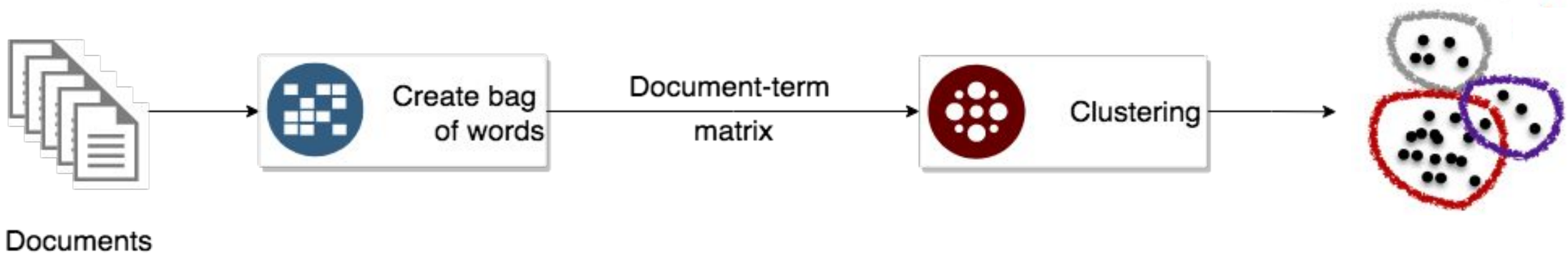
Análise de Mídias Sociais e Mineração de Texto

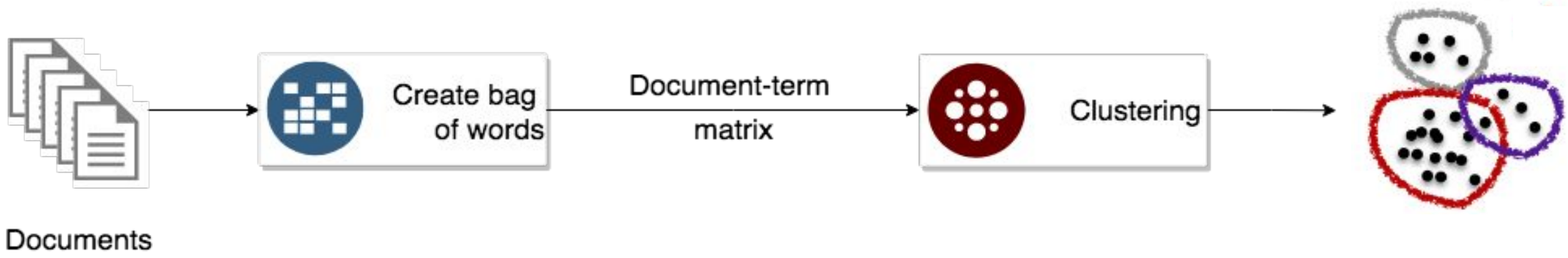
Modelagem de Tópicos

Laura de Oliveira F. Moraes

Processo NLP



Processo NLP



terms/tokens/words

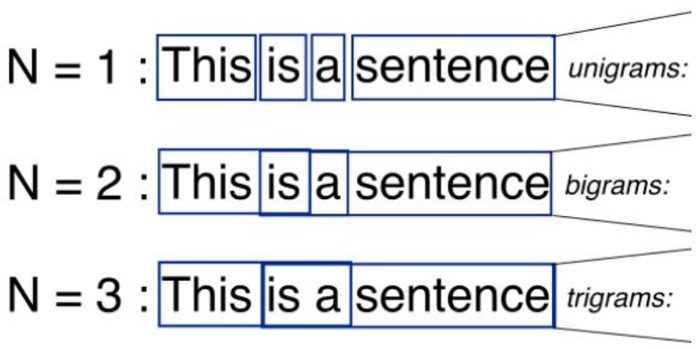
documents

	for	if	append
1	2	1	1
2	0	1	1
3	1	2	0

Create bag of words

Tokenizer — Frequência mínima no documento /stopwords — Contagem — Vetorização

Separa o texto em:



- Entre 5% a 50%
- Remove palavras comuns como preposições, artigos, etc

Como contar:

