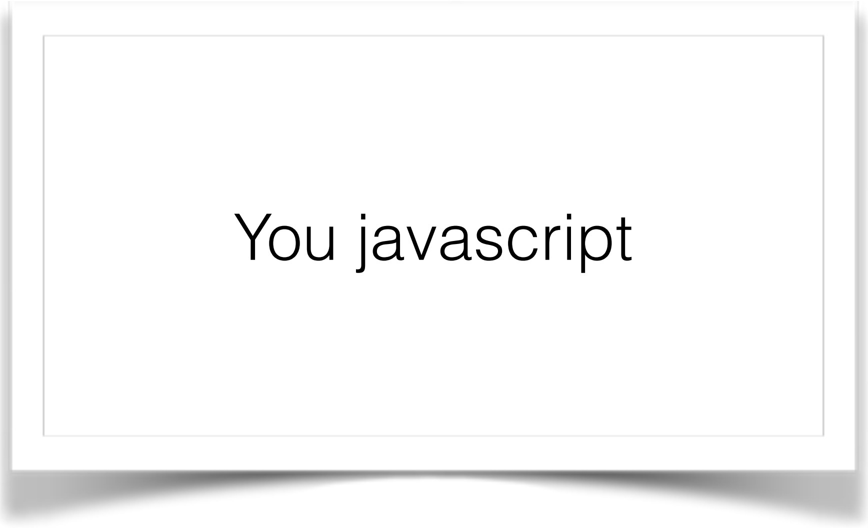
You

1. Sample a topic

Topic Models

Where do documents come from?
“The generative story”

instructions: stencil, instructions, part, step, rubric, handin…
UI: html, javascript, debug, display, elements…
systems: mac, windows, linux, chrome, firefox, os…
fillers: I, you, when, the, and, a



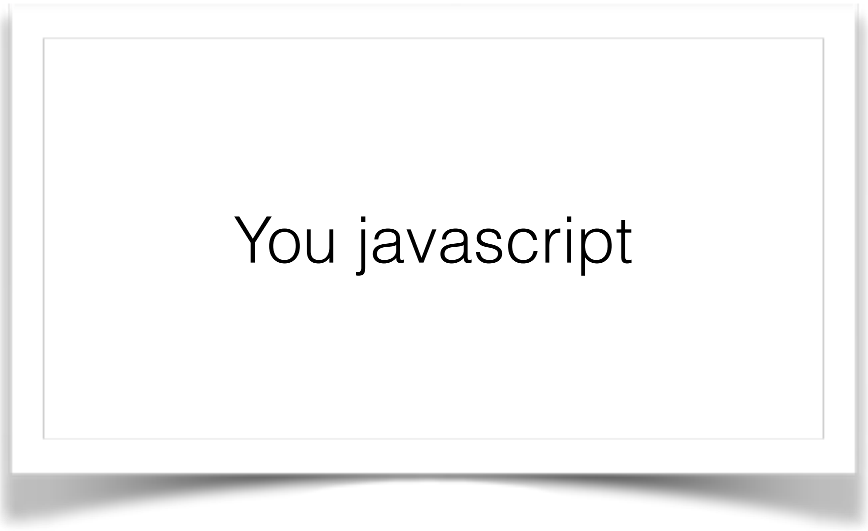
You javascript

2. Sample a word from that topic

Topic Models

Where do documents come from?
"The generative story"

instructions: stencil, instructions, part, step, rubric, handin…
UI: html, javascript, debug, display, elements…
systems: mac, windows, linux, chrome, firefox, os…
fillers: I, you, when, the, and, a



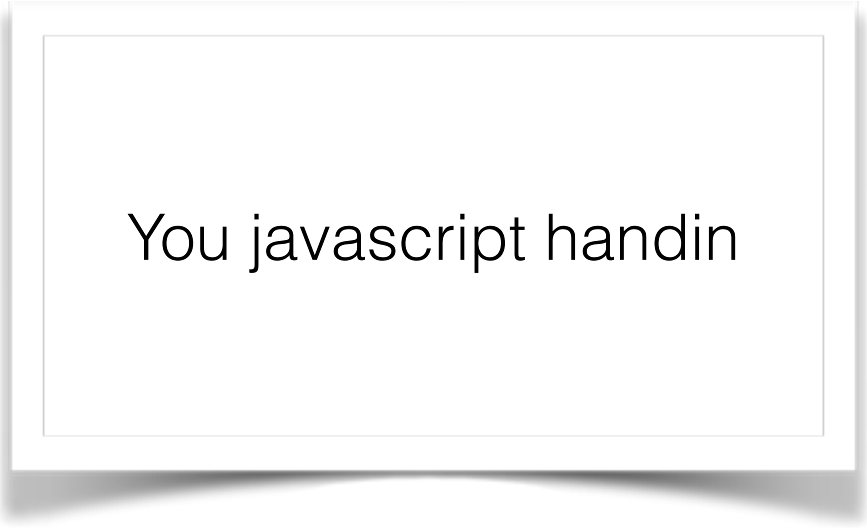
You javascript

1. Sample a topic

Topic Models

Where do documents come from?
“The generative story”

instructions: stencil, instructions, part, step, rubric, handin…
UI: html, javascript, debug, display, elements…
systems: mac, windows, linux, chrome, firefox, os…
fillers: I, you, when, the, and, a



You javascript handin

2. Sample a word from that topic

Topic Models

“Latent Semantic Analysis” (LSA)

$$P(w_i) = \sum_{j=1}^{T} P(w_i \mid z_i = j) P(z_i = j)$$

Topic Models

“Latent Semantic Analysis” (LSA)

$$P(w_i) = \sum_{j=1}^{T} P(w_i \mid z_i = j) P(z_i = j)$$

“latent” variable (not observed)

Topic Models

“Latent Semantic Analysis” (LSA)

$$P(w_i) = \sum_{j=1}^{T} P(w_i \mid z_i = j) P(z_i = j)$$

words are determined by topic
(and are conditionally independent of each other)

