lda.py とデータセット（その２）

木村　健

平成 30 年 12 月 5 日

1　このドキュメントについて

本ドキュメントは持橋先生が改良し、さらにそれを木村が少し改良した lda.py について、いくつかハイパーパラメータについての知見やコーパス（ja.text8）のバグが発見されたのでそれについて学習結果を踏まえて解説する。

それぞれの分類結果のファイル名と各種パラメーターについては、最後に表を列挙するので急いでいる方はそちらを当たられたい。

2　NIPS

NIPS についてはもうすでに前処理された結果が提供されているので、特にこれ以上改良しない。

3　The 20 Newsgroups data set

正規表現”[a-zA-Z]+”を通るものだけ扱ってはどうか、という持橋先生のご指摘に則り、そのような正規表現フィルターを実装してみた。なお最低頻度の閾値が今見たら 50 になっていたが、これはもっと下げてもいいかもしれない。ただあとで述べるがデータ（分類結果）を観たところ単語ではないものが多く散見され、むしろ閾値をあげた方がいいのかもと思ってしまう結果であった。

4　ja.text8

まず、最低頻度が高すぎるであろうという意見があった。頻度の低い単語の中に重要な情報が含まれていることは珍しいことではなく、ただ perplexity のためだけに最低頻度を高く設定してしまうのは間違いだということで、一気に 2 以上だけ扱う、という風にした。

また ja.text8 のコーパス自体（僕の変更した分）に誤りがあり、連接する文章の切れ目（普通は「。」と「先頭単語」）にスペースが入っていなくて、「。あ」のような単語が出来てしまっていた。コーパスごと修正した。このためコーパスに収録の文数が少しだけ少なくなった。収録文書数は 36353 文書である。

5　hyper parameter alpha

$\alpha = 0.01$ とした。（過去の実験で一番良かった値。）

6　hyper parameter beta

$\beta = 0.01$ とした。

7　結果ファイル（重要なものだけ抜粋）

file prefix	alpha	beta	N	K	En/Ja	date	min. perp.	desc.
ja.0717least2	0.01	0.01	1000	1000	Ja	Jul 18	9695	最低頻度 2 以上
ja.0720.2000	0.01	0.01	1000	2000	Ja	Jul 22	11439	K=2000 のデータ
ja.text8.cool	0.01	0.01	100	10	Ja	Jun 22	1273	coolcutter 適用後
20news.regexp	0.01	0.01	1000	1000	En	Jul 25	2177	正規表現フィルタ
ja.text8.kanji	0.01	0.01	1000	2000	Ja(new)	Jul 26	10189	コーパス修正

結果ファイルについては NLP リポジトリの/topics 以下を参照されたい。prefix に.topics.csv をつけたファイル名がデータファイル名である。

- コーパスのバグにより perplexity が 1000 ほど上がっていた
- 最低頻度を 2 にすると K が大きい時にいろんな単語が入ってきてデータとして充実する
- 正規表現フィルターは 20news に入れたのだが、却って非単語が目立つ形となった
- 日本語の stop words は頻度順にトップ 100 を除いている
- 英語の stop words はライブラリ提供のものを使った

参考文献

[1] Kevin P. Murphy: Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning series) The MIT Press, 2012.

[2] Blei, David M. and Ng, Andrew Y. and Jordan, Michael I.: Latent Dirichlet Allocation, J. Mach. Learn. Res. 3/1, volume 3, 2003.

[3] Thomas L. Griffiths, Mark Steyvers.: Finding Scientific Topics. PNAS (101) pp. 5228-5235, 2004.