Modélisation probabiliste de collections textuelles et distributions de mots

Stéphane Clinchant*†, Eric Gaussier†

*Xerox Research Centre Europe
Stephane.Clinchant@xrce.xerox.com
†Laboratoire d'Informatique de Grenoble
Université J. Fourier, Grenoble 1
Eric.Gaussier@imag.fr

Résumé. Nous examinons dans cet article les liens entre modèles probabilistes de documents textuels et observations empiriques sur la distribution des mots au sein d'une collection. Nous proposons une caractérisation formelle de ces observations, et introduisons la distribution beta négative binomiale. Cette distribution (connue sous diverses dénominations mais dont la dérivation que nous proposons est nouvelle) permet de rendre compte des observations empiriques et fournit un modèle non paramétrique dont le bon comportement est validé en catégorisation de textes.

1 Introduction

Plusieurs modèles probabilistes de collections textuelles ont récemment été proposés, dans un cadre apprentissage automatique ou recherche d'information Hofmann (1999); Blei et al. (2003); Buntine et Jakulin (2004); Madsen et al. (2005); Elkan (2006). Ces modèles sont souvent définis d'un point de vue statistique. Au cours des années, cependant, plusieurs observations empiriques sur la façon dont les mots se comportent dans les documents ont été faites (depuis le travail de G. Zipf en 1949 jusqu'à des études plus récentes, comme par exemple celles décrites dans Church et Gale (1995); Katz (1996)). Dans quelle mesure les modèles probabilistes sont en accord avec ces observations empiriques est une question toujours ouverte, et que nous abordons dans cet article. En particulier, nous commençons par passer en revue les observations empiriques sur les distributions de mots, en les caractérisant formellement (section 2). Nous examinerons ensuite quelques-uns des modèles probabilistes les plus populaires au regard de ces caractéristiques formelles (section 3), avant de dériver une loi (BNB) qui rend compte de toutes ces caractéristiques (section 4). La section 5 valide expérimentalement cette nouvelle loi sur une tâche de catégorisation de textes.