1. Uma vez por documento **OU**
2. Cada ocorrência no documento

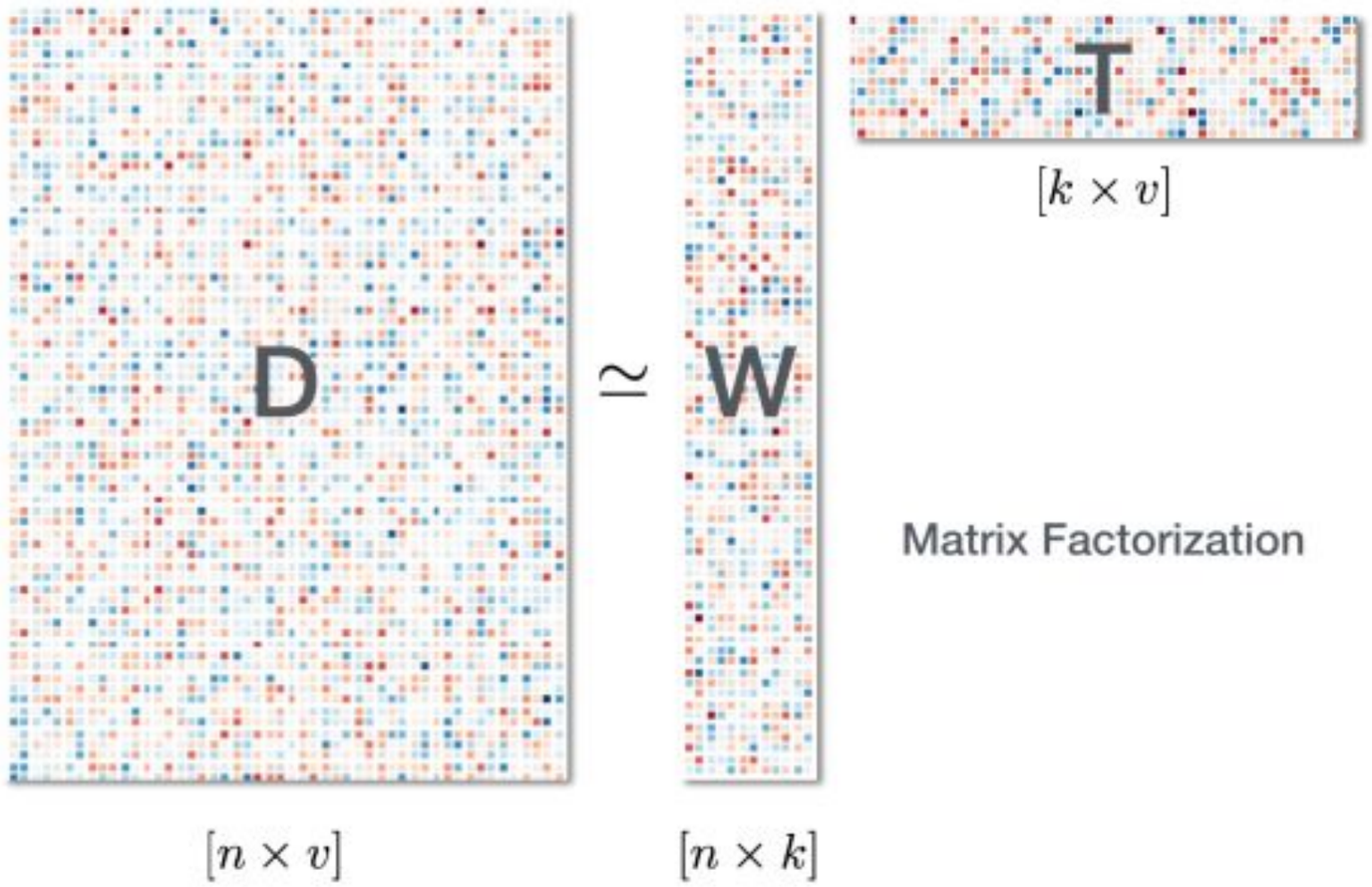
Normalização da contagem:

1. Contagem Normal **OU**
2. TF-IDF

Clustering *Methods for Topic Modeling*

1. Non-negative Matrix Factorization (**NMF**)

(Lee et al 1999, Cichocki et al 2009)



When constructing the document-term matrix, the terms counts can be interpreted as a set of the visible variables generated from an underlying set of hidden variables (topics). With this in mind, we can model the (hidden) topics as linear combinations of terms (a sum of each term weighted by its corresponding importance) and the documents as a weighted-sum of the topics it belongs to.

This factorization has a special property of only allowing non-negative values in its entries. By imposing this constraint, the resulting factorized matrices can be described essentially as weighted-union of sets, which is well-suited for human interpretability.

