# lda.py の test sets での学習について

木村　健

平成 30 年 6 月 27 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作成された lda.py について、三種類のテストセットで実際どのようにトピック付けがなされるかを報告するものである。

## 2　LDA のタスクと入力パラメータ

LDA[2] はドキュメント群（ここでは簡単のため英語とする）の各単語について潜在的トピックを推定する技法である。いくつのトピックに分けるかは最初に指定する。常に同じ順序でトピックが抽出されると限らない。unsupervised learning である（教師値を用いない）。

LDA の開始に先立って次のパラメータが必要となる (LDA.py の index.html より引用加筆) Gibbs sampling という技法を利用していることを付け加えておく。

- K: topics: number of topics in LDA
- N: iters: number of Gibbs iterations
- $\alpha$: Dirichlet hyperparameter on topics
- $\beta$: Dirichlet hyperparameter on words

K がトピック数で、はじめに与える。例えば $K = 10$ とする。N はイテレーションの回数でこれを大きくするほど perplexity が改善（小さくなる）するがある程度まで行くとあまり動かなくなる。

## 3　NIPS

Abstract: This data set contains the distribution of words in the full text of the NIPS conference papers published from 1987 to 2015.

Data Set Characteristics:

Text

Number of Instances:

11463

Area:

Computer

Attribute Characteristics:

Integer

Number of Attributes:

5812

Date Donated
2016-11-23
Associated Tasks:
Clustering
Missing Values?
N/A
Number of Web Hits:
28333
Data Set Information:

The dataset is in the form of a 11463 x 5812 matrix of word counts, containing 11463 words and 5811 NIPS conference papers (the first column contains the list of words). Each column contains the number of times each word appears in the corresponding document. The names of the columns give information about each document.

NIPS は学会である NIPS に投稿された論文について、documents x words の CSV(密行列) を提供するもので、低頻度単語の omit や正規化を行ったデータセットである。

## 4 The 20 Newsgroups data set

The 20 Newsgroups data set is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. To the best of my knowledge, it was originally collected by Ken Lang, probably for his Newsweeder: Learning to filter netnews paper, though he does not explicitly mention this collection. The 20 newsgroups collection has become a popular data set for experiments in text applications of machine learning techniques, such as text classification and text clustering.

Organization The data is organized into 20 different newsgroups, each corresponding to a different topic. Some of the newsgroups are very closely related to each other (e.g. comp.sys.ibm.pc.hardware / comp.sys.mac.hardware), while others are highly unrelated (e.g misc.forsale / soc.religion.christian). Here is a list of the 20 newsgroups, partitioned (more or less) according to subject matter:

- comp.graphics
- comp.os.ms-windows.misc
- comp.sys.ibm.pc.hardware
- comp.sys.mac.hardware
- comp.windows.x rec.autos
- rec.motorcycles
- rec.sport.baseball
- rec.sport.hockey sci.crypt
- sci.electronics
- sci.med
- sci.space
- misc.forsale
- talk.politics.misc
- talk.politics.guns
- talk.politics.mideast
- talk.religion.misc

- alt.atheism
- soc.religion.christian

Data The data available here are in .tar.gz bundles. You will need tar and gunzip to open them. Each subdirectory in the bundle represents a newsgroup; each file in a subdirectory is the text of some newsgroup document that was posted to that newsgroup.

Below are three versions of the data set. The first (”19997”) is the original, unmodified version. The second (”bydate”) is sorted by date into training(60%) and test(40%) sets, does not include cross-posts (duplicates) and does not include newsgroup-identifying headers (Xref, Newsgroups, Path, Followup-To, Date). The third (”18828”) does not include cross-posts and includes only the ”From” and ”Subject” headers.

20news-19997.tar.gz - Original 20 Newsgroups data set 20news-bydate.tar.gz - 20 Newsgroups sorted by date; duplicates and some headers removed (18846 documents) 20news-18828.tar.gz - 20 Newsgroups; duplicates removed, only ”From” and ”Subject” headers (18828 documents) I recommend the ”bydate” version since cross-experiment comparison is easier (no randomness in train/test set selection), newsgroup-identifying information has been removed and it’s more realistic because the train and test sets are separated in time.

昔懐かしい NetNews の記事を集めたコーパス。bydate を使用。30 頻度以下の単語の omit と正規化を実施。プログラムを作り lda.py に入力できる形に編集。

## 5 ja.text8

ja.wikipedia.org のデータを集め、100MB で crop した日本語コーパス。ただし、元のコーパスは全て一行に入れられており、ドキュメントごとになってなかったので、repo を fork して新しいスクリプトを起草。元データが古く wikipedia サイトにももうなかったので、比較的新しい（2018/06/01）データでコーパスを再編成した。30 頻度以下の単語の omit をしている。コーパス自体は分かち書き（スペース区切り）がなされており、一行 1 ドキュメントである。

original ja.text8 is a small (100MB) text corpus from the web (japanese wikipedia). it modified by kimrin for one line per one article format.

You can download ja.text8 corpus from the following link(original corpus):

this repository contains ja.text8.20180601.100MB that is body of the new corpus.

Requirements Python 3.x MeCab

License CC-BY-SA

## 6 hyper parameter $\alpha$

ハイパーパラメータ $\alpha$ を変化させて perplexity の変化を見た。コーパスは ja.text8 である。
意外とデフォルトの $50/K$ が最良ではないことがわかった。

## 7 hyper parameter $\beta$

ハイパーパラメータ $\beta$ を変化させて perplexity の変化を見た。コーパスは ja.text8 である。
こちらは意外と変化は少なかった。$\beta = 0.01$ が最良でいいと思われる。

## 8　後処理の有無による perplexity の変化

コーパスは 20newsgroups である。このコーパスは英語かつ平文からデータを作っている関係で、前処理（ストップワードの除去）により perplexity に変化が出るか測定した。

