Big Data Computing

Master's Degree in Computer Science

2020-2021

Gabriele Tolomei

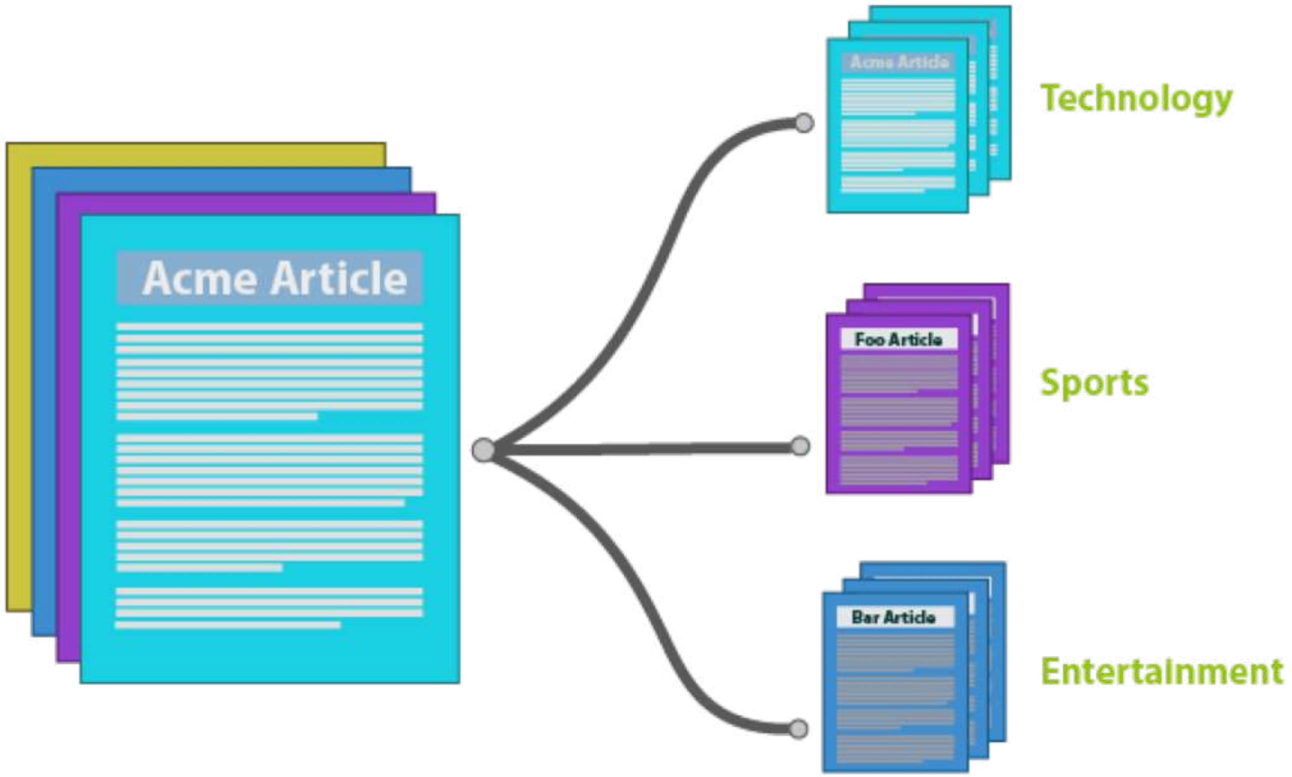
Department of Computer Science

Sapienza Università di Roma

tolomei@di.uniroma1.it

SAPIENZA
UNIVERSITÀ DI ROMA

Our Running Example: Document Clustering



source: https://towardsdatascience.com/applying-machine-learning-to-classify-an-unsupervised-text-document-e7bb6265f52

Our Running Example: Document Clustering

- **Problem:** Group together documents on the same **topic**

Our Running Example: Document Clustering

- Problem: Group together documents on the same topic
- Documents with similar sets of words may be about the same topic

Our Running Example: Document Clustering

- Problem: Group together documents on the same topic
- Documents with similar sets of words may be about the same topic
- Key Issues:
  - Representing documents (in the space of words)
  - Measuring document similarity (in the space of words)