Topic Models

“Latent Semantic Analysis” (LSA)

$$P(w_i) = \sum_{j=1}^{T} P(w_i \mid z_i = j) P(z_i = j)$$

documents are a distribution over topics

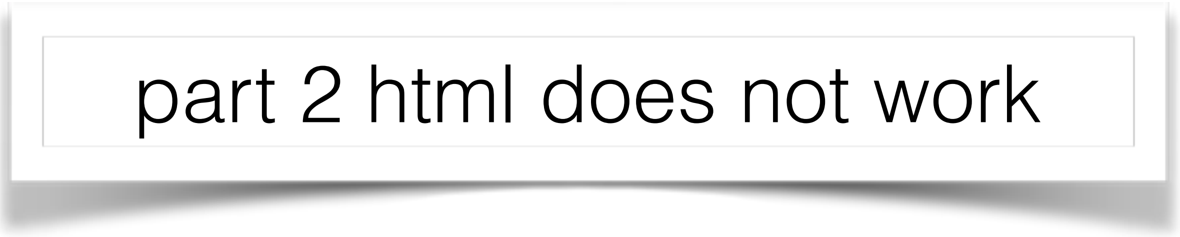
Topic Models

“Latent Semantic Analysis” (LSA)

$$P(w_i) = \sum_{j=1}^{T} P(w_i \mid z_i = j) P(z_i = j)$$