## 9　木村改善案と元の lda.py の比較

結論：ほとんど一緒で、topics の randomness を考えるとほぼ誤差。

## 10　coolcutter.py による素性除去

素性数や df の clamp を行うツールで前処理した結果である。
こちらは $perplexity = 1200$ ぐらいで落ち着いた。

## 11　木村改善案と元の lda.py の比較（topics）

分類結果について図示する（元=OLD, 木村=NEW、COOL のついているものは coolcutter.py 済み）

## 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.

NIPS10

| init : python/lda.py -K 10 -N 100 ../data/NIPS_1987-2015.csv.pseudo.svm ../data/NIPS10 | | | | |
|---|---|---|---|---|
| host : MacBook-Pro.local | | | | |
| start : Thu Jun 14 22:43:38 2018 | | | | |
| iterations | K=10 | K=20 | K=50 | K=100 |
| 1 | 28834.9082 | 60561.1734 | 170743.3882 | 397853.5293 |
| 2 | 28108.0886 | 57149.8029 | 146194.2989 | 295580.8796 |
| 3 | 27346.8758 | 53966.4703 | 128168.961 | 233965.4657 |
| 4 | 26589.3846 | 51054.2666 | 113284.3825 | 189549.1172 |
| 5 | 25802.5011 | 48188.76 | 99828.4335 | 153667.7885 |
| 6 | 24926.2281 | 45102.8765 | 86181.8947 | 121468.8293 |
| 7 | 23880.0226 | 41322.3362 | 70534.5305 | 90778.5598 |
| 8 | 22352.9244 | 36208.5254 | 54128.8739 | 63989.4578 |
| 9 | 20047.577 | 29828.7838 | 40143.0991 | 44948.7133 |
| 10 | 17021.914 | 23598.4583 | 29963.0367 | 33097.0637 |
| 11 | 14054.6958 | 18822.1933 | 23306.616 | 25800.2341 |
| 12 | 11797.7545 | 15553.5051 | 19152.6023 | 21244.4353 |
| 13 | 10283.854 | 13359.0385 | 16464.8368 | 18286.3269 |
| 14 | 9288.7967 | 11830.1489 | 14638.011 | 16250.4143 |
| 15 | 8585.8641 | 10742.1121 | 13344.4936 | 14805.9958 |
| 16 | 8075.279 | 9945.7554 | 12384.0562 | 13706.7171 |
| 17 | 7687.5501 | 9361.8631 | 11643.3915 | 12849.991 |
| 18 | 7385.8749 | 8916.0146 | 11039.5754 | 12162.3981 |
| 19 | 7151.6814 | 8567.2188 | 10540.1518 | 11593.7413 |
| 20 | 6959.0999 | 8293.8319 | 10131.7466 | 11117.0028 |
| 21 | 6809.9899 | 8068.5409 | 9793.1359 | 10713.2183 |
| 22 | 6682.2019 | 7891.5895 | 9504.3257 | 10361.7387 |
| 23 | 6574.6219 | 7735.2807 | 9258.7355 | 10057.1458 |
| 24 | 6484.6739 | 7602.8762 | 9043.3781 | 9798.0864 |
| 25 | 6405.2102 | 7483.2727 | 8852.5666 | 9561.4407 |
| 26 | 6335.708 | 7380.4519 | 8680.3172 | 9344.9161 |
| 27 | 6276.5094 | 7288.8513 | 8534.0912 | 9140.6313 |
| 28 | 6222.3034 | 7199.7256 | 8390.4781 | 8972.3439 |
| 29 | 6170.9601 | 7131.6758 | 8268.736 | 8802.6225 |
| 30 | 6125.236 | 7070.3986 | 8149.2859 | 8659.831 |
| 31 | 6079.4007 | 7011.9696 | 8052.6962 | 8525.8484 |
| 32 | 6040.3321 | 6957.125 | 7962.9183 | 8399.7218 |
| 33 | 5999.5914 | 6909.7016 | 7886.4099 | 8282.2719 |
| 34 | 5964.4286 | 6868.4204 | 7804.7048 | 8171.4879 |
| 35 | 5930.9429 | 6829.2914 | 7730.0851 | 8073.014 |
| 36 | 5896.6001 | 6791.7905 | 7664.4907 | 7978.8914 |
| 37 | 5867.8216 | 6753.2919 | 7601.8489 | 7892.2253 |
| 38 | 5838.1038 | 6719.5769 | 7540.7118 | 7808.8194 |
| 39 | 5811.8857 | 6682.336 | 7488.7731 | 7734.6732 |
| 40 | 5785.9288 | 6658.7951 | 7438.726 | 7662.7404 |
| 41 | 5763.2123 | 6631.1997 | 7388.0271 | 7592.3688 |
| 42 | 5745.2443 | 6597.7218 | 7334.5646 | 7530.9066 |
| 43 | 5725.5971 | 6570.1152 | 7297.6555 | 7477.8077 |
| 44 | 5706.2922 | 6547.0056 | 7264.8228 | 7416.7324 |
| 45 | 5691.8579 | 6527.2781 | 7226.3101 | 7354.866 |
| 46 | 5675.4911 | 6506.8152 | 7193.372 | 7305.3836 |
| 47 | 5659.5709 | 6491.2822 | 7155.0513 | 7248.5904 |
| 48 | 5645.7791 | 6471.0052 | 7130.0163 | 7205.3184 |
| 49 | 5627.1142 | 6448.57 | 7091.2016 | 7150.267 |
| 50 | 5615.8415 | 6432.4879 | 7057.578 | 7115.2826 |
| 51 | 5604.4386 | 6419.1523 | 7028.7784 | 7075.4984 |
| 52 | 5593.9013 | 6406.1466 | 6999.1555 | 7037.8677 |
| 53 | 5579.0178 | 6390.8828 | 6977.3143 | 6989.8477 |
| 54 | 5568.7634 | 6375.4753 | 6950.1926 | 6955.9727 |
| 55 | 5558.5658 | 6359.2294 | 6936.9584 | 6924.7466 |
| 56 | 5549.0377 | 6353.7597 | 6912.4566 | 6896.162 |
| 57 | 5541.801 | 6341.3393 | 6887.5623 | 6866.8042 |
| 58 | 5535.0364 | 6327.3387 | 6864.356 | 6833.8483 |
| 59 | 5528.0365 | 6315.6668 | 6841.1467 | 6800.1035 |
| 60 | 5517.9843 | 6308.1131 | 6827.76 | 6771.789 |
| 61 | 5506.4676 | 6302.7596 | 6809.1205 | 6742.8903 |
| 62 | 5503.0139 | 6291.6771 | 6795.4831 | 6718.9712 |
| 63 | 5493.9951 | 6283.7364 | 6780.4896 | 6692.52 |
| 64 | 5490.2074 | 6268.7477 | 6763.4531 | 6669.0011 |
| 65 | 5485.0019 | 6262.6876 | 6741.6491 | 6640.46 |
| 66 | 5476.5099 | 6254.2848 | 6724.9989 | 6613.1556 |
| 67 | 5469.8586 | 6245.7964 | 6711.3113 | 6588.3154 |
| 68 | 5463.1124 | 6238.3523 | 6696.303 | 6569.0884 |
| 69 | 5456.134 | 6230.5958 | 6686.5073 | 6550.1613 |
| 70 | 5455.9986 | 6220.4632 | 6672.3246 | 6522.634 |
| 71 | 5455.1388 | 6213.3709 | 6649.5557 | 6500.1296 |
| 72 | 5449.2729 | 6203.9355 | 6633.1297 | 6481.1833 |
| 73 | 5441.5002 | 6196.5956 | 6627.9052 | 6457.6918 |
| 74 | 5438.0817 | 6191.0885 | 6616.3784 | 6437.5897 |
| 75 | 5431.1419 | 6185.559 | 6592.4138 | 6422.6093 |
| 76 | 5424.8301 | 6179.792 | 6583.7583 | 6404.9864 |
| 77 | 5424.8665 | 6173.9321 | 6571.0092 | 6389.9542 |
| 78 | 5423.8421 | 6166.7763 | 6560.7071 | 6374.0933 |
| 79 | 5419.6174 | 6158.5279 | 6558.8259 | 6358.9526 |
| 80 | 5412.8509 | 6157.8749 | 6544.6093 | 6342.0948 |
| 81 | 5413.7085 | 6153.673 | 6535.1397 | 6327.3562 |
| 82 | 5409.5682 | 6149.1059 | 6521.9699 | 6314.2107 |
| 83 | 5405.0091 | 6147.4947 | 6516.3543 | 6297.1743 |
| 84 | 5397.9619 | 6144.4722 | 6509.2431 | 6285.0674 |
| 85 | 5392.8852 | 6137.403 | 6497.7723 | 6271.91 |
| 86 | 5392.1721 | 6133.174 | 6480.4817 | 6260.7374 |
| 87 | 5388.4955 | 6129.0153 | 6475.1923 | 6248.9294 |
| 88 | 5385.0586 | 6124.5746 | 6466.1467 | 6238.115 |
| 89 | 5384.1582 | 6122.7548 | 6456.4069 | 6223.6403 |
| 90 | 5383.0092 | 6118.5426 | 6451.52 | 6215.0189 |
| 91 | 5375.8517 | 6115.8778 | 6437.147 | 6197.8592 |
| 92 | 5373.2423 | 6115.0714 | 6428.3057 | 6184.9973 |
| 93 | 5369.6253 | 6112.3263 | 6419.2982 | 6177.4825 |
| 94 | 5370.9614 | 6107.4688 | 6412.1027 | 6162.5689 |
| 95 | 5367.0249 | 6104.7981 | 6401.1534 | 6150.2906 |
| 96 | 5367.1832 | 6100.9041 | 6397.9858 | 6142.9634 |
| 97 | 5363.0555 | 6098.3404 | 6390.4196 | 6131.38 |
| 98 | 5363.7384 | 6099.9039 | 6388.4446 | 6118.2918 |
| 99 | 5357.2518 | 6096.4992 | 6380.4921 | 6113.3529 |
| 100 | 5355.8266 | 6091.5071 | 6371.6038 | 6098.5244 |
| ========== | | | | |
| finish : Thu Jun 14 22:45:48 2018 | | | | |

