Documents as words

Sandeep Soni

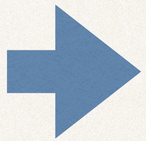
09/11/2023

Class logistics

- PS1 will be released today
- Piazza signup: https://piazza.com/emory/fall2023/qtm340
- Start forming project groups (2-3 students); details about the project submissions updated on class website

Hw3 review

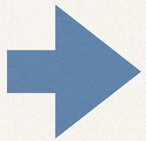
@mark,This is my first tweet after the rebranding,2022-07-23
@roger,Finally, I can start tweeting!,2023-01-11



@mark	This is my first tweet after the rebranding	2022-07-23
@roger	Finally, I can start tweeting!	2023-01-11

Hw3 review

@mark,This is my first tweet after the rebranding,2022-07-23
@roger,Finally, I can start tweeting!,2023-01-11



@mark	This is my first tweet after the rebranding	2022-07-23
@roger	Finally, I can start tweeting!	2023-01-11

Write a regular expression to extract the body of the tweet

Hw3 review

	`@(\w+),(.+),(\d{4}-\d{2}-\d{2})`	`@[^,]+,(.*?),\d{4}-\d{2}-\d{2}`	`^(@\w+),\s*(.*),\s*(\d{4}\-\d{2}\-\d{2})$`
`@mark,This is my first tweet after the rebranding,2022-07-23`			
`@roger,Finally, I can start tweeting!,2023-01-11`			
`@barack,.@potus Happy new year! #newyear, 2023-01-01`			

Question for the day

"How to represent a document as a function of its words?"

Agenda

- How to represent documents by their words?
- What are the variations of this representation?

Vocabulary

Vocabulary

You know my methods, Watson

To Sherlock Holmes she is always *the* woman

"Excellent! I cried. "Elementary," said he.

Vocabulary

You know my methods, Watson

To Sherlock Holmes she is always *the* woman

"Excellent! I cried. "Elementary," said he.

You
Sherlock
Holmes
methods
Watson
,
!
I
…
…
said
he

Vocabulary

You know my methods, Watson

To Sherlock Holmes she is always *the* woman

"Excellent! I cried. "Elementary," said he.

```
import spacy
spacy.load (“en_core_web_sm”)
vocab ={token.text for token in nlp (text)}
```

You
Sherlock
Holmes
methods
Watson
,
!
I
…
…
said
he

Vocabulary

You know my methods, Watson

To Sherlock Holmes she is always *the* woman

"Excellent! I cried. "Elementary," said he.

```
import spacy
spacy.load (“en_core_web_sm”)
vocab ={token.text for token in nlp (text)}
```

You
Sherlock
Holmes
methods
Watson
,
!
I
…
…
said
he

You may want to preprocess the text (lowercasing, lemmatizing, punctuation removal) but you can construct a lexicon or vocabulary from the corpus

How can we represent documents?

How can we represent documents?

- For some applications, we may not need all the linguistic structure for our analysis

How can we represent documents?

- For some applications, we may not need all the linguistic structure for our analysis
- e.g. Is a research paper from Emory or Georgia Tech?

How can we represent documents?

- For some applications, we may not need all the linguistic structure for our analysis
- e.g. Is a research paper from Emory or Georgia Tech?
- Can I get away with losing syntactic and semantic info. from text?

Representing documents

Representing documents

To Sherlock Holmes
she is
always *the* woman

Document

Representing documents

To Sherlock Holmes
she is
always *the* woman

Document

To
Sherlock
Holmes
she
is
always
the
woman

Vocabulary of document

Representing documents

- Use the vocabulary of a given document as its representation
- We have thrown away information about the position and ordering of the words in the document

