lightlda.sh

木村　健

平成 30 年 12 月 6 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からご提供いただいた、lightlda を簡単に使う lightlda.sh について、既存のデータセットで試したのでその報告をする。

なお分類結果については S3, BitBucket に置く予定なので参照していただきたい。

2　lightlda[4] とは

非常に高速に LDA を行うプログラムである（詳細はウェブページを。。。）。ただ入出力の形式が独自で、少し（lda.py を使っていた身からすると）使いずらい。そこで、lightlda.sh という一連のスクリプトを使うと、なんと、ほぼ lda.py のインターフェイスで LDA を行うことができる（素晴らしい）。

今回はその lightlda.sh に対して新しい Python ファイルの追加と、lightlda.sh(bash ファイル) の追加を行い、visLDA.py と同様の結果を得ることに成功した。

また LDA の各トピックごとのソーティング（上下関係）も二種類の尺度を用いて分類できたのでこれを紹介する。

3　preLDAJSON.py

以前作ったファイルだが、All About コーパスの JSON 形式（AA2.json という巨大なファイル）から lda.py に必要な svm ファイルと単語を列挙した words ファイルを抽出する。

各記事には記事番号が付与されている。また記事のページごとにその下の階層でページ番号が付与されている。各ページは見出しとテキストの集合からなり、今回はページ単位で一つのドキュメントとして svm ファイルを生成した。

4　lightlda2pickle.py

lightlda.sh の実行結果取り出しから直接 visLDA.py のような csv 結果ファイルを取り出せないことがわかり、lightlda.sh の最終段階に pickle ファイルを生成する lightlda2pickle.py ファイルを新しく追加してより多くの情報を取り出すようにした。

k をトピックの id、w を単語 id、d をドキュメント id とする。param ファイルには、*alpha*, *beta*, *iterations*, *topics* が入っている。

lightlda2pickle.py の生成物は pickle4.gz という gzip された Python Pickle file (protocol = 4) である。

中に入っているオブジェクトは dict 一つだけであり、それぞれの key と content は次の通り。今 dict オブジェクトを model として参照しているとすると、

- model['alpha'] = $alpha$ (single value)
- model[‘topics’] = $topics$ (single value)
- model[‘C(w)’] = 単語の頻度 $C(w)$ (長さ=$nlex$).
- model[’ N’] = N (total occurence).
- model[‘nlex’] = $nlex$ (語彙数).
- model[’ C(k)’] = それぞれの topics の頻度 $C(k)$
- model[’ P(w|k)’] = キー (w,k) を持つ dict: dict の中身は $P(w|k)$
- model[’ P(w|k)/P(w)’] = キー (w,k) を持つ dict: dict の中身は $\frac{P(w|k)}{P(w)}$

最後の二つが visualize に使われる確率で、それぞれ次のように（便宜上）呼ぶことにする。

- $P(w|k)$: normal probability
- $\frac{P(w|k)}{P(w)}$: maniac probability

$$P(w|k) = \frac{P(k|w)P(w)}{P(k)}. \tag{1}$$

$$\frac{p(w|k)}{P(w)} = P(k|w)P(w). \tag{2}$$

$$P(k|w) = \frac{C(k,w)}{C(w)}. \tag{3}$$

$$C(w) = \frac{C(k,w)}{P(k|w)} -> normalized -> P(w). \tag{4}$$

$$\mathrm{or} C(w) = \sum_k C(k,w) -> normalized -> p(w). \tag{5}$$

$$\mathrm{or} P(w|k) = \frac{C(w,k)}{C(k)}. \tag{6}$$

$$\tag{7}$$

5 結果

結果をレポジトリの topics フォルダに置いたので参照して欲しい。特に maniac の方は「マニアック（笑）」と思える内容の単語が上に来ていることがわかる。

6 lightlda, lightlda.sh の変更箇所

6.1 lightlda の変更箇所

/lightlda/multiverso (/lightlda を展開フォルダとしたとき) の Makefile を変更（Ubuntu 18.04).

```
diff --git a/Makefile b/Makefile
index c7bf4f5..e42d214 100644
--- a/Makefile
+++ b/Makefile
@@ -18,7 +18,7 @@ THIRD_PARTY_LIB = $(THIRD_PARTY)/lib


 INC_FLAGS = -I$(HEADERS_DIR)
```

```
 INC_FLAGS += -I$(THIRD_PARTY_INC)
-LD_FLAGS = -L$(THIRD_PARTY_LIB) -lzmq -lmpich -lmpl
+LD_FLAGS = -L$(THIRD_PARTY_LIB) -lzmq -lmpich -lmpl -lpthread

 LIB_SRC_DIR = $(PROJECT)/src/multiverso
 SERVER_SRC_DIR = $(PROJECT)/src/multiverso_server
```

あと LD_LIBRARY_PATH に　zmq のライブラリが見えるようにする必要があったと思う（うろ覚え）。

6.2 lightlda.sh の変更箇所

シェバングを/bin/sh から/bin/bash に。最後で、

```
python3 ../lightlda2pickle.py .
```

を追加。

全体的に lightlda2pickle.py は追加。

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.

[4] LightLDA: Big Topic Models on Modest Computer Clusters: Yuan, Jinhui and Gao, Fei and Ho, Qirong and Dai, Wei and Wei, Jinliang and Zheng, Xun and Xing, Eric Po and Liu, Tie-Yan and Ma, Wei-Ying, Proceedings of the 24th International Conference on World Wide Web, pp.1351–1361,2015.