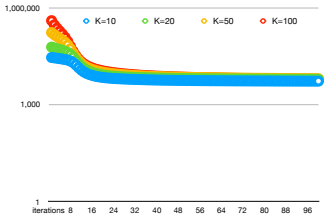


図 1　NIPS

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | topic0 | topic1 | topic2 | topic3 | topic4 | topic5 | topic6 | topic7 | topic8 | topic9 |
| 2 | ilya(22313) | learners(32320) | theor(17050) | mode(15690) | stat(20157) | mode(14702) | netw(30857) | algorithm(16586) | mode(39579) | matrices(29581) |
| 3 | imagery(14707) | trained(25375) | bound(16619) | timal(13416) | learners(18082) | modelling(7698) | neur(23464) | granule(14980) | dat(24774) | probit(15690) |
| 4 | objec(10713) | dat(25095) | lesser(14735) | figueiredo(8225) | timal(12606) | wordnet(7616) | inpainting(20841) | trec(10908) | distributing(23940) | algorithm(15555) |
| 5 | feature(9239) | sessions(17466) | fun(14513) | responds(7005) | algorithm(11018) | woods(6644) | networking(18888) | sessions(10326) | modelling(22718) | dat(12943) |
| 6 | mode(7742) | classifica(16396) | case(11533) | sign(6915) | policies(10811) | sessions(5436) | outperforms(12782 | node(9870) | gauss(16988) | line(11871) |
| 7 | feat(7562) | ker(16375) | probabilities(11463) | informatik(6708) | valuations(10374) | sequel(5189) | united(11024) | nocedal(9660) | parameterizing(13905 | spans(11528) |
| 8 | uses(6984) | tesauro(11232) | losing(9547) | stimuli(5881) | fun(9154) | top(5097) | neurones(10664) | timal(9179) | prints(12642) | optimiza(11289) |
| 9 | objectives(6782) | class(11209) | bounds(9532) | nodes(5388) | action(8983) | dat(5089) | learners(10086) | num(8464) | likeli(12439) | meth(11126) |
| 10 | recogni(6761) | erroneous(10517) | algorithm(9500) | visits(5320) | optima(7391) | num(4436) | lay(8202) | variable(6203) | bayesian(12043) | methodology(9935) |
| 11 | figueiredo(5657) | example(10125) | functioning(9240) | cells(5100) | probit(6928) | speculate(4425) | neuromorphic(8122 | algorithms(6030) | lof(10603) | gradi(9529) |
| 12 | loc(4782) | uses(9945) | distributing(9122) | frequencies(4910) | revow(6530) | use(4333) | syst(7574) | variation(5874) | posted(10216) | solu(9070) |
| 13 | detecting(4704) | perform(9620) | followed(8284) | differences(4633) | statements(6516) | uses(4149) | timal(7561) | clus(5790) | uses(9859) | nord(8893) |
| 14 | sessions(4624) | feat(9195) | pronounced(8229) | brain(4554) | regressors(5796) | contents(4100) | weightings(7509) | probit(5783) | proceeds(9552) | clustered(8653) |
| 15 | visible(4620) | use(8391) | erroneous(8020) | activity(4494) | contributions(5743 | informatik(3931) | hid(7482) | structurally(5282) | mead(9465) | converts(8564) |
| 16 | based(4179) | lab(8289) | sam(7962) | spiegelhalter(4187) | deciding(5062) | langford(3869) | figueiredo(7356) | graphon(5259) | infer(9401) | vec(8117) |
| 17 | usage(4058) | num(7811) | learners(7419) | signalling(4163) | stems(4983) | docu(3371) | weighing(6453) | clusterings(4982) | variable(8488) | dimension(7384) |
| 18 | visits(4052) | probit(7727) | rand(7419) | response(3912) | actions(4975) | firm(3330) | uniqueness(6350) | rand(4979) | distributionally(8147) | ranging(7173) |
| 19 | differences(3977) | feature(7711) | give(7411) | fields(3910) | sessions(4600) | probabilities(3276) | fun(6254) | infer(4692) | samples(8016) | analysis(6947) |
| 20 | segmental(3906) | fun(7594) | lof(7381) | film(3830) | uses(4505) | based(3273) | inputoutput(5684) | sixth(4448) | probabilities(7918) | algorithms(6770) |
| 21 | use(3881) | classified(7583) | lem(7185) | tardos(3814) | algorithms(4496) | sequence(3123) | systematically(510 | edelman(4409) | mixing(7783) | fun(6730) |

図 2　分類結果 (NIPS)

20news10

| init : python/lda.py -K 10 -N 100 ../data/20news.svm ../data/20news10 | | | | |
|---|---|---|---|---|
| host : MacBook-Pro.local | | | | |
| start : Thu Jun 14 23:15:44 2018 | | | | |
| iterations | K=10 | K=20 | K=50 | K=100 |
| 1 | 4816.9906 | 10229.4809 | 29526.8934 | 70085.3425 |
| 2 | 4609.9864 | 9271.2066 | 22807.6559 | 43372.204 |
| 3 | 4450.1037 | 8617.0145 | 19429.7332 | 33346.6996 |
| 4 | 4297.6149 | 8075.0771 | 16997.7106 | 26340.1311 |
| 5 | 4040.5517 | 7281.476 | 13777.6532 | 17715.7605 |
| 6 | 3463.8594 | 5581.9921 | 9493.7037 | 10547.7313 |
| 7 | 2662.7988 | 3823.7214 | 6278.6096 | 6881.2574 |
| 8 | 2017.7783 | 2862.89 | 4413.6537 | 5253.8824 |
| 9 | 1647.6506 | 2366.3603 | 3384.9055 | 4313.3511 |
| 10 | 1433.0599 | 2065.9294 | 2775.0284 | 3646.1837 |
| 11 | 1287.9515 | 1849.8985 | 2382.4345 | 3147.8253 |
| 12 | 1178.7159 | 1678.536 | 2112.7309 | 2771.7347 |
| 13 | 1095.8196 | 1538.3076 | 1917.5538 | 2489.1525 |
| 14 | 1031.4561 | 1427.0262 | 1773.0404 | 2279.1738 |
| 15 | 979.8862 | 1336.009 | 1661.0765 | 2113.3669 |
| 16 | 938.4404 | 1264.3987 | 1571.2865 | 1978.5991 |
| 17 | 905.3241 | 1205.9585 | 1501.145 | 1872.079 |
| 18 | 879.956 | 1157.1557 | 1438.7812 | 1784.3829 |
| 19 | 858.2175 | 1116.898 | 1387.9851 | 1708.557 |
| 20 | 838.6782 | 1082.7857 | 1343.1525 | 1648.3706 |
| 21 | 823.3959 | 1051.6527 | 1303.4391 | 1593.1111 |
| 22 | 809.3754 | 1026.5074 | 1268.5161 | 1544.3493 |
| 23 | 796.5903 | 1002.6479 | 1237.2257 | 1501.9633 |
| 24 | 785.1454 | 981.6751 | 1210.4385 | 1465.8333 |
| 25 | 774.5846 | 963.0247 | 1187.5984 | 1432.1324 |
| 26 | 764.538 | 946.2971 | 1164.3907 | 1402.536 |
| 27 | 755.8496 | 932.5745 | 1142.4258 | 1376.4301 |
| 28 | 748.5931 | 920.5568 | 1124.0151 | 1351.1551 |
| 29 | 741.0555 | 909.2316 | 1105.7683 | 1329.7036 |
| 30 | 734.6562 | 898.8651 | 1089.9422 | 1307.3551 |
| 31 | 728.2433 | 889.3882 | 1076.0167 | 1288.8999 |
| 32 | 722.3549 | 881.3617 | 1061.985 | 1271.0201 |
| 33 | 716.8778 | 873.6972 | 1049.2169 | 1256.1161 |
| 34 | 711.359 | 866.2851 | 1037.3297 | 1242.0478 |
| 35 | 706.8726 | 859.0688 | 1026.5005 | 1228.1106 |
| 36 | 703.0242 | 852.9457 | 1015.6332 | 1215.6185 |
| 37 | 699.4097 | 847.5022 | 1006.4468 | 1202.8154 |
| 38 | 695.5895 | 841.8681 | 996.0899 | 1191.0862 |
| 39 | 692.4746 | 837.7594 | 987.5151 | 1179.3885 |
| 40 | 689.4255 | 833.3554 | 980.2196 | 1169.3556 |
| 41 | 687.0549 | 829.5336 | 972.6243 | 1160.5408 |
| 42 | 684.3633 | 825.7575 | 965.67 | 1151.3372 |
| 43 | 680.978 | 822.4133 | 958.801 | 1142.6349 |
| 44 | 678.834 | 820.0105 | 952.4719 | 1134.3319 |
| 45 | 676.3321 | 816.551 | 947.7362 | 1127.4424 |
| 46 | 674.4186 | 813.1693 | 942.2508 | 1119.5347 |
| 47 | 672.2796 | 810.4345 | 937.2209 | 1112.5193 |
| 48 | 670.0307 | 807.9462 | 931.7273 | 1106.3314 |
| 49 | 668.0407 | 805.6166 | 927.2193 | 1099.7661 |
| 50 | 666.6097 | 803.8673 | 922.7438 | 1092.247 |
| 51 | 664.4537 | 802.0111 | 917.6371 | 1086.7287 |
| 52 | 662.9904 | 799.0414 | 912.6848 | 1078.9969 |
| 53 | 661.7184 | 797.7393 | 908.4856 | 1074.6068 |
| 54 | 660.4512 | 796.4366 | 905.264 | 1069.2171 |
| 55 | 658.7858 | 794.59 | 900.0513 | 1062.9725 |
| 56 | 657.4181 | 792.3301 | 896.7417 | 1057.9178 |
| 57 | 656.4683 | 791.0525 | 894.4497 | 1053.0367 |
| 58 | 655.2376 | 789.4837 | 891.8824 | 1048.4574 |
| 59 | 654.1623 | 787.6305 | 888.5332 | 1042.5021 |
| 60 | 653.669 | 785.8069 | 885.221 | 1038.0723 |
| 61 | 652.0529 | 784.9124 | 881.7036 | 1034.3563 |
| 62 | 651.8137 | 783.0927 | 878.863 | 1030.2148 |
| 63 | 650.7546 | 782.1885 | 876.212 | 1025.5224 |
| 64 | 650.1613 | 780.4948 | 873.949 | 1021.0909 |
| 65 | 649.8995 | 780.1356 | 870.4299 | 1016.7769 |
| 66 | 648.8786 | 779.1096 | 868.5471 | 1013.2581 |
| 67 | 648.1664 | 777.1417 | 866.777 | 1011.3042 |
| 68 | 647.4689 | 775.7517 | 864.9971 | 1008.59 |
| 69 | 646.8359 | 774.2701 | 862.4964 | 1005.8166 |
| 70 | 646.6387 | 774.3469 | 860.4241 | 1001.6029 |
| 71 | 646.744 | 772.6676 | 857.7747 | 999.0484 |
| 72 | 646.0338 | 771.8772 | 857.269 | 996.301 |
| 73 | 644.8099 | 771.1912 | 856.3836 | 992.6516 |
| 74 | 644.9121 | 770.8234 | 854.2225 | 989.5546 |
| 75 | 644.7649 | 769.658 | 853.2943 | 986.994 |
| 76 | 644.9692 | 768.3188 | 851.2948 | 985.0941 |
| 77 | 644.5771 | 767.6931 | 847.4843 | 982.059 |
| 78 | 644.1445 | 766.9494 | 846.1932 | 980.501 |
| 79 | 643.6702 | 765.9539 | 845.2719 | 977.0369 |
| 80 | 643.2793 | 764.7164 | 843.1635 | 974.0158 |
| 81 | 643.8623 | 763.9083 | 842.5981 | 971.7851 |
| 82 | 643.159 | 762.9837 | 840.6298 | 969.352 |
| 83 | 643.0495 | 762.6474 | 839.2993 | 966.5515 |
| 84 | 642.6977 | 761.8546 | 837.648 | 964.8817 |
| 85 | 642.2821 | 761.2457 | 835.9757 | 964.6909 |
| 86 | 642.1108 | 760.4839 | 834.4134 | 962.9233 |
| 87 | 641.8485 | 760.1327 | 833.5939 | 959.4271 |
| 88 | 641.556 | 759.3816 | 831.8814 | 958.6107 |
| 89 | 641.1224 | 758.7743 | 829.956 | 957.3997 |
| 90 | 641.0887 | 758.0517 | 829.3204 | 955.687 |
| 91 | 640.5934 | 758.1078 | 827.79 | 954.4301 |
| 92 | 640.7238 | 757.7022 | 827.2466 | 952.8073 |
| 93 | 640.5768 | 756.4649 | 826.5706 | 950.8608 |
| 94 | 640.0753 | 755.9914 | 825.0786 | 948.6142 |
| 95 | 640.0859 | 755.3541 | 824.4705 | 948.1992 |
| 96 | 640.4069 | 753.8771 | 823.2102 | 946.7244 |
| 97 | 640.2602 | 753.1402 | 822.8689 | 944.8147 |
| 98 | 639.9151 | 752.3027 | 822.3489 | 942.3723 |
| 99 | 639.7869 | 752.1882 | 820.9995 | 941.151 |
| 100 | 639.9225 | 752.2005 | 819.938 | 940.3563 |
| ========== | | | | |
| finish : Thu Jun 14 23:17:23 2018 | | | | |

