Jul – Nov 2022

IT458

Classical IR Models for Unstructured Text

Bag of Words (BoW)

Dr. Sowmya Kamath S, Dept of IT, NITK Surathkal

10-Aug-22

Bag of Words Representation (BoW)

- A bag-of-words is a representation of text that describes the occurrence of words within a document.

- Consists of
  - A vocabulary of known words.
  - A measure of the presence of known words.

Bag of Words Representation

- In BoW,
  - A document is typically represented by a *bag of words* -
    - Bag = a *set* that allows multiple occurrences of the same element.
    - unordered words with frequencies.

Bag of Words Representation

- User can specify a set of desired terms with optional weights:
  - Weighted query terms:
    - E.g. **Q = < database 0.5; text 0.3; information 0.2 >**
  - Unweighted query terms (but order matters):
    - E.g. **Q = < database; text; information >**

Bag of Words Representation

- Standard case:
  - Vocabulary (bag) = set of all the words occurring in the collection's documents.
  - Each document is represented by the words it contains.

Bag of Words Representation

- Standard case:
  - Vocabulary (*bag*) = set of all the words occurring in the collection's documents
  - Each document is represented by the words it contains.

That's one small step for a man, a giant leap for mankind → { that's, one, small, step, for (2), a (2), man, giant, leap, mankind }

Bag of Words Representation

- Each document in the collection can be represented by an **incidence vector**, for matching.

vocabulary (index terms) →

	that's	one	small	step	for	a	man	giant	leap	mankind	Gandhi's	was	turning	point	India
That's one small step for a man, a giant leap for mankind	1	1	1	1	2	2	1	1	1	1	0	0	0	0	0
Gandhi's small step was a turning point for India	0	0	1	1	1	1	0	0	0	0	1	1	1	1	1

incidence matrix
(term-document matrix)

Bag of Words Representation

- Each document in the collection can be represented by an **incidence vector**, for matching.

vocabulary (index terms) →

	that's	one	small	step	for	a	man	giant	leap	mankind	Gandhi's	was	turning	point	India
That's one small step for a man, a giant leap for mankind	1	1	1	1	2	2	1	1	1	1	0	0	0	0	0
Gandhi's small step was a turning point for India	0	0	1	1	1	1	0	0	0	0	1	1	1	1	1

incidence matrix
(term-document matrix)

Sentence 1 = [1,1,1,1,2,2,1,1,1,1,0,0,0,0,0]
Sentence 2 = [0,0,1,1,1,1,0,0,0,0,1,1,1,1,1]

.
.
.

Bag of Words Representation

- Major issues:
  - As the vocabulary size increases, sparseness increases.
  - BoW is an orderless document representation

Bag of Words Representation

Dealing with sparseness

- Simple text cleaning techniques that can be used for preprocessing :
  - Ignoring case
  - Ignoring punctuation
  - Ignoring frequent words that don't contain much information
    - called stop words, like "a," "of," etc.
  - Fixing misspelled words.
  - Reducing words to their stem.
    - e.g. "play" from "playing", "played", "plays" using stemming algorithms.

Bag of Words Representation

- Solution: different ways for word order representation in the vocabulary

- Options -
  - Unigram
  - N-gram – bigram, trigram…
    - N = 2,3,4….

- N-gram BoW model - used to store the spatial information within the text.

Bag of Words Representation

- *Example*: a **unigram** Bag of Words Representation

```
Doc1 = [ "That's", "one", "small", "step", "for":2,
"a":2, "man", "giant", "leap", "for", "mankind"]
```

- *Example*: a **bigram** Bag of Words Representation

```
Doc1 = [ "That's one", "one small", "small step", "step
for", "for a", "a man", "man a", "a giant", "giant
leap", "leap for", "for mankind"]
```

Bag of Words - Applications

- Spam filtering.