To Sherlock Holmes
she is
always *the* woman

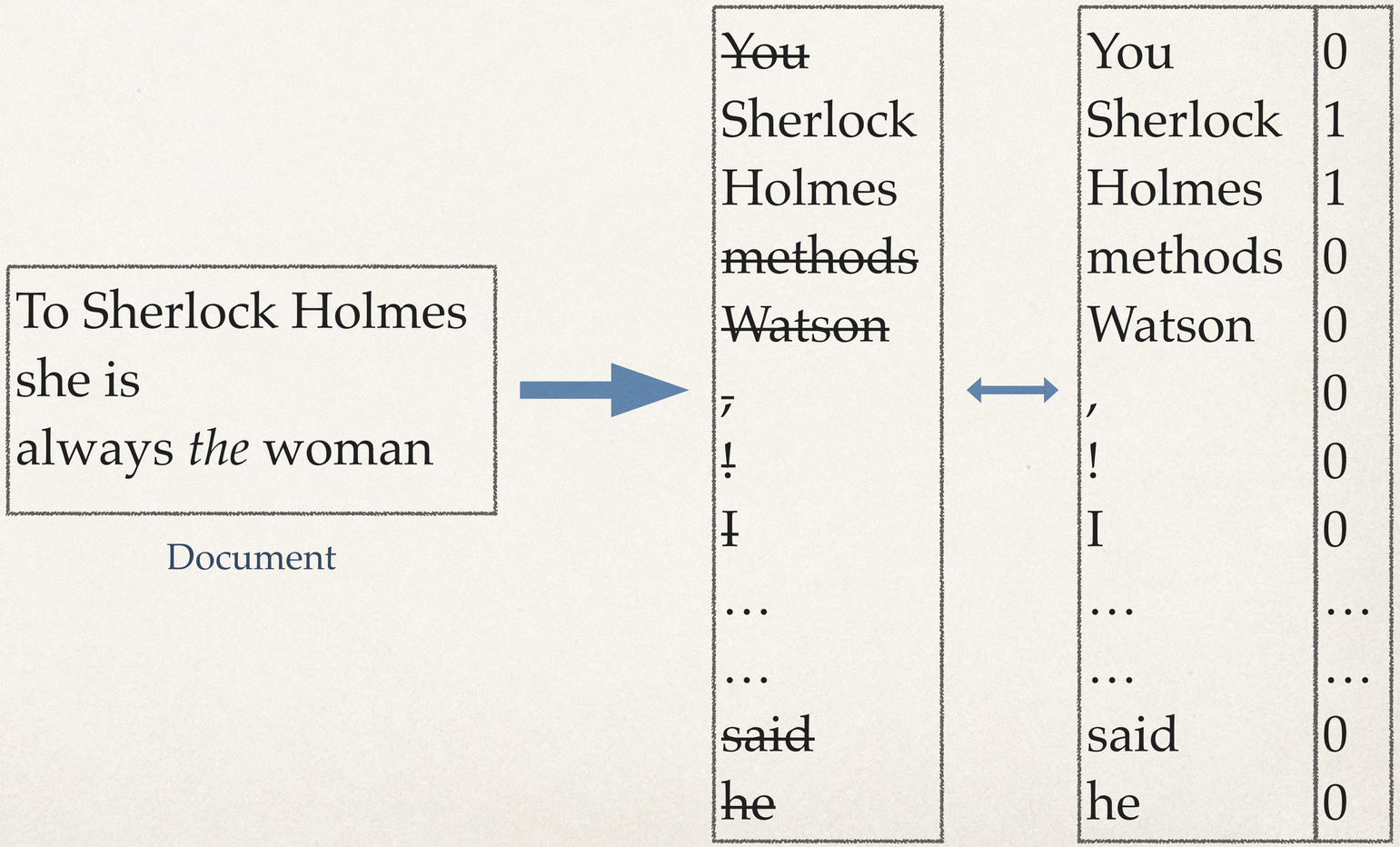
Document

To
Sherlock
Holmes
she
is
always
the
woman

Vocabulary of document

Representing documents

- You can represent a document with a shared vocabulary



Representing documents

- What if words occur more than once in a document?