set parameters to maximize probability of observations

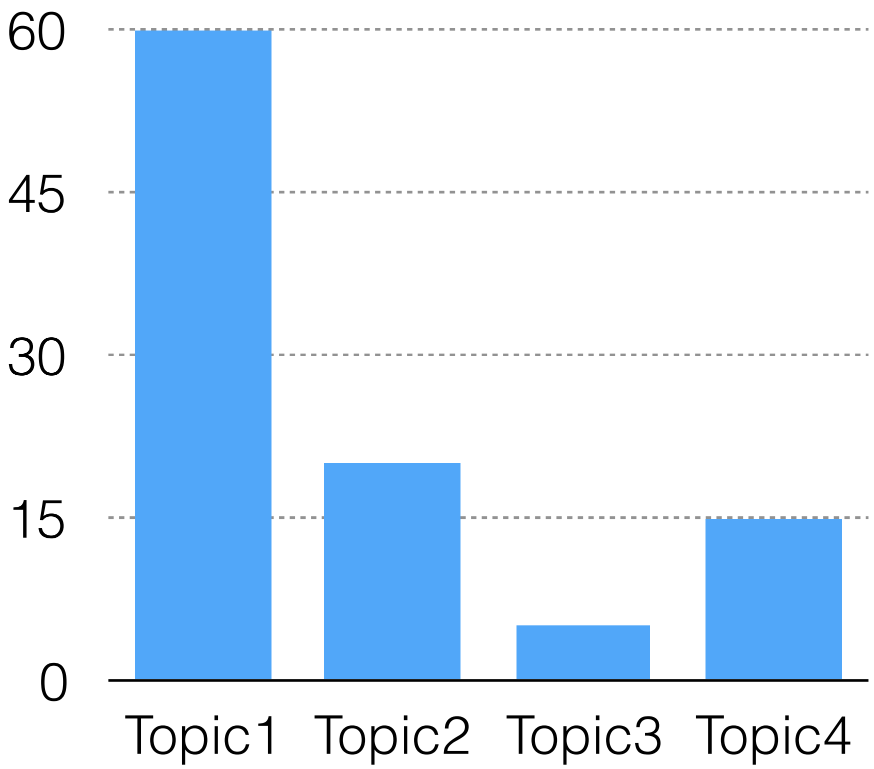
Topic Models



part 2 html does not work

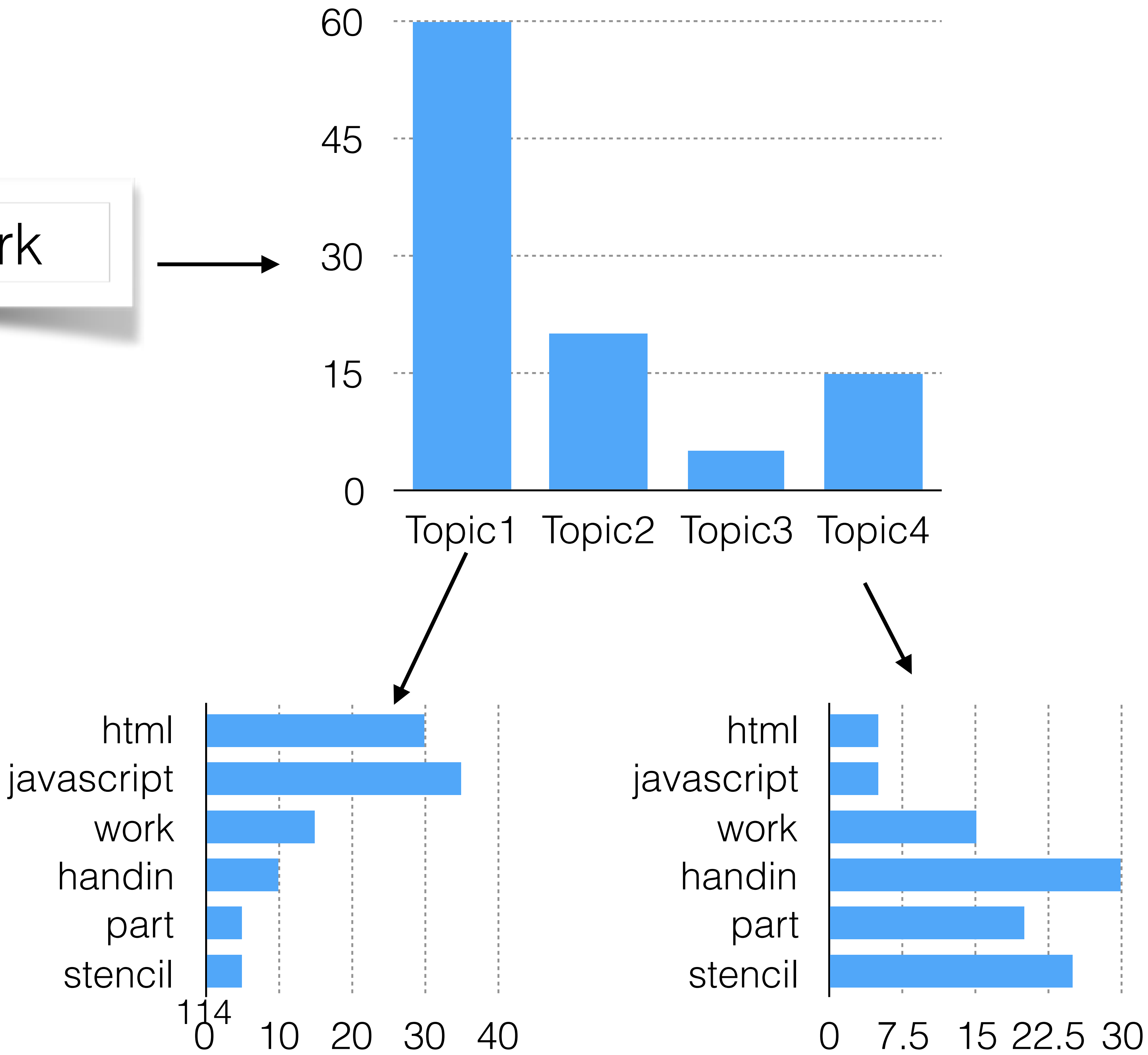
Topic Models

part 2 html does not work



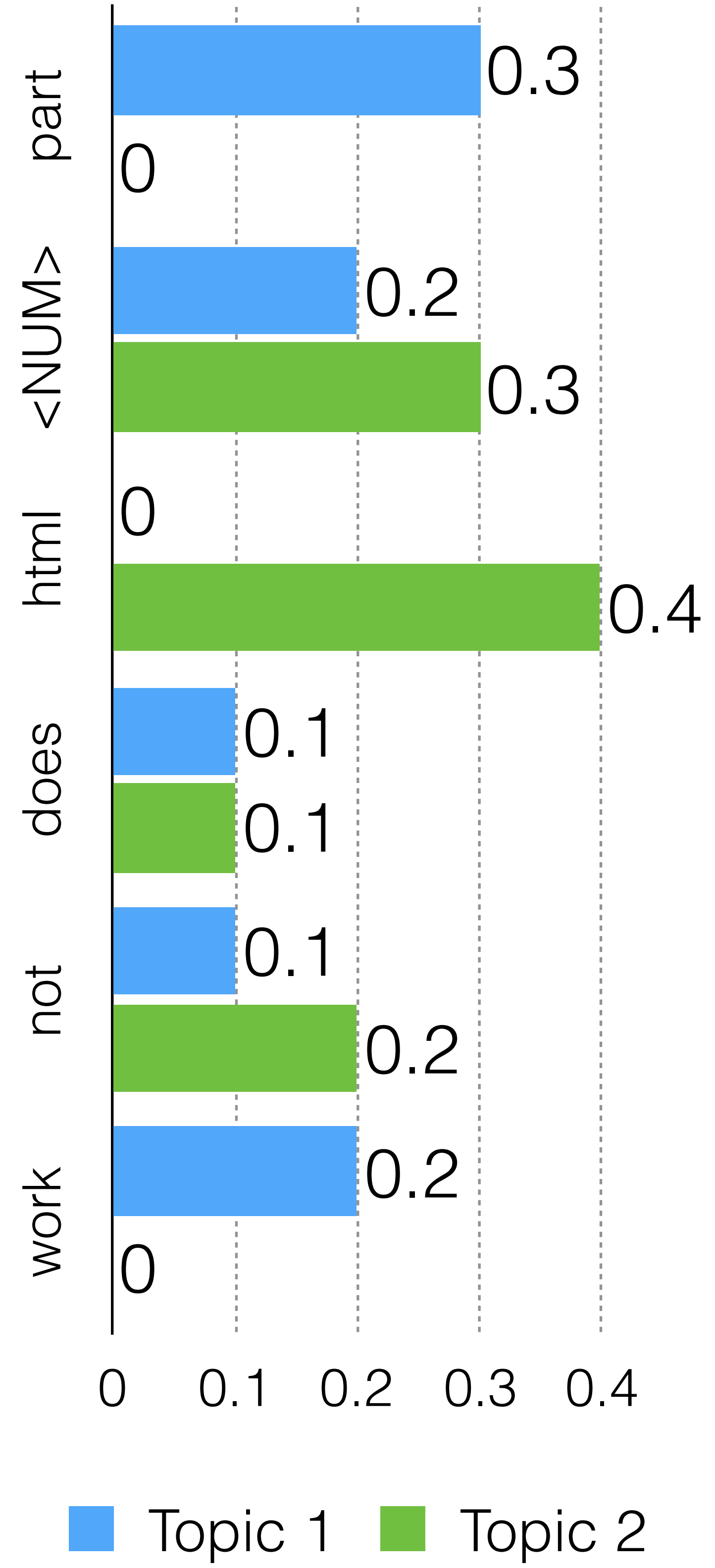
Topic Models

part 2 html does not work



Clicker Question!