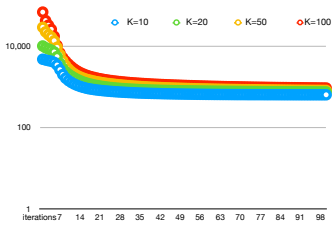


図 3　20newsgroups

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | topic0 | topic1 | topic2 | topic3 | topic4 | topic5 | topic6 | topic7 | topic8 | topic9 |
| 2 | .(56190) | ,(167328) | *(87090) | _(74661) | ,(74175) | >(264396) | I(62037) | "(32772) | =(101593) | .(65710) |
| 3 | ,(51675) | @(38888) | .(35877) | /(28206) | .(62292) | 'AX(61917) | .(32049) | ((30689) | =-(38883) | .(46386) |
| 4 | n(32507) | |(34458) | .(13002) | .(21259) | n(33501) | <(10440) | ?(23653) | "(28285) | .(18398) | nd(22926) |
| 5 | "(29384) | :(28753) | t(11044) | )(12855) | nd(32963) | .(9643) | n'(13631) | @(27489) | )(16040) | n(16804) |
| 6 | nd(20500) | edu(19890) | n(6679) | ((11686) | "(16084) | '(9476) | .(13407) | #(27108) | ((14540) | "(16235) |
| 7 | e(11194) | wre(12561) | ge(4976) | .(10880) | re(12620) | )(5587) | e(12810) | =+(26799) | .(14523) | )(12268) |
| 8 | ;(7726) | com(12525) | )(4782) | :(9675) | T(10679) | MAX(4551) | f(12233) | -251.54 | 1(10968) | ((11201) |
| 9 | r(7598) | ln(10861) | :(4668) | X(6374) | en(4935) | ?(3372) | r(9850) | /(24635) | .(10597) | T(10316) |
| 10 | l(7492) | rcle(10841) | ((4425) | n(5048) | e(4638) | 3(3314) | b(9207) | )(24568) | 2(7651) | e(7121) |
| 11 | T(6283) | ((9344) | e(3650) | l(4757) | d(4221) | Q(2590) | would(9141) | -24151 | 0(5475) | :(6913) |
| 12 | ll(6200) | c(9153) | nd(3411) | e(3179) | ce(4072) | -2412 | n(8409) | .(22798) | 4(4115) | ue(5368) |
| 13 | re(5940) | )(8732) | T(2594) | nd(3097) | ((3796) | [(1576) | ll(8016) | -21983 | 3(4073) | l(4748) |
| 14 | ((5191) | <(5040) | edu(2489) | f(2857) | ll(3786) | ](1342) | re(7931) | <(21549) | 5(3080) | b(4480) |
| 15 | )(5086) | =-–(4945) | l(2159) | drve(2803) | r(3560) | P(1246) | hng(7795) | .(21541) | 6(2305) | yem(3728) |
| 16 | people(4817) | NNTP-Pong-Ho(3069) | l(2049) | fle(2465) | )(3322) | G9V(1185) | ge(7665) | t(21076) | 10(2091) | m(3601) |
| 17 | en(4586) | Nnp-Pong-Ho(2840) | yer(2011) | r(2267) | l(3228) | "(1087) | nd(7544) | =(21039) | =+(1955) | "(3003) |
| 18 | ng(4256) | g(2399) | @(1901) | ](2183) | p(3174) | B8F(997) | kw(7382) | .(20819) | =-–(1856) | r(2877) |
| 19 | y(4075) | Reply-To(2251) | c(1804) | crd(2139) | A(3029) | N(834) | lke(7321) | ](20072) | 7(1583) | A(2835) |
| 20 | b(3937) | cc(1860) | =-–(1589) | wnw(2115) | fr(2900) | 7(832) | wn(6187) | .(19038) | |(1564) | fle(2644) |
| 21 | God(3658) | EDU(1779) | ply(1489) | |(2070) | dy(2878) | A86(727) | hnk(5742) | [(18830) | 8(1559) | d(2612) |

