Practical issues with pLSA

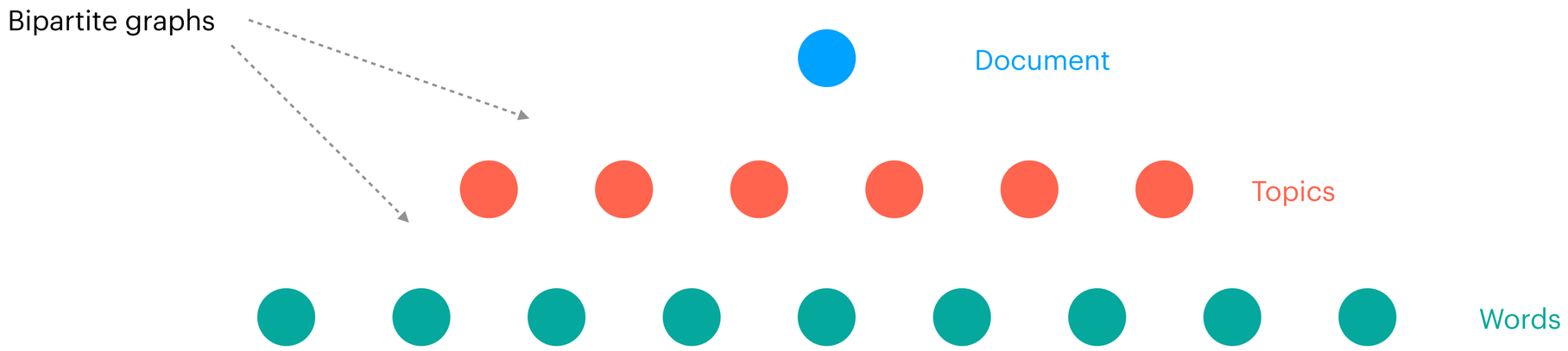
- There are a few problems that arise when using pLSA in practice:
    - There is no probabilistic interpretation at the document level; each document is represented as a list of numbers and there is no generative model for them.
    - Parameters grows linearly with the size of the corpus, which leads to overfitting
    - No natural way to assign probability to a document outside of the training set
    - Unlike NMF, there is no clear way to impose sparsity on our latent representation

[1] Blei, 2003

Latent Dirichlet Allocation (LDA)

- LDA is a generative model of text documents that defines a latent space, $\mathbf{z}$, which describes the relationship between words and documents probabilistically.

- In LDA, documents and words are related only through $\mathbf{z}$



[1] Blei, 2003