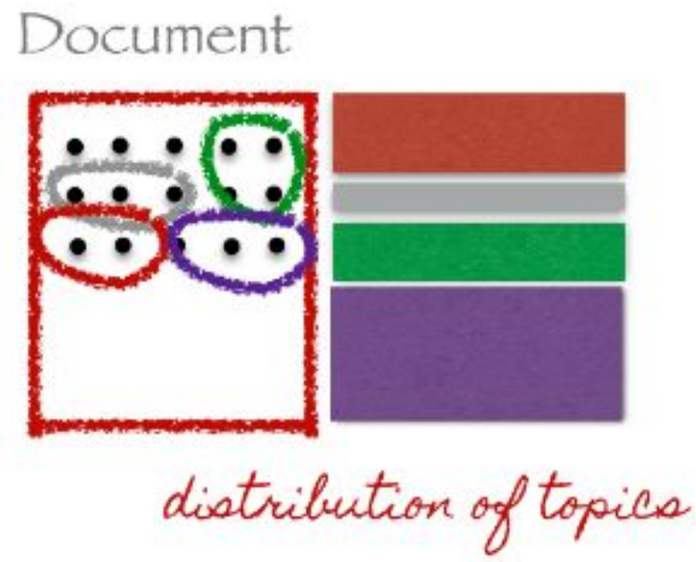
Figure reproduced from
http://tech.opentable.com/2015/01/12/finding-key-themes-from-free-text-reviews/

Clustering

Methods for Topic Modeling

2. Latent Dirichlet Allocation (**LDA**)

(Blei et al 2003. Steyvers et al 2010)



LDA is a generative probabilistic model that describes how documents in a collection are created. It assumes that:

1. A document is a mixture of topics

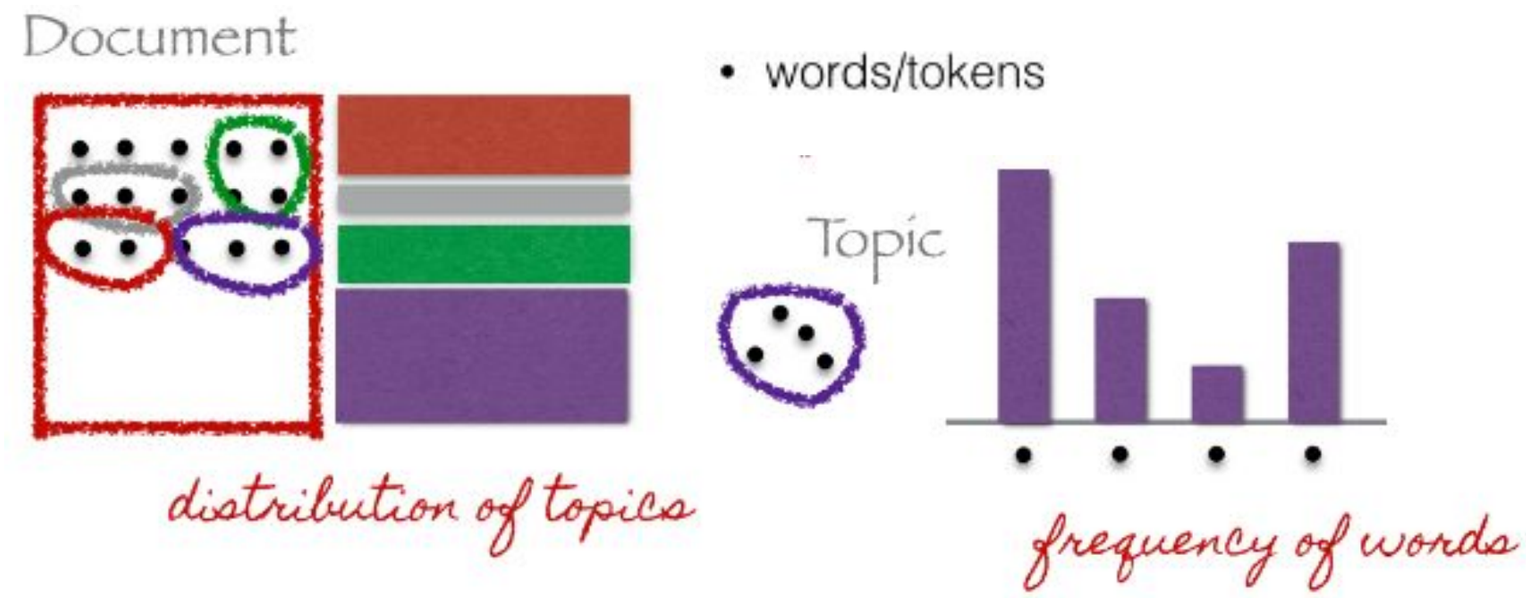
Figure reproduced from http://chdoig.github.io/pygotham-topic-modeling/#/

