

## Latent Dirichlet Allocation

In der Statistik ist Latent Dirichlet Allocation (LDA) ein generatives Wahrscheinlichkeitsmodell für Dokumente wie Texte und Bilder. Dabei wird jedes Dokument von verschiedenen Themen die zugrunde liegen betrachtet. Zum Beispiel werden Beobachtungen von Wörtern in Dokumenten gesammelt, sodass jedes Dokument eine Mischung aus einer kleinen Anzahl von Themen ist und dass jedes Wort der Schöpfung der Themen des Dokuments dient. LDA ist ein Beispiel für ein Wahrscheinlichkeitsmodell und wurde zunächst als ein graphisches Modell zum Thema Entdeckung von David Blei, Andrew Ng, und Michael Jordan im Jahr 2002 vorgestellt.

## Themen im LDA

In LDA, kann jedes Dokument als eine Mischung von verschiedenen Themen dargestellt werden. Dies ist vergleichbar mit einer Wahrscheinlichkeits-latent semantischen Analyse (pLSA; probabilistic latent semantic analysis), außer dass in LDA das Thema der Verteilung angenommen wird, dass ein Dirichlet vorhat. In der Praxis führt dies zu mehr Mischungen von Themen in einem Dokument. Es wurde jedoch festgestellt, dass dies dem pLSA Modell entspricht.

Zum Beispiel könnte ein LDA-Modell ein Thema haben wie Katze und Hund klassifiziert werden können. Zwar ist die Klassifikation willkürlich, da das Thema diese Worte nicht benennen kann. Darüber hinaus hat ein Thema Wahrscheinlichkeiten zur Erzeugung verschiedener Worte: die Worte Milch, miauen, Kätzchen und das kann klassifiziert und interpretiert werden vom Betrachter als mit hoher Wahrscheinlichkeit als Thema „Katze“ interpretiert. Das Hunde-Thema erzeugt ebenfalls Wahrscheinlichkeiten für jedes Wort: Hund und Knochen zum Beispiel. Worte ohne besondere Relevanz, wie die (siehe Funktion Wort), wird etwa noch Wahrscheinlichkeit zwischen Klassen (oder können in einer separaten Kategorie zugeordnet werden).

Ein Dokument wird durch Auswahl einer Verteilung über Themen erzeugt (d.h. meist über Hund, meist über Katze, oder ein bisschen von beidem), und angesichts dieser Verteilung, wird das Thema des jeweiligen bestimmten Wortes kommissioniert. Dann werden Wörter zu ihren Themen generiert. (Beachten Sie, dass Worte als unabhängig angesichts der Themen sind)