図 4　分類結果 (20news)

ja.text8_10

| init : python/lda.py -K 10 -N 100 ../data/ja.text8.svm ../data/ja.text8_10 | | | | |
|---|---|---|---|---|
| host : MacBook-Pro.local | | | | |
| start : Thu Jun 14 23:35:33 2018 | | | | |
| iterations | K=10 | K=20 | K=50 | K=100 |
| 1 | 49729.3304 | 109255.8528 | 340525.678 | 877794.7126 |
| 2 | 46834.2386 | 95324.6833 | 237574.3108 | 451548.8087 |
| 3 | 44364.9398 | 85304.6239 | 186475.6139 | 305123.5794 |
| 4 | 42116.6721 | 77390.8504 | 153358.4934 | 225250.7258 |
| 5 | 39870.2979 | 70050.704 | 126619.1985 | 169063.2407 |
| 6 | 37230.5342 | 62408.9848 | 101428.9896 | 122134.9733 |
| 7 | 33592.2686 | 52929.2297 | 75422.285 | 81499.0834 |
| 8 | 28250.2354 | 41011.8534 | 51666.7741 | 52247.0747 |
| 9 | 21980.598 | 29273.9833 | 34876.7788 | 34992.6293 |
| 10 | 16744.72 | 20908.9629 | 25086.5799 | 25649.9158 |
| 11 | 13252.6058 | 15963.8059 | 19583.9298 | 20434.9657 |
| 12 | 11128.9123 | 13163.3241 | 16356.1275 | 17328.8313 |
| 13 | 9819.0584 | 11500.613 | 14304.4508 | 15276.1306 |
| 14 | 8989.9992 | 10445.6493 | 12884.8845 | 13851.428 |
| 15 | 8439.638 | 9724.8281 | 11876.0837 | 12799.396 |
| 16 | 8055.1414 | 9185.9982 | 11097.1236 | 11982.7871 |
| 17 | 7776.6695 | 8780.7368 | 10502.0826 | 11353.8224 |
| 18 | 7557.4839 | 8462.1715 | 10022.1132 | 10823.4318 |
| 19 | 7390.956 | 8204.0489 | 9628.5126 | 10386.4086 |
| 20 | 7248.2802 | 7991.5203 | 9294.1051 | 10019.2521 |
| 21 | 7129.2706 | 7804.4208 | 9028.3284 | 9716.6371 |
| 22 | 7037.3467 | 7656.5842 | 8794.5548 | 9444.4361 |
| 23 | 6955.0379 | 7527.4507 | 8588.8315 | 9194.681 |
| 24 | 6884.2287 | 7417.9333 | 8417.2327 | 8979.1473 |
| 25 | 6825.3276 | 7316.5467 | 8260.2898 | 8797.9176 |
| 26 | 6771.4701 | 7228.5495 | 8127.7416 | 8625.3347 |
| 27 | 6728.7462 | 7159.0058 | 8011.6394 | 8467.9539 |
| 28 | 6683.8245 | 7086.5438 | 7894.3608 | 8330.4629 |
| 29 | 6644.5314 | 7028.3115 | 7796.3284 | 8204.5311 |
| 30 | 6608.2608 | 6973.1106 | 7706.5011 | 8094.1676 |
| 31 | 6578.9615 | 6917.4037 | 7631.0986 | 7979.0275 |
| 32 | 6552.7514 | 6876.5058 | 7556.5859 | 7889.6832 |
| 33 | 6526.2502 | 6835.5013 | 7487.9415 | 7812.1044 |
| 34 | 6504.8724 | 6798.0071 | 7418.9992 | 7728.2124 |
| 35 | 6490.2856 | 6761.776 | 7353.8739 | 7646.0761 |
| 36 | 6468.465 | 6724.8996 | 7299.2172 | 7570.9464 |
| 37 | 6450.3667 | 6695.1636 | 7251.3787 | 7499.8026 |
| 38 | 6439.0904 | 6668.3107 | 7198.9065 | 7439.6195 |
| 39 | 6419.0806 | 6639.6599 | 7150.7046 | 7382.1309 |
| 40 | 6408.66 | 6614.0633 | 7106.4106 | 7327.4439 |
| 41 | 6393.1797 | 6593.4034 | 7062.9728 | 7270.1758 |
| 42 | 6379.6231 | 6566.5932 | 7017.8223 | 7213.7625 |
| 43 | 6369.9722 | 6546.4647 | 6973.7323 | 7168.8779 |
| 44 | 6360.2147 | 6523.1019 | 6936.5177 | 7119.8114 |
| 45 | 6347.5266 | 6509.832 | 6902.8115 | 7076.9615 |
| 46 | 6337.4969 | 6491.0485 | 6873.8217 | 7042.1431 |
| 47 | 6324.4677 | 6465.927 | 6843.3144 | 7000.8714 |
| 48 | 6314.8094 | 6450.4418 | 6808.4703 | 6960.0037 |
| 49 | 6308.6367 | 6437.975 | 6776.1875 | 6914.7794 |
| 50 | 6300.0018 | 6426.1925 | 6752.8127 | 6880.0744 |
| 51 | 6291.8413 | 6409.7849 | 6725.0531 | 6853.0341 |
| 52 | 6280.5745 | 6400.6716 | 6705.6313 | 6827.0176 |
| 53 | 6271.4416 | 6389.1418 | 6677.3454 | 6798.0865 |
| 54 | 6256.8683 | 6376.5259 | 6654.551 | 6763.1704 |
| 55 | 6252.0311 | 6359.8811 | 6628.891 | 6737.5491 |
| 56 | 6244.9061 | 6352.1655 | 6610.0895 | 6711.4285 |
| 57 | 6241.361 | 6339.0754 | 6590.8981 | 6679.1003 |
| 58 | 6230.9497 | 6325.0206 | 6564.2774 | 6662.2381 |
| 59 | 6227.8238 | 6316.0969 | 6543.0578 | 6633.7591 |
| 60 | 6217.812 | 6300.9378 | 6522.8004 | 6608.3143 |
| 61 | 6212.4175 | 6294.9171 | 6502.0248 | 6588.1934 |
| 62 | 6206.2402 | 6286.0269 | 6484.7568 | 6565.2161 |
| 63 | 6200.445 | 6276.7611 | 6473.9599 | 6541.325 |
| 64 | 6189.9705 | 6272.6712 | 6445.8578 | 6525.2062 |
| 65 | 6184.5458 | 6260.305 | 6435.202 | 6512.0004 |
| 66 | 6179.1481 | 6254.2859 | 6424.4781 | 6492.1772 |
| 67 | 6178.578 | 6246.096 | 6405.4198 | 6477.0299 |
| 68 | 6174.7454 | 6238.1846 | 6389.7158 | 6456.1537 |
| 69 | 6170.2703 | 6228.9418 | 6371.7247 | 6444.0549 |
| 70 | 6163.5465 | 6220.4011 | 6362.3225 | 6424.0828 |
| 71 | 6161.7789 | 6218.1326 | 6345.7216 | 6407.7406 |
| 72 | 6155.9624 | 6214.9338 | 6336.3251 | 6389.1802 |
| 73 | 6152.3058 | 6207.2153 | 6325.1499 | 6374.7876 |
| 74 | 6147.5046 | 6200.1278 | 6314.6997 | 6359.9719 |
| 75 | 6143.1023 | 6193.648 | 6298.3069 | 6349.0785 |
| 76 | 6142.4223 | 6188.2998 | 6288.4069 | 6335.1133 |
| 77 | 6138.5873 | 6179.1052 | 6281.9532 | 6324.8443 |
| 78 | 6135.1951 | 6172.2443 | 6266.4917 | 6315.2219 |
| 79 | 6131.224 | 6164.9788 | 6258.5609 | 6302.4866 |
| 80 | 6129.2588 | 6159.0844 | 6245.9886 | 6295.148 |
| 81 | 6125.355 | 6157.3227 | 6234.0373 | 6283.427 |
| 82 | 6121.545 | 6150.4548 | 6223.7977 | 6274.942 |
| 83 | 6118.8429 | 6145.7063 | 6219.692 | 6265.8354 |
| 84 | 6115.4351 | 6140.5924 | 6214.9644 | 6253.877 |
| 85 | 6114.4625 | 6135.2185 | 6211.8522 | 6245.7059 |
| 86 | 6110.6314 | 6129.8766 | 6199.2362 | 6234.3273 |
| 87 | 6110.138 | 6123.0576 | 6190.9903 | 6224.1234 |
| 88 | 6107.4816 | 6119.8447 | 6182.6328 | 6210.877 |
| 89 | 6103.1488 | 6112.3223 | 6172.4976 | 6201.8912 |
| 90 | 6100.3749 | 6110.3757 | 6169.0926 | 6192.1034 |
| 91 | 6098.379 | 6101.5963 | 6167.4155 | 6185.6253 |
| 92 | 6095.2146 | 6097.3547 | 6157.5921 | 6168.0896 |
| 93 | 6094.584 | 6096.2734 | 6149.0142 | 6163.7752 |
| 94 | 6090.7585 | 6088.9997 | 6143.2361 | 6153.107 |
| 95 | 6087.2314 | 6086.8531 | 6143.6032 | 6145.2849 |
| 96 | 6086.3982 | 6082.3456 | 6135.4084 | 6134.4857 |
| 97 | 6083.8646 | 6075.1263 | 6123.8266 | 6124.7921 |
| 98 | 6083.3334 | 6070.8317 | 6120.1605 | 6120.9729 |
| 99 | 6082.9171 | 6063.8566 | 6116.9374 | 6113.4932 |
| 100 | 6082.6255 | 6056.6926 | 6112.7188 | 6098.0836 |
| ========== | | | | |
| finish : Thu Jun 14 23:38:37 2018 | | | | |