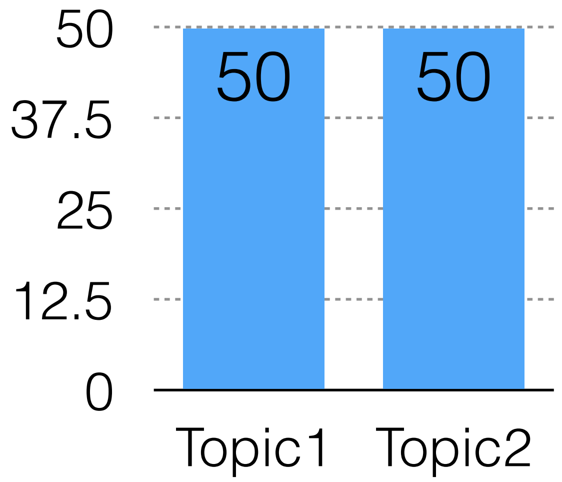
Clicker Question!



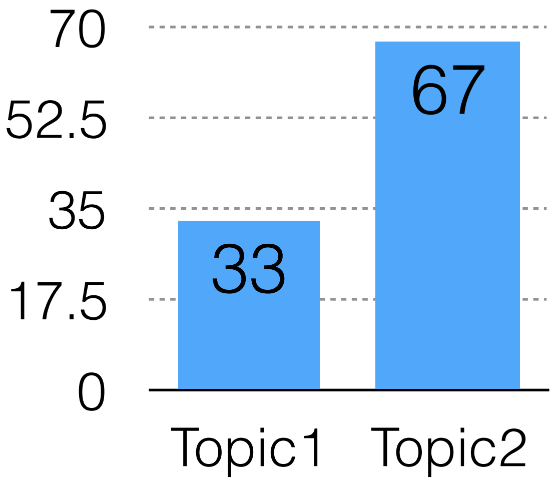
Which is the best parameter setting for the observed data?

$$P(w_i) = \sum_{j=1}^{T} P(w_i \mid z_i = j) P(z_i = j)$$

part <NUM> html does not work

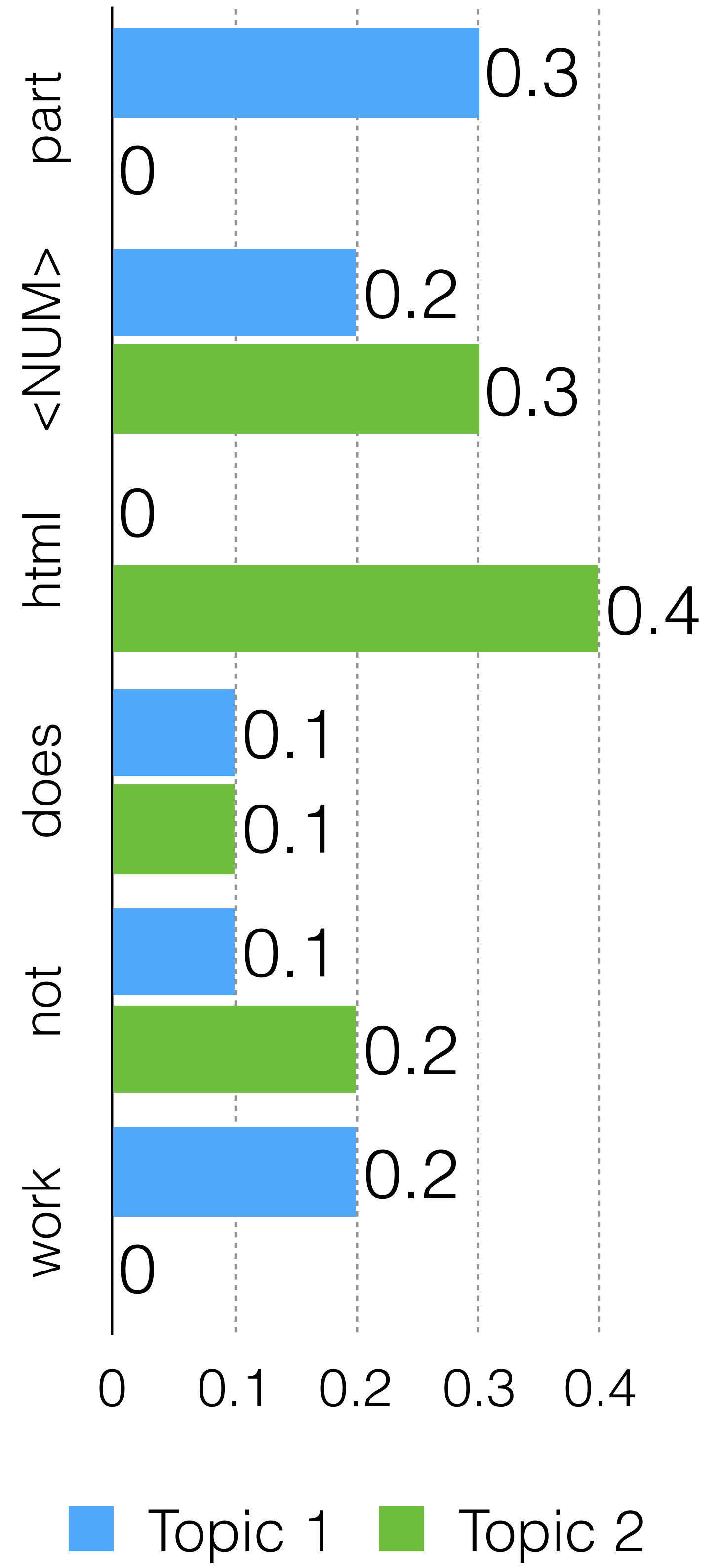


(a)