Our Running Example: Document Clustering

- Problem: Group together documents on the same topic
- Documents with similar sets of words may be about the same topic
- Key Issues:
  - Representing documents (in the space of words)
  - Measuring document similarity (in the space of words)

NOTE

A dual problem is topic clustering, where topics (i.e., set of words co-occurring in many documents) are clustered within the space of documents

Document Representation

- Different ways of representing documents (in the space of words)

Document Representation

- Different ways of representing documents (in the space of words)
  - As a set of words (disregarding the order and multiplicity)

Document Representation

- Different ways of representing documents (in the space of words)
  - As a set of words (disregarding the order and multiplicity)
  - As a bag-of-words (i.e., a multiset disregarding the order yet keeping multiplicity)

Document Representation

- Different ways of representing documents (in the space of words)
  - As a set of words (disregarding the order and multiplicity)
  - As a bag-of-words (i.e., a multiset disregarding the order yet keeping multiplicity)
  - As a bag-of-*n*-grams (i.e., the more general case of bag-of-words)

Document Representation

- Different ways of representing documents (in the space of words)
  - As a set of words (disregarding the order and multiplicity)
  - As a bag-of-words (i.e., a multiset disregarding the order yet keeping multiplicity)
  - As a bag-of-*n*-grams (i.e., the more general case of bag-of-words)
  - More advanced representations derived from Neural Language Models (e.g., word2vec)

Document Representation

- Different ways of representing documents (in the space of words)
  - As a set of words (disregarding the order and multiplicity)
  - As a bag-of-words (i.e., a multiset disregarding the order yet keeping multiplicity)
  - As a bag-of-*n*-grams (i.e., the more general case of bag-of-words)
  - More advanced representations derived from Neural Language Models (e.g., word2vec)
- The choice of document representation affects the similarity measure

Document Representation: Set of Words

doc 1

John likes to watch movies. Mary likes movies too.

doc 2

Mary also likes to watch football games.

Document Representation: Set of Words

doc 1

```
John likes to
watch movies.
Mary likes
movies too.
```

doc 2

```
Mary also likes
to watch
football games.
```

{John, likes, to, watch, movies, Mary, too}

{Mary, also, likes, to, watch, football, games}

Document Representation: Bag-of-Words

We keep multiplicity

doc 1

```
John likes to
watch movies.
Mary likes
movies too.
```

doc 2

```
Mary also likes
to watch
football games.
```

Document Representation: Bag-of-Words

We keep multiplicity

doc 1

```
John likes to
watch movies.
Mary likes
movies too.
```

{
John:1, likes:2, to:1, watch:1,
movies:2, Mary:1, too:1
}

doc 2

```
Mary also likes
to watch
football games.
```

{
Mary:1, also:1, likes:1, to:1,
watch:1, football:1, games:1
}

Document Representation: Bag-of-Words

Bag-of-Words (BoW) model is just a preliminary step for more complex document representations

Document Representation: Bag-of-Words

Bag-of-Words (BoW) model is just a preliminary step for more complex document representations

2 key components

vocabulary

weighting scheme

Bag-of-Words: Vocabulary

doc 1

```
John likes to
watch movies.
Mary likes
movies too.
```

doc 2

```
Mary also likes
to watch
football games.
```

Bag-of-Words: Vocabulary

doc 1

```
John likes to
watch movies.
Mary likes
movies too.
```

doc 2

```
Mary also likes
to watch
football games.
```

V = {also, football, games, John, likes, Mary, movies, to, too, watch}

Bag-of-Words: Weighting Scheme

doc 1

```
John likes to
watch movies.
Mary likes
movies too.
```

A |V|-dimensional vector, where the *i*-th component indicates the multiplicity of the *i*-th word of the vocabulary

V = {also, football, games, John, likes, Mary, movies, to, too, watch}

Bag-of-Words: Weighting Scheme

doc 1

```
John likes to
watch movies.
Mary likes
movies too.
```

A |V|-dimensional vector, where the *i*-th component indicates the multiplicity of the *i*-th word of the vocabulary

(0, 0, 0, 1, 2, 1, 2, 1, 1, 1)

V = {also, football, games, John, likes, Mary, movies, to, too, watch}

Bag-of-Words: Weighting Scheme

doc 2

```
Mary also likes
to watch
football games.
```

A |V|-dimensional vector, where the *i*-th component indicates the multiplicity of the *i*-th word of the vocabulary