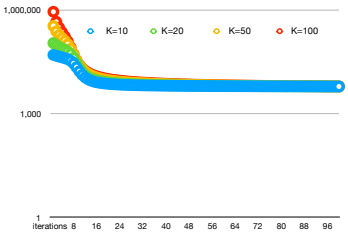




図 5　ja.text8

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | topic0 | topic1 | topic2 | topic3 | topic4 | topic5 | topic6 | topic7 | topic8 | topic9 |
| 2 | 2(16445) | 駅(14728) | 者(14715) | 時代(6613) | 軍(10102) | 的(21756) | ][(39860) | -8474 | 放送(10316) | 市(20622) |
| 3 | 1(14582) | 1(10415) | 人(10799) | として(6454) | 第(8817) | れる(21654) | [(35888) | 日本(7187) | など(9268) | 県(20550) |
| 4 | 3(11369) | -9413 | 日本(9366) | られ(6084) | 2(7592) | や(19407) | 人(8956) | 2(6880) | や(8760) | -11694 |
| 5 | -9206 | 2(8644) | 会(8255) | あっ(5604) | 1(6649) | ない(18523) | 作品(8578) | 3(6782) | 1(7670) | 町(11291) |
| 6 | 戦(8478) | 線(8287) | 的(7721) | 氏(5548) | この(6148) | など(14333) | 映画(7120) | 1(6555) | 番組(7388) | ((8062) |
| 7 | 4(8089) | 車(6168) | として(7074) | なっ(5332) | ため(5794) | もの(13782) | という(6973) | ((6301) | として(6661) | 郡(7873) |
| 8 | 5(6256) | 形(5717) | 大学(6885) | 家(5264) | へ(5372) | よう(13526) | ない(6796) | 曲(5821) | なっ(5654) | 区(6541) |
| 9 | 選手(6215) | なっ(5551) | 第(5549) | 世(5014) | あっ(5212) | ″(12857) | だ(6701) | 4(5417) | 3(5507) | 現在(5980) |
| 10 | 位(6174) | 式(5410) | 研究(5050) | れる(4732) | 機(5174) | この(10915) | なっ(5826) | として(5380) | 2(4670) | )(5625) |
| 11 | .(6140) | ため(5174) | や(4875) | 国(4720) | 3(4879) | ため(9735) | その(5694) | 賞(5208) | れる(4560) | 部(4704) |
| 12 | 6(6045) | 鉄道(5116) | など(4612) | 後(4608) | なっ(4684) | その(9698) | 作(5686) | 活動(4618) | 4(4299) | 地域(4655) |
| 13 | 出場(5744) | 3(5094) | 国(4257) | 王(4120) | 後(4529) | 性(9619) | や(5547) | 音楽(4362) | 会社(4267) | 村(4626) |
| 14 | チーム(5628) | 間(4206) | 委員(3907) | その(3923) | として(4307) | あり(9055) | として(5067) | -4129 | 開発(4070) | 昭和(4531) |
| 15 | 大会(5585) | として(4176) | 教育(3879) | 2(3826) | 4(3892) | として(8709) | 中(4842) | 10(4039) | まで(3865) | 地(4209) |
| 16 | 試合(5519) | へ(4104) | 社会(3843) | 3(3623) | -3876 | という(8667) | たち(4639) | 5(4005) | 者(3858) | 道(4185) |
| 17 | 回(5340) | 4(4094) | 学校(3661) | 城(3563) | 5(3846) | によって(8268) | よう(4445) | アルバム(3990 | ゲーム(3713) | 号(3966) |
| 18 | シーズン(4968) | まで(3750) | なっ(3649) | 三(3485) | アメリカ(3831) | なる(8179) | か(4274) | :(3982) | 販売(3600) | あり(3661) |
| 19 | なっ(4857) | 両(3702) | あっ(3610) | 代(3369) | まで(3745) | られる(8174) | だっ(4240) | .(3922) | ため(3579) | 川(3585) |
| 20 | 7(4682) | 列車(3547) | 法(3589) | この(3331) | ,(3719) | 場合(7481) | 本(4234) | 回(3650) | :(3542) | 州(3455) |
| 21 | 優勝(4662) | あっ(3432) | 主義(3423) | ため(3262) | られ(3681) | られ(7169) | 家(4217) | 6(3510) | また(3427) | 地区(3408) |

図 6　分類結果 (ja.text8)