To Sherlock Holmes she is always *the* woman. I have seldom heard him mention her under any other name. In his eyes she eclipses and predominates the whole of her sex.

You	0
Sherlock	1
Holmes	1
methods	0
Watson	0
the	1
!	0
I	1
…	…
…	…
said	0
he	0

Document

Representing documents

- What if words occur more than once in a document?
- Indicate the count of any word within a document

To Sherlock Holmes she is always *the* woman. I have seldom heard him mention her under any other name. In his eyes she eclipses and predominates the whole of her sex.

Document

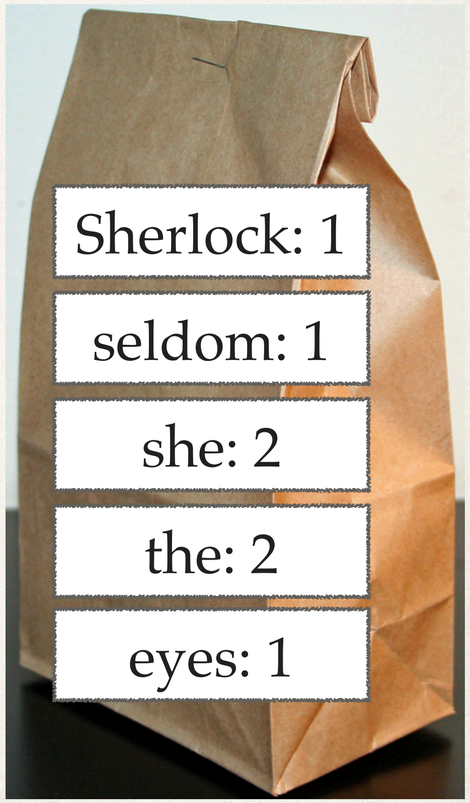
You	0
Sherlock	1
Holmes	1
methods	0
Watson	0
the	2
!	0
I	1
…	…
…	…
said	0
he	0

Document as a bag of words!

- This is the bag of words representation of a document

To Sherlock Holmes she is always *the* woman. I have seldom heard him mention her under any other name. In his eyes she eclipses and predominates the whole of her sex.

Document



Exercise

- Suppose V={I, am, fine}
- Assume that we convert all documents to lowercase
- Ignore punctuation

Doc	i	am	fine
Fine			
I am fine			
I am fine, alright			
Fine I am			
I am fine. I AM FINE. FINE			

Exercise

- Suppose V={I, am, fine}
- Assume that we convert all documents to lowercase
- Ignore punctuation

Doc	i	am	fine
Fine	0	0	1
I am fine	1	1	1
I am fine, alright	1	1	1
Fine I am	1	1	1
I am fine. I AM FINE. FINE	2	2	3

Exercise

- This is the document-term matrix
- Every document is a vector; words are dimensions.
- Out of vocabulary words are not represented in this scheme

Doc	i	am	fine
Fine	0	0	1
I am fine	1	1	1
I am fine, alright	1	1	1
Fine I am	1	1	1
I am fine. I AM FINE. FINE	2	2	3