Clustering

Methods for Topic Modeling

2. Latent Dirichlet Allocation (**LDA**)

(Blei et al 2003. Steyvers et al 2010)



LDA is a generative probabilistic model that describes how documents in dataset were created. It assumes that:

1. A document is a mixture of topics
2. A topic is a distribution over words

Figure reproduced from http://chdoig.github.io/pygotham-topic-modeling/#/

Clustering

Methods for Topic Modeling

2. Latent Dirichlet Allocation (**LDA**)

(Blei et al 2003. Steyvers et al 2010)

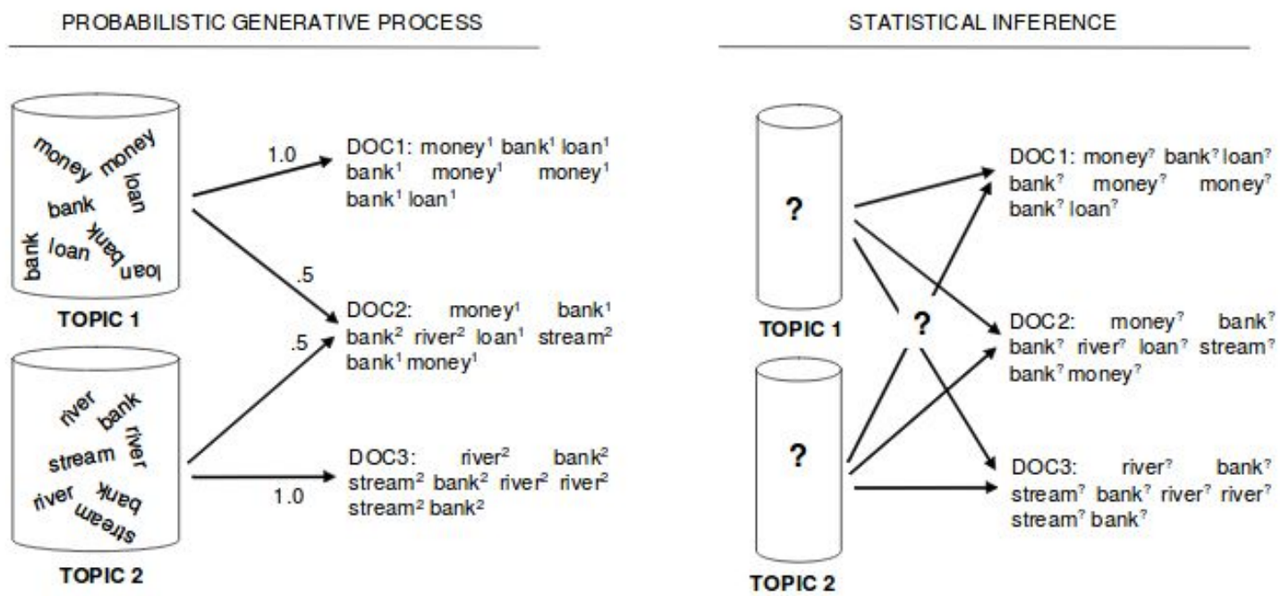


Figure reproduced from Steyvers et al 2010

To generate a document, you simply sample from the distributions:

1. Sample a topic for the document
2. Sample a word from the topic

Doing this iteratively, you will generate a document.

However, once you have a dataset, meaning a collection of documents, what we need is to discover these distributions. The LDA algorithm tries to backtrack this probabilistic model to find a set of topics that are likely to have generated the collection.

TF-IDF

TF-IDF is a measure of originality of a word by comparing the number of times a word appears in a doc with the number of docs the word appears in.

$$\text{TF-IDF} = TF(t, d) \times IDF(t)$$

- $TF(t, d)$: Term frequency — Number of times term t appears in a doc, d
- $IDF(t)$: Inverse document frequency

$$\log \frac{1 + n}{1 + df(d, t)} + 1$$

- n: # of documents
- $df(d, t)$: Document frequency of the term t