V = {also, football, games, John, likes, Mary, movies, to, too, watch}

Bag-of-Words: Weighting Scheme

doc 2

```
Mary also likes
to watch
football games.
```

A |V|-dimensional vector, where the *i*-th component indicates the multiplicity of the *i*-th word of the vocabulary

(1, 1, 1, 0, 1, 1, 0, 1, 0, 1)

V = {also, football, games, John, likes, Mary, movies, to, too, watch}

Bag-of-Words: A Formal Perspective

$$D = \{d_1, \ldots, d_N\} = \text{ collection of } N \text{ documents}$$

$$V = \{w_1, \ldots, w_{|V|}\} = \textbf{ vocabulary} \text{ of } |V| \text{ words extracted from } D$$

$$\mathbf{d}_i = (f(w_1, i), \ldots, f(w_{|V|}, i)) = |V|\text{-dimensional vector representing } d_i$$

$f : V \times D \longmapsto \mathbb{R}$ is a function that maps each word of a document to a real value (**weighting scheme**)

Bag-of-Words: A Formal Perspective

One-Hot (binary) weighting scheme

$$f(w_j, i) = \begin{cases} 1 & \text{if } w_j \text{ appears in } d_i \\ 0 & \text{otherwise} \end{cases}$$

Bag-of-Words: A Formal Perspective

Term-Frequency weighting scheme

$$f(w_j, i) = tf(w_j, i)$$

tf computes the number of times word w_j occurs in document d_i

Bag-of-Words: A Formal Perspective

$\boxed{\text{TF-IDF weighting scheme}}$

$$f(w_j, i) = tf(w_j, i) * idf(w_j)$$

$$idf(w_j) = \log\left(\frac{N}{1 + n_j}\right)$$

n_j is the number of documents in D containing the word w_j

Bag-of-Words: A Formal Perspective

TF-IDF weighting scheme

$$f(w_j, i) = tf(w_j, i) * idf(w_j)$$

$$idf(w_j) = \log\left(\frac{N}{\boxed{1} + n_j}\right)$$

Any idea why?

n_j is the number of documents in D containing the word w_j

Bag-of-Words: Limitations and Improvements

- 2 main limitations of BoW model:
  - High dimensionality → sparseness
  - No sequential information nor semantics included → unigram model

Bag-of-Words: Limitations and Improvements

- 2 main limitations of BoW model:
  - High dimensionality → sparseness
  - No sequential information nor semantics included → unigram model
- Possible improvements:
  - Use *n*-grams rather than unigrams to capture sequentiality between consecutive words (i.e., context)
  - Even better, use so-called Neural Language Models like word2vec

Document Representation: Bag-of-*n*-grams

Example: bigrams (n=2)

doc 1

```
John likes to
watch movies.
Mary likes
movies too.
```

doc 2

```
Mary also likes
to watch
football games.
```

Document Representation: Bag-of-*n*-grams

Example: bigrams (n=2)

doc 1

```
John likes to
watch movies.
Mary likes
movies too.
```

doc 2

```
Mary also likes
to watch
football games.
```

{"John likes", "likes to", "to watch", "watch movies", "Mary likes", "likes movies", "movies too"}

{"Mary also", "also likes", "likes to", "to watch", "watch football", "football games"}

Document Similarity

- We have examined a number of possible document representations

Document Similarity

- We have examined a number of possible document representations
- Depending on those, several similarity measures can be used

Document Similarity

- We have examined a number of possible document representations
- Depending on those, several similarity measures can be used
- For example, if documents are represented as:
  - set of words → Jaccard coefficient

Document Similarity

- We have examined a number of possible document representations
- Depending on those, several similarity measures can be used
- For example, if documents are represented as:
  - set of words → Jaccard coefficient
  - one-hot bag-of-words → Euclidean distance

Document Similarity

- We have examined a number of possible document representations
- Depending on those, several similarity measures can be used
- For example, if documents are represented as:
  - set of words → Jaccard coefficient
  - one-hot bag-of-words → Euclidean distance
  - tf or tf-idf bag-of-words → Cosine similarity