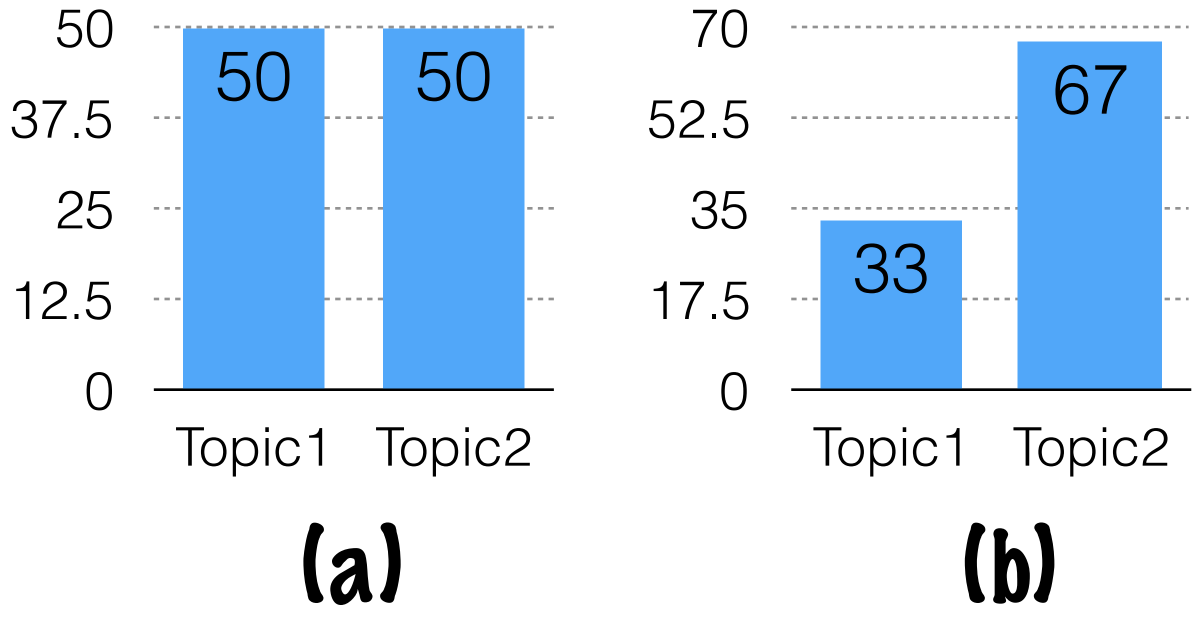
(b)

Clicker Question!

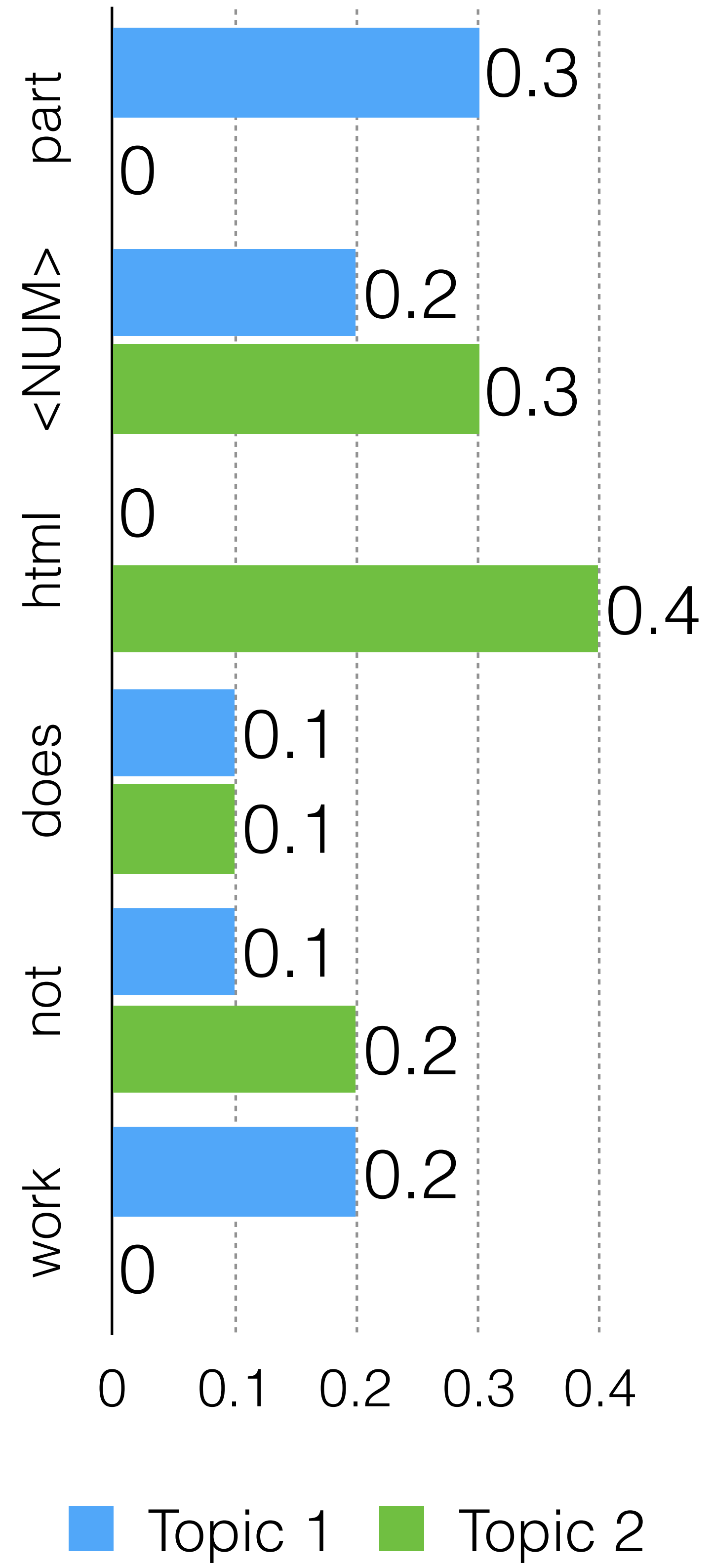


a: $(0.3+0.2+0+0.1+0.1+0.2)\times 0.5$

part <NUM> html does not work



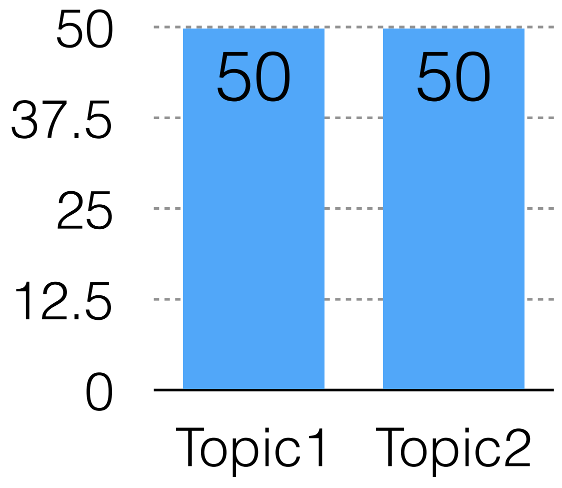
Clicker Question!