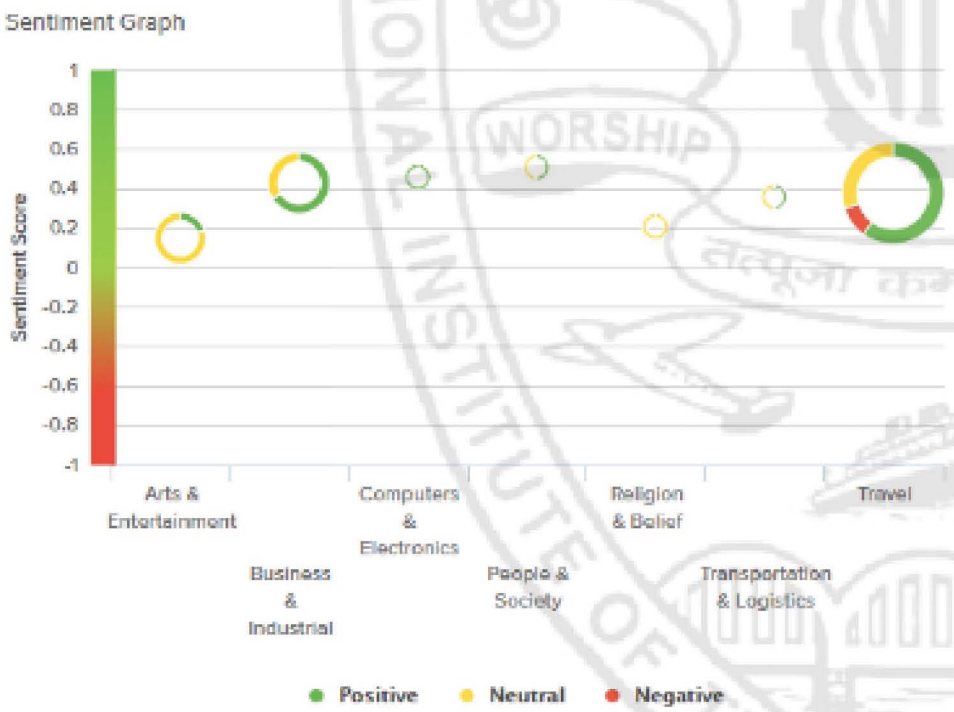
- E-mail messages modeled as an unordered collection of words.
- Use probability distributions to determine which bag a new incoming email is more likely to be.

