Accès à l'information - Suites du cours 2

François Yvon

30 novembre 2016

PLSA, etc

— Essayez de retrouver la forme de la fonction auxiliaire.
— comment procéderiez vous pour « lisser » les comptes dans PLSA ? Essayer de retrouver les formules en ajoutant des distributions à priori pour les α et les β
— Essayer de réimplanter PLSA dans le langage de votre choix - ce qui signifie essentiellement réimplanter les formules de la planche numéro. Pour les données voir ci-dessous.

EM

— Redémontrer les propriétés 1-5 de la fonction $F(q, \boldsymbol{\theta})$.

Regarder des films

Pour ceux qui préfèrent l'anglais, l'exposé de Thomas Hoffman ici permet de voir les contextes d'applications en RI de PLSA, en plus de la présentation de la méthode par son inventeur : `http://videolectures.net/slsfs05_hofmann_lsvm/?q=plsa%20hofmann%20latent%20semantic%20analysis`

Lire un article

Eessayer de lire (avec un oeil maintenant acéré) tout l'article :
Thomas Hofmann. *U*nsupervised Learning by Probabilistic Latent Semantic Analysis.
Machine Learning 42(1/2) : 177-196 (2001)

Téléchargeable ici : `http://www.cs.helsinki.fi/u/vmakinen/stringology-k04/hofmann-unsupervised_learning_by_probabilistic_latent_semantic_analysis.pdf`

Faire des expériences

Installer Gensim : `https://radimrehurek.com/gensim/`.

1. Hofmann prétend (p11) :

   > By symmetry this also holds for different occurrences of the same word. As a result of this coupling, the probabilities $P(z_k|d_i)$ and $P(z_k|w_j)$ tend to be "sparse", i.e., for given d_i or w_j typically only few entries are significantly different from zero.

   Vérifier que cela est effectivement le cas sur des données réelles.

Vous pouvez trouver des matrices de comptes directement utilisables ici : `http://archive.ics.uci.edu/ml/datasets/Bag+of+Words`.