Applications: language identification

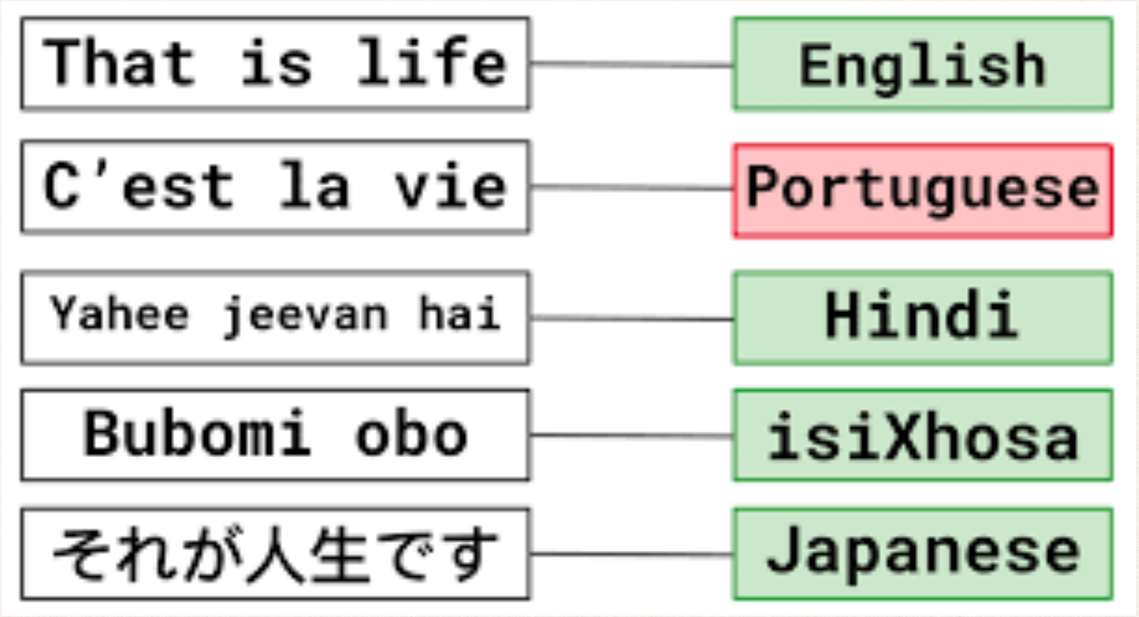


Image credit Connor O' Sullivan

Applications: language identification

- Input:
  Document as
  bag of words

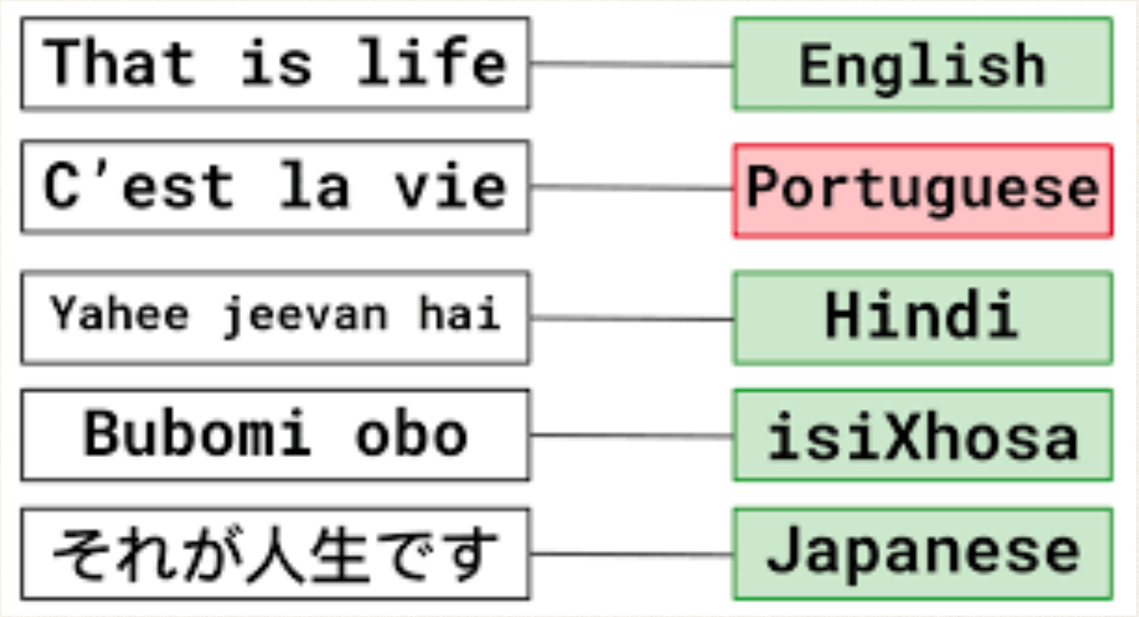


Image credit Connor O' Sullivan

Applications: language identification

- Input:
  Document as
  bag of words
- Output: Id of
  the language

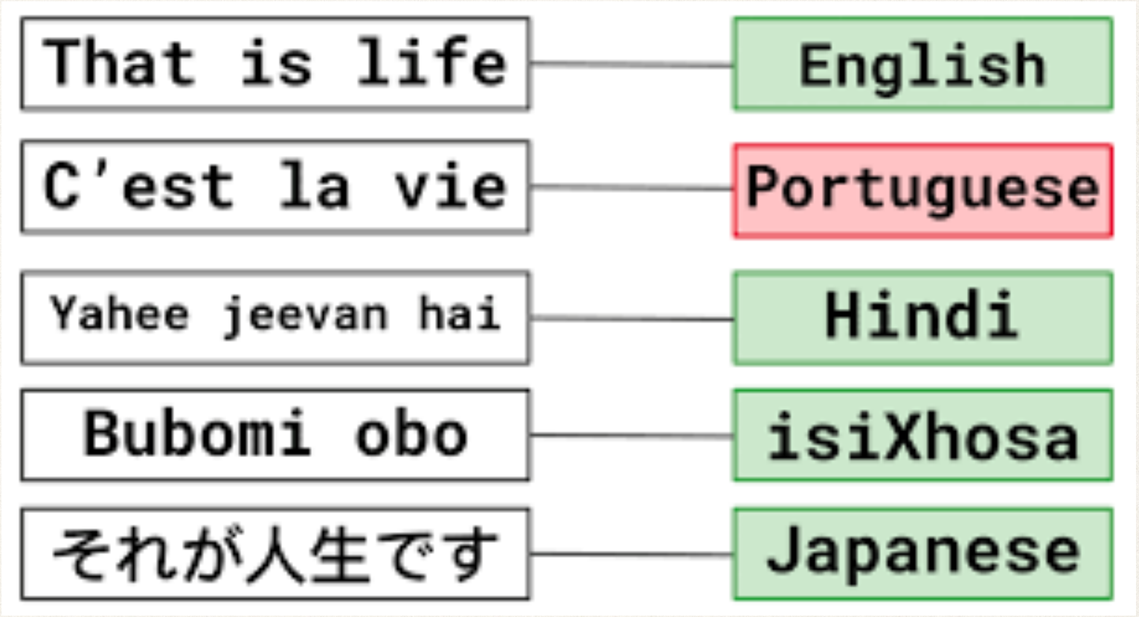


Image credit Connor O' Sullivan

Applications: language identification

- Input:
  Document as
  bag of words
- Output: Id of
  the language

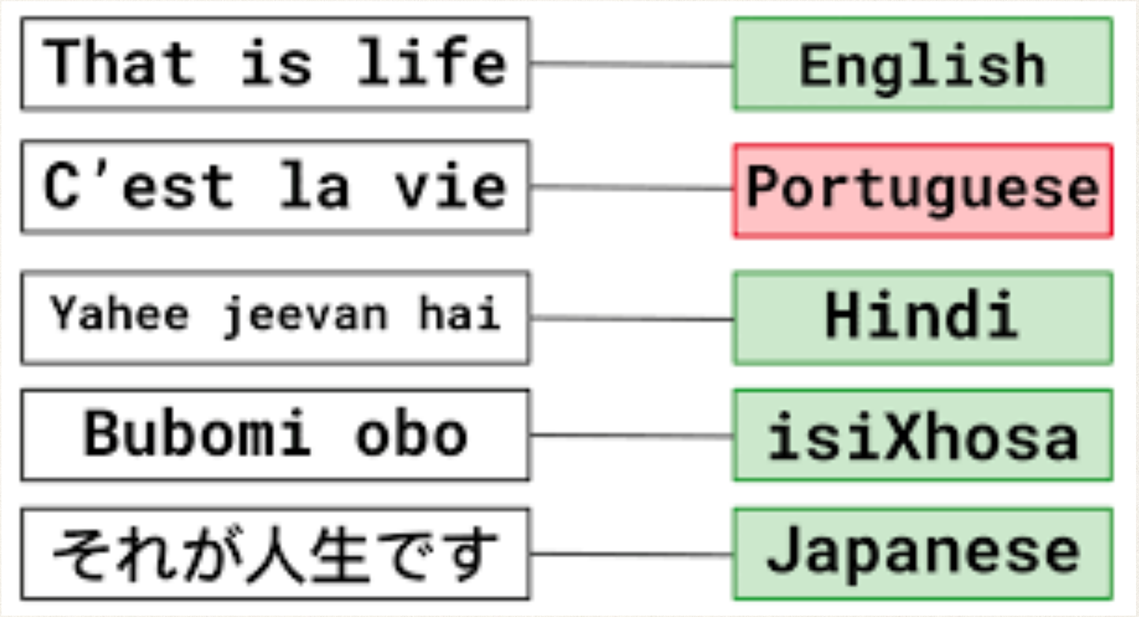


Image credit Connor O' Sullivan

C. Huang and L. Lee, "Contrastive approach towards text source classification based on top-bag-of-word similarity," in Proceedings of PACLIC 2008, 2008, pp. 404–410.

Relative Frequency

the	i	…	…	singer	run
0.23	0.18	…	…	0.04	0.08
0.18	0.34	…	…	0.09	0.03
0.1	0.08	…	…	0	0.01

- Instead of raw counts, we can calculate proportion of occurrences over all the counts