Bag of Words - Applications

- Topic Analysis/Sentiment Analysis/Profanity detection

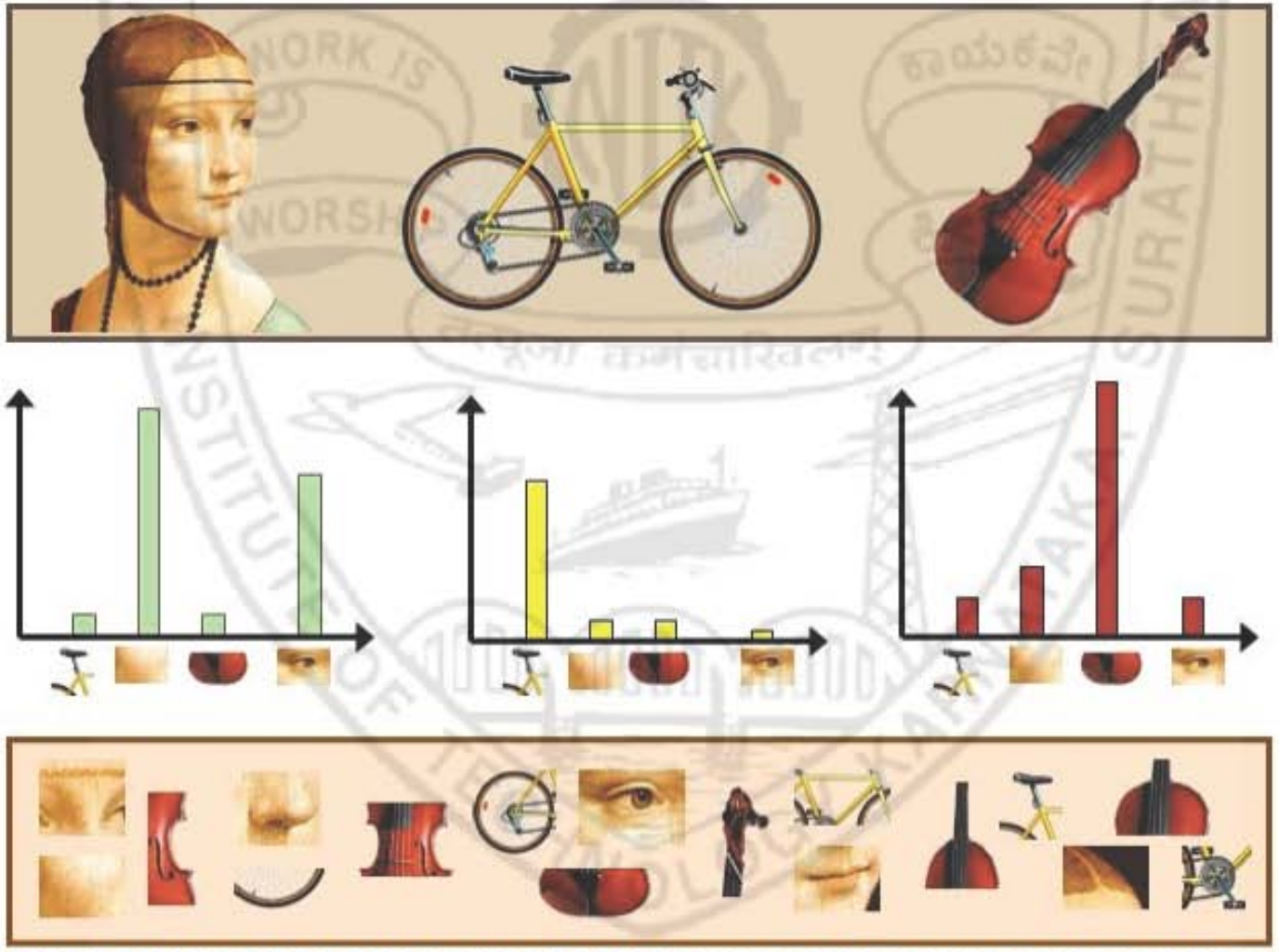
Bag of Words - Applications

- Topic Analysis / Sentiment Analysis / Profanity detection



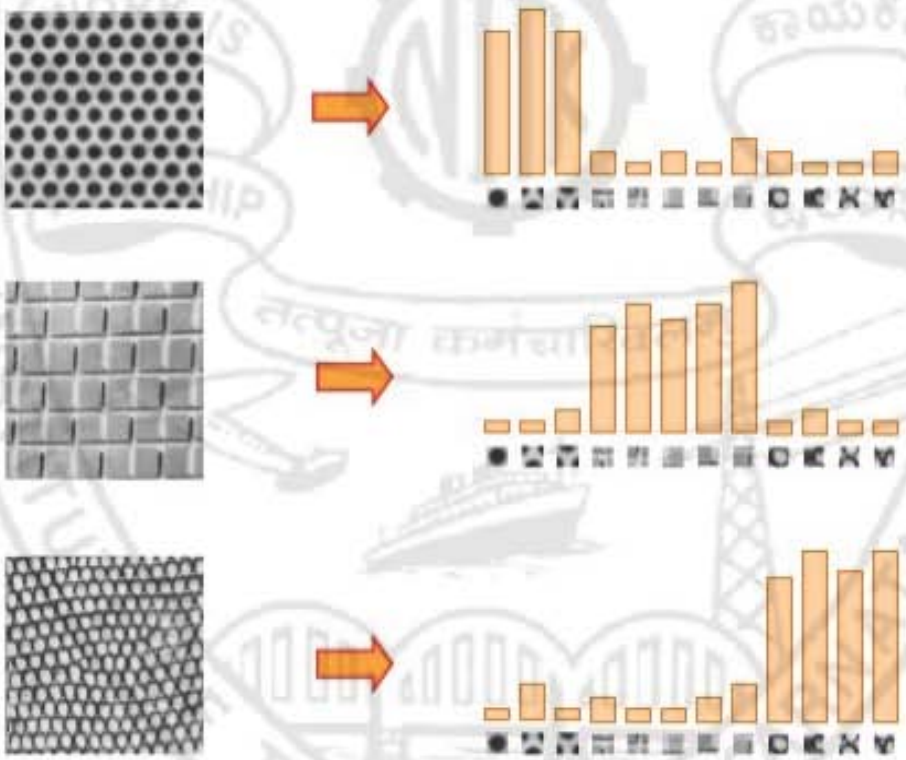
Bag of Words - Applications

- In Computer Vision Object Recognition, Categorization.



Bag of Words - Applications

- In Computer Vision Texture Modeling.



Bag of Words Representation

- Pros:
  - Simple set-theoretic representation of documents
  - Efficient storage and retrieval of individual terms
  - well-suited for some specialized, small-scale applications.
  - Very popular as a text vectorizer for ML applications

Bag of Words Representation

- Cons:
  - As the vocabulary size increases, sparseness increases.
  - Word order gets lost
  - Very different documents could have similar representations
  - Document structure (e.g. headings) and metadata is ignored

More reading...

- Harris, Zellig S. "Distributional structure." *Word* 10.2-3 (1954): 146-162.
- Rosenfeld, A., Huang, H., & Schneider, V. (1969). An application of cluster detection to text and picture processing. *IEEE Transactions on Information theory*, *15*(6), 672-681.
- Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive science*, *22*(4), 425-469.