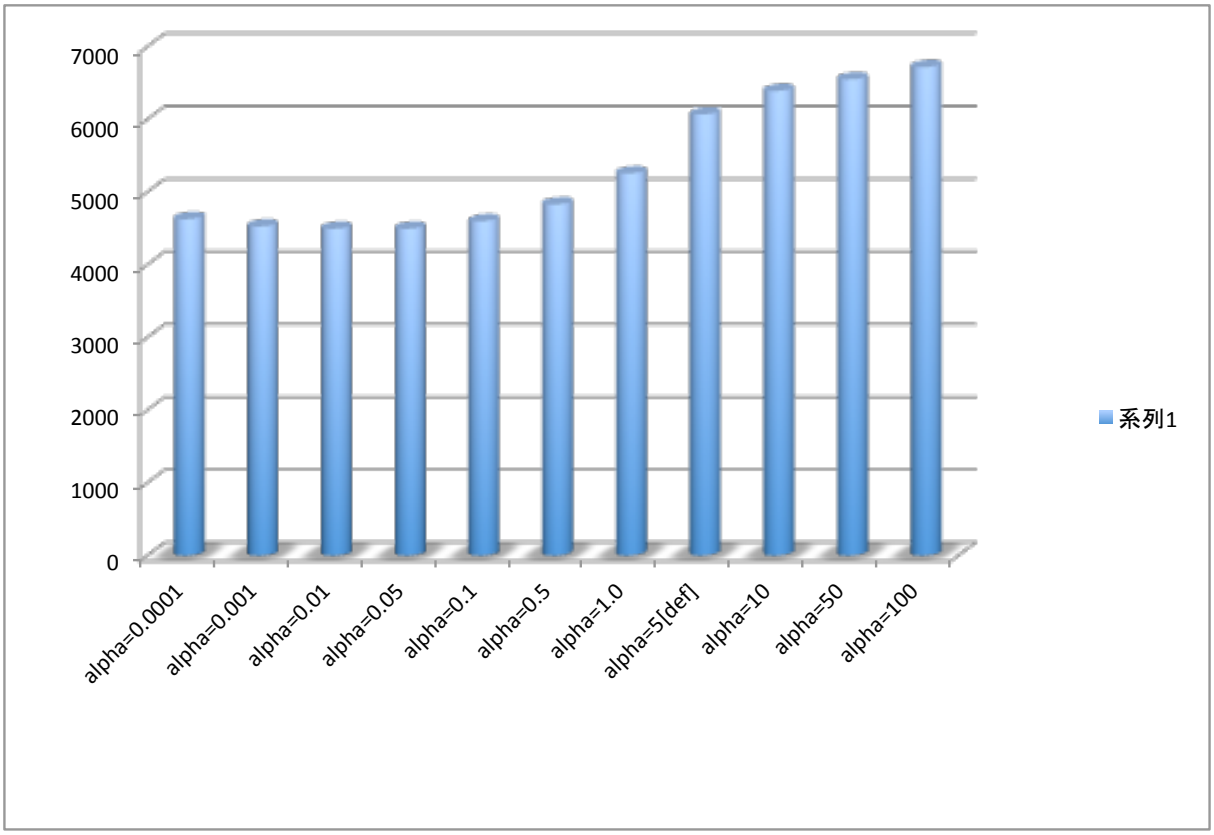


図 7　alpha を変化させた時の perplexity の変化

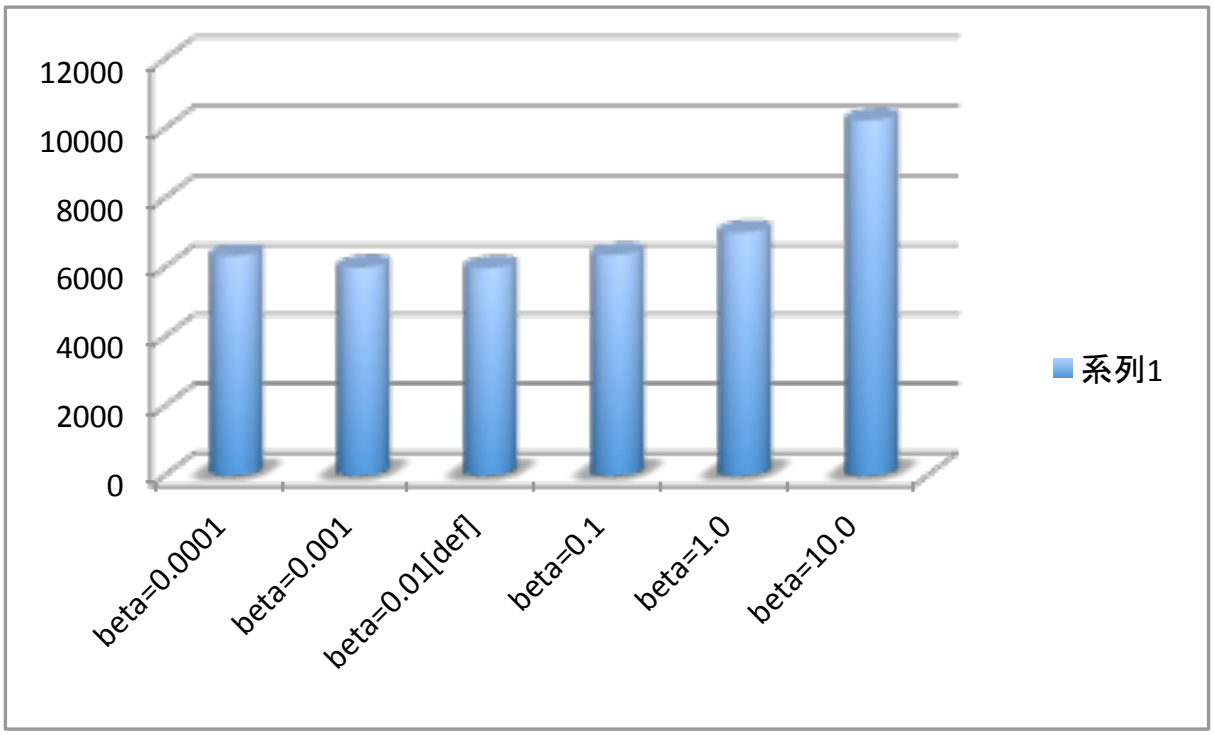


図 8　beta を変化させた時の perplexity の変化

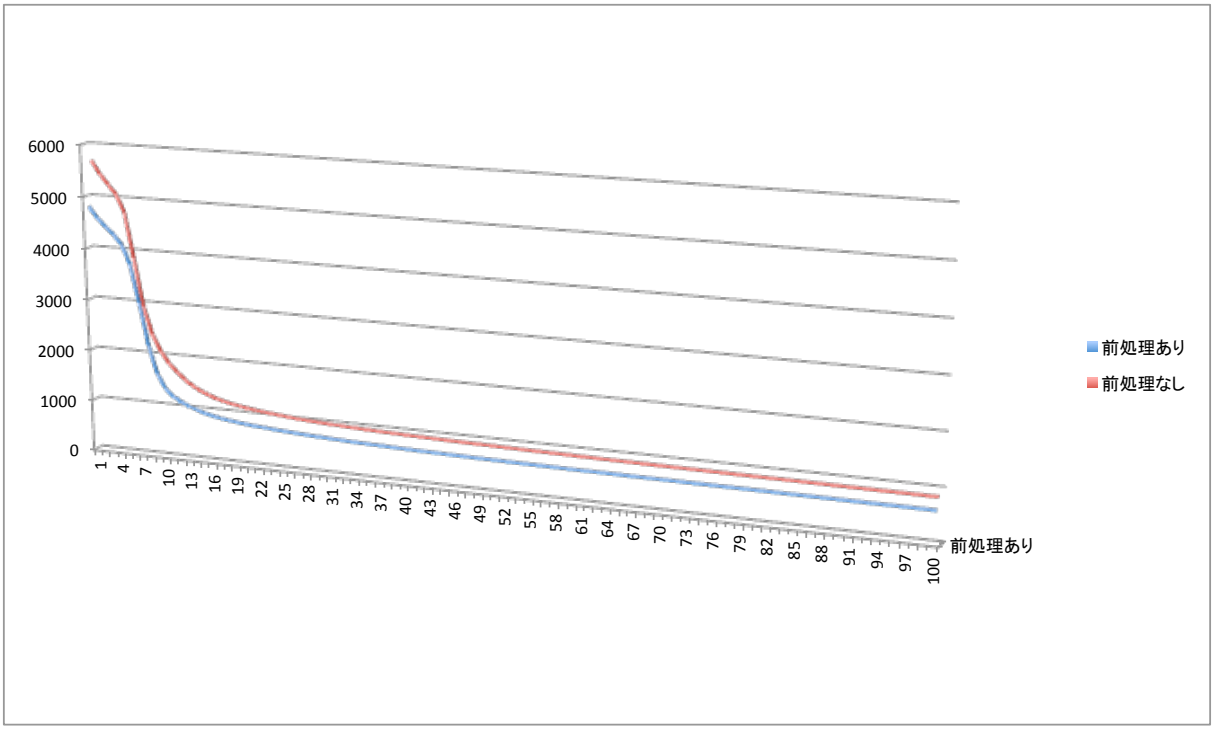


図 9　前処理を変化させた時の perplexity の変化

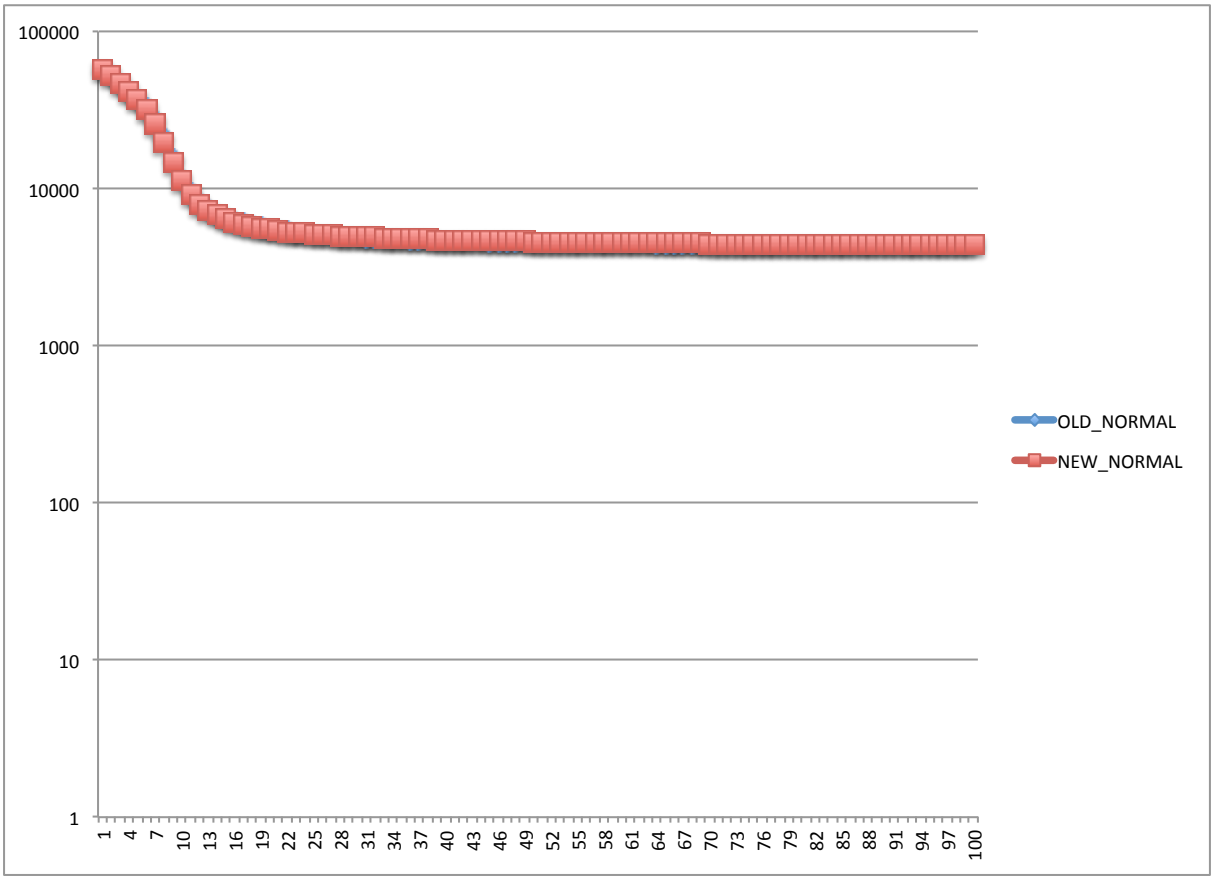


図 10　木村改善案と従来の lda.py

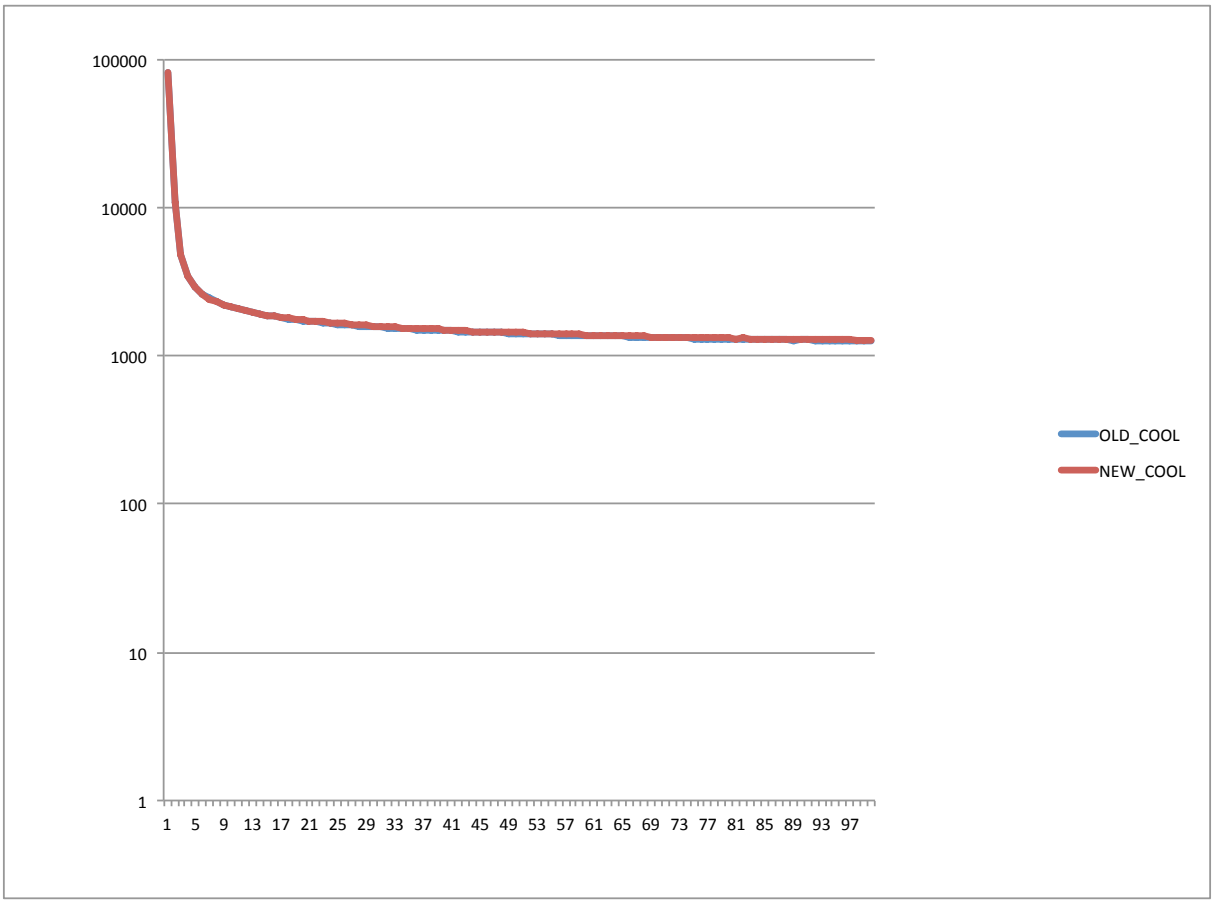


図 11　木村改善案と従来の lda.py(coolcutter.py 済み)

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | topic0 | topic1 | topic2 | topic3 | topic4 | topic5 | topic6 | topic7 | topic8 | topic9 |
| 2 | 駅(14651) | −7864 | 氏(5450) | 人(10023) | れる(13928) | 放送(6986) | 』(23650) | 2(12114) | 』(12094) | 日本(8980) |
| 3 | 市(10947) | 機(6964) | として(4718) | 軍(8915) | や(12873) | 者(6646) | 『(21410) | 1(10757) | 『(11006) | 県(7039) |
| 4 | 県(9432) | 2(6740) | 時代(4353) | 的(6796) | 的(11924) | や(6133) | −13474 | 3(8726) | ない(10799) | 市(6501) |
| 5 | 町(8111) | 1(5971) | られ(3759) | この(6630) | ″(10809) | など(5996) | 2(7475) | 戦(7951) | 人(9087) | 学校(6276) |
| 6 | 線(7523) | 車(5456) | なっ(3495) | 第(6322) | など(10570) | ない(5727) | ((7132) | −7745 | という(8829) | 大学(6124) |
| 7 | 1(6758) | ため(4589) | 国(3470) | なっ(6244) | ない(9732) | れる(4898) | として(7099) | 4(6293) | 的(8605) | 会(5128) |