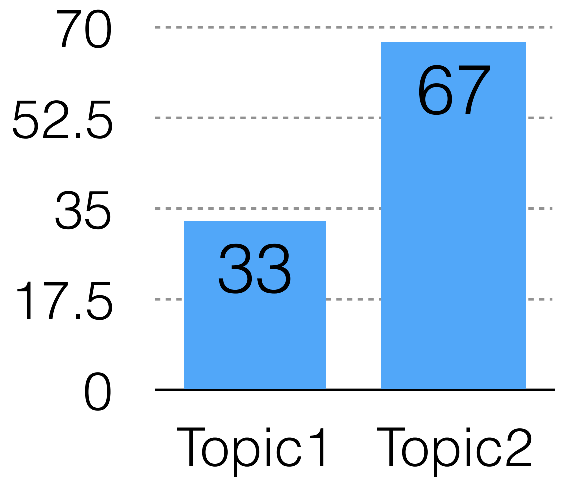
a: (0.3+0.2+0+0.1+0.1+0.2)x0.5

(0+0.3+0.4+0.1+0.2)x0.5

part <NUM> html does not work

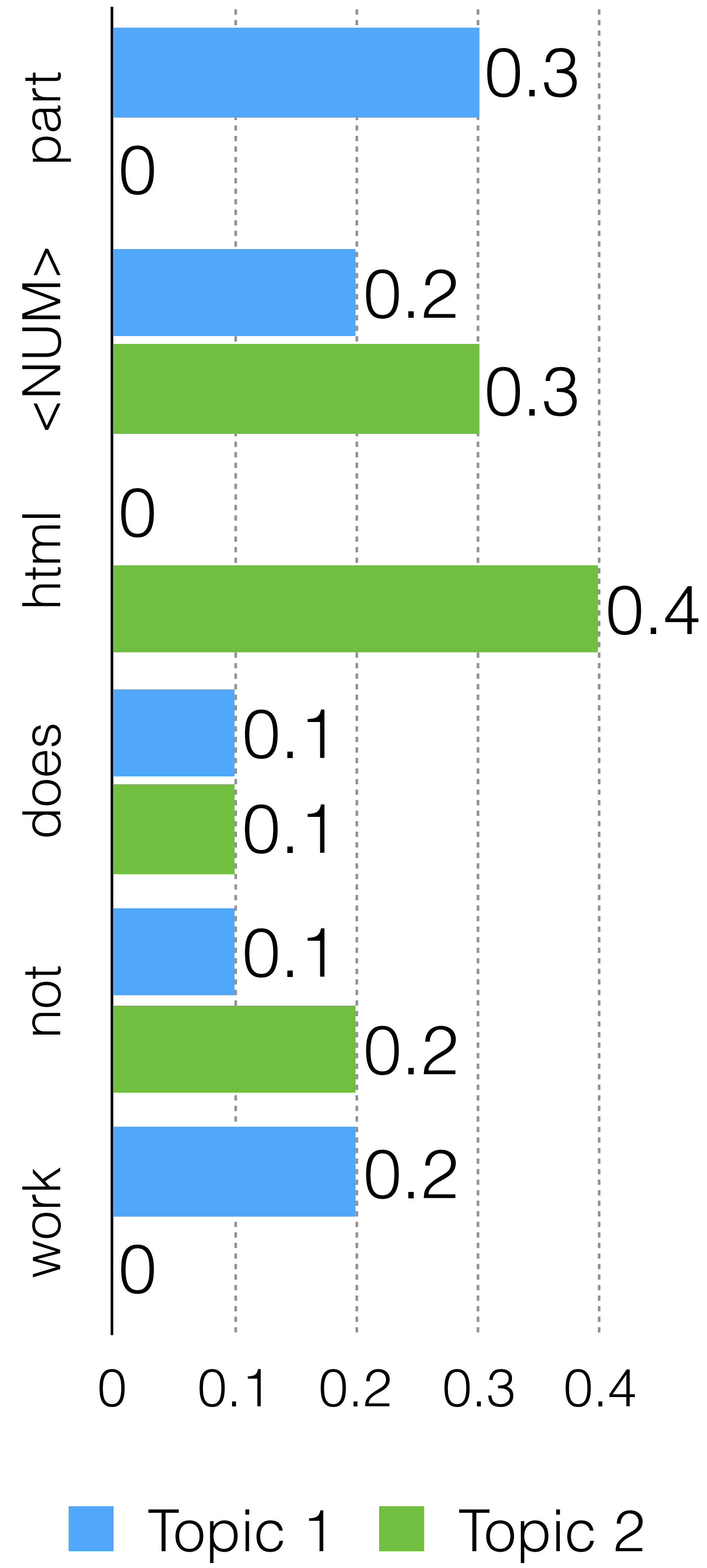


(a)



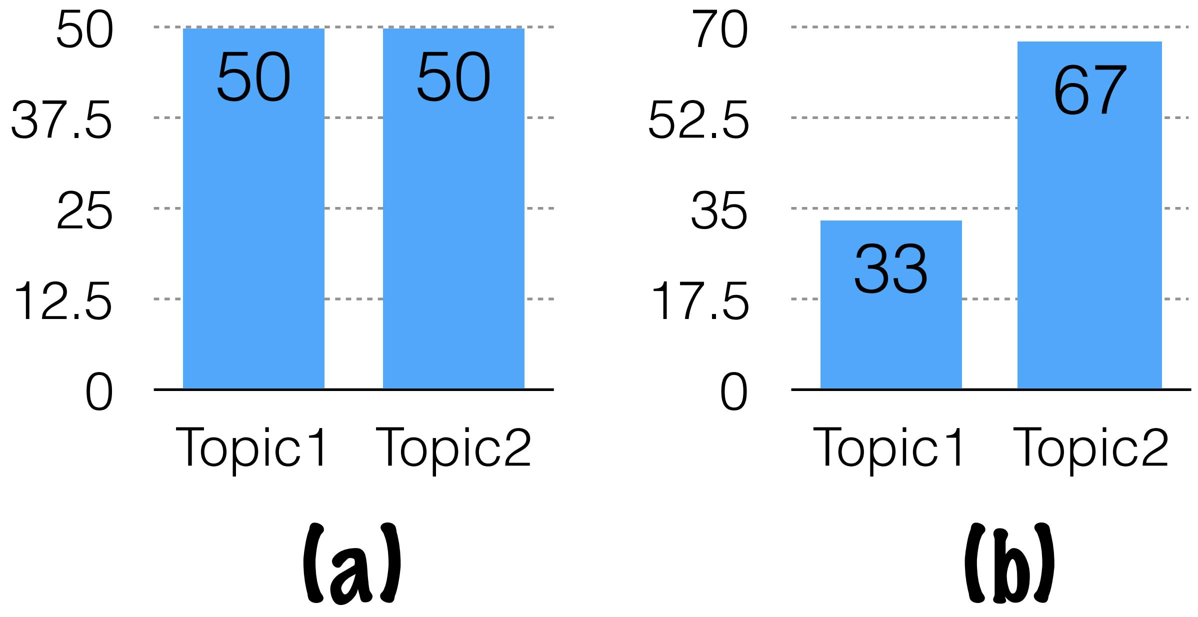
(b)

Clicker Question!

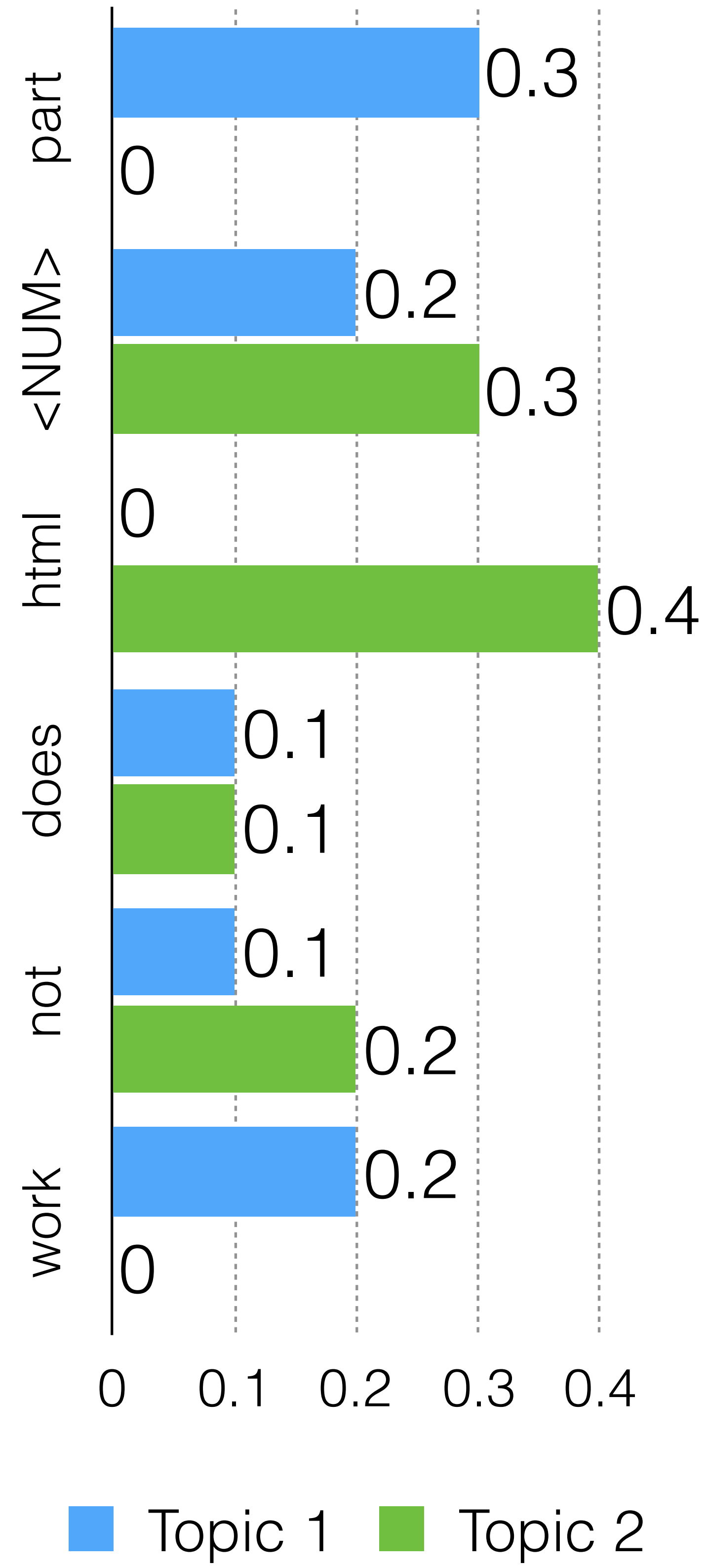


a: (0.3+0.2+0+0.1+0.1+0.2)x0.5
(0+0.3+0.4+0.1+0.2)x0.5
= 0.45 + 0.5
= 0.95

part <NUM> html does not work



Clicker Question!



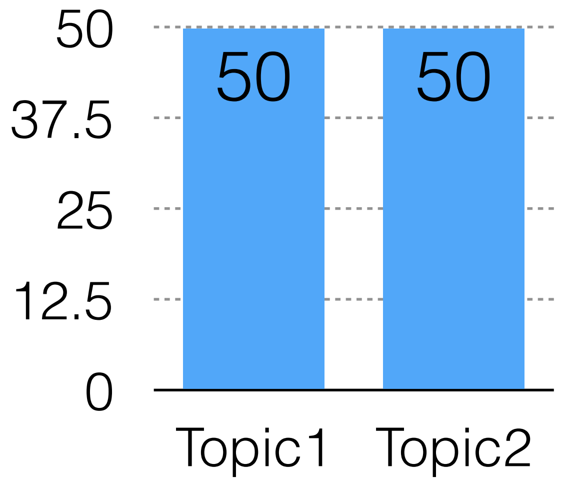
b: (0.3+0.2+0+0.1+0.1+0.2)x0.33

(0+0.3+0.4+0.1+0.2)x0.67

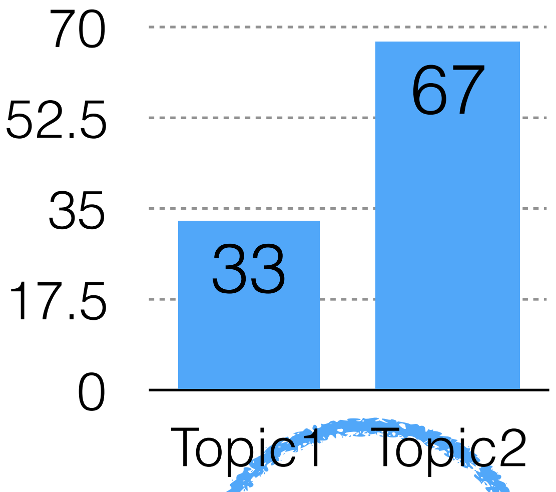
= 0.297 + 0.67

= 0.967

part <NUM> html does not work



(a)



[✓ (b)]

Topic Models

c Models

	the	congress	parliame	US	UK
doc1	1	1	1	1	0
doc2	1	0	1	0	1
doc3	1	1	0	1	0
doc4	1	0	1	0	1

d1	-0.60	-0.39	0.70	0.00
d2	-0.48	0.50	-0.12	-0.71
d3	-0.43	-0.58	-0.69	0.00
d4	-0.48	0.50	-0.12	0.71

U

3.06	0.00	0.00	0.00	0.00
0.00	1.81	0.00	0.00	0.00
0.00	0.00	0.57	0.00	0.00
0.00	0.00	0.00	0.00	0.00

D

the	congress	parliament	US	UK
-0.65	-0.34	-0.51	-0.34	-0.31
0.02	-0.54	0.34	-0.54	0.56
-0.42	0.02	0.79	0.02	-0.44
-0.63	0.27	0.00	0.37	0.63
-0.04	0.73	0.00	-0.68	0.04

V

	the	congress	parliame	US	UK
doc1	1	1	1	1	0
doc2	1	0	1	0	1
doc3	1	1	0	1	0
doc4	1	0	1	0	1

c Models

component = "topic"

d1	-0.60	-0.39	0.70	0.00
d2	-0.48	0.50	-0.12	-0.71
d3	-0.43	-0.58	-0.69	0.00
d4	-0.48	0.50	-0.12	0.71

U

3.06	0.00	0.00	0.00	0.00
0.00	1.81	0.00	0.00	0.00
0.00	0.00	0.57	0.00	0.00
0.00	0.00	0.00	0.00	0.00

D

the	congress	parliament	US	UK
-0.65	-0.34	-0.51	-0.34	-0.31
0.02	-0.54	0.34	-0.54	0.56
-0.42	0.02	0.79	0.02	-0.44
-0.63	0.27	0.00	0.37	0.63
-0.04	0.73	0.00	-0.68	0.04

V

	the	cong ress	parli ame	US	UK
doc1	1	1	1	1	0
doc2	1	0	1	0	1
doc3	1	1	0	1	0
doc4	1	0	1	0	1

c Models

component = "topic" = distribution over words

d1	-0.60	-0.39	0.70	0.00
d2	-0.48	0.50	-0.12	-0.71
d3	-0.43	-0.58	-0.69	0.00
d4	-0.48	0.50	-0.12	0.71

U

3.06	0.00	0.00	0.00	0.00
0.00	1.81	0.00	0.00	0.00
0.00	0.00	0.57	0.00	0.00
0.00	0.00	0.00	0.00	0.00

D

the	cong ress	parlia ment	US	UK
-0.65	-0.34	-0.51	-0.34	-0.31
0.02	-0.54	0.34	-0.54	0.56
-0.42	0.02	0.79	0.02	-0.44
-0.63	0.27	0.00	0.37	0.63
-0.04	0.73	0.00	-0.68	0.04

V

	the	congress	parliame	US	UK
doc1	1	1	1	1	0
doc2	1	0	1	0	1
doc3	1	1	0	1	0
doc4	1	0	1	0	1

c Models

document = distribution over topics

d1	-0.60	-0.39	0.70	0.00
d2	-0.48	0.50	-0.12	-0.71
d3	-0.43	-0.58	-0.69	0.00
d4	-0.48	0.50	-0.12	0.71

U

3.06	0.00	0.00	0.00	0.00
0.00	1.81	0.00	0.00	0.00
0.00	0.00	0.57	0.00	0.00
0.00	0.00	0.00	0.00	0.00

D

the	congress	parliament	US	UK
-0.65	-0.34	-0.51	-0.34	-0.31
0.02	-0.54	0.34	-0.54	0.56
-0.42	0.02	0.79	0.02	-0.44
-0.63	0.27	0.00	0.37	0.63
-0.04	0.73	0.00	-0.68	0.04

V

k bye