Word importance

- How important is a given word to the document?
- A word is important if:
  - It frequently occurs in a document
  - Occurs only in a small set of other documents

Word importance

- How important is a given word to the document?
- A word is important if:
  - It frequently occurs in a document (term frequency)
  - Occurs only in a small set of other documents (inverse document frequency)

TF-IDF

- Term frequency (TF) of word w: Relative frequency of word in a document
- Inverse document frequency (IDF) of word w: Total documents in corpus divided by number of documents in which w occurs
- TF-IDF (w, d) = TF (w,d) * IDF (w,d)

TF and IDF Variations

Variants of term frequency (tf) weight

weighting scheme	tf weight
binary	$0, 1$
raw count	$f_{t,d}$
term frequency	$f_{t,d} \Big/ \sum_{t' \in d} f_{t',d}$
log normalization	$\log(1 + f_{t,d})$
double normalization 0.5	$0.5 + 0.5 \cdot \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$
double normalization K	$K + (1 - K)\frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$

Source: Wikipedia

Variants of inverse document frequency (idf) weight

weighting scheme	idf weight ($n_t = \lvert\{d \in D : t \in d\}\rvert$)
unary	1
inverse document frequency	$\log \frac{N}{n_t} = -\log \frac{n_t}{N}$
inverse document frequency smooth	$\log\left(\frac{N}{1+n_t}\right) + 1$
inverse document frequency max	$\log\left(\frac{\max_{\{t' \in d\}} n_{t'}}{1+n_t}\right)$
probabilistic inverse document frequency	$\log \frac{N - n_t}{n_t}$

Source: Wikipedia

TF-IDF Variations

Variants of term frequency-inverse document frequency (tf–idf) weights

weighting scheme	tf-idf
count-idf	$f_{t,d} \cdot \log \frac{N}{n_t}$
double normalization-idf	$\left(0.5 + 0.5 \frac{f_{t,q}}{\max_t f_{t,q}}\right) \cdot \log \frac{N}{n_t}$
log normalization-idf	$(1 + \log f_{t,d}) \cdot \log \frac{N}{n_t}$

Source: Wikipedia

In class

- Bag of words demo