| 8 | 郡(5831) | として(4522) | れる(3446) | として(6168) | よう(7749) | 1(4603) | 1(7013) | 選手(6164) | その(8474) | として(4670) |
| 9 | なっ(5334) | 3(4406) | 』(3393) | 者(5683) | もの(7714) | 番組(4344) | 3(6403) | 出場(5591) | や(7979) | 研究(4165) |
| 10 | .(4537) | なっ(4300) | あっ(3368) | あっ(5489) | として(7639) | として(4270) | 映画(5921) | 試合(5506) | よう(7607) | など(3919) |
| 11 | 2(4463) | 4(4004) | 3(3271) | 州(5244) | ため(6597) | 場合(3618) | 作品(5710) | チーム(5157) | なっ(7363) | 東京(3827) |
| 12 | 鉄道(4009) | 型(3762) | 2(3217) | 2(5236) | 性(6563) | なっ(3605) | 曲(5633) | 位(5128) | れる(6726) | 昭和(3823) |
| 13 | ,(3925) | 形(3539) | 後(3163) | 1(4990) | この(6312) | ため(3520) | 発売(5191) | 大会(4892) | ため(6648) | 者(3650) |
| 14 | や(3879) | 艦(3445) | 家(3126) | や(4745) | ((6043) | 3(3351) | 4(5091) | なっ(4789) | この(6565) | 第(3563) |
| 15 | −3808 | .(3296) | 城(3073) | 国(4666) | なる(6003) | 2(2681) | )(4764) | シーズン(4667) | だ(6395) | −3492 |
| 16 | 区(3807) | 用(3159) | 『(2941) | へ(4661) | 場合(5576) | について(2533 | 作(4505) | リーグ(4647) | 者(6308) | 高等(3264) |
| 17 | あり(3766) | あっ(3118) | など(2558) | ため(4392) | あり(5402) | まで(2525) | 第(4453) | 5(4592) | として(6124) | 教授(3127) |
| 18 | まで(3398) | 式(2966) | 県(2556) | 世(4318) | その(5263) | 4(2365) | 出演(4422) | 優勝(4565) | もの(5306) | 卒業(3045) |
| 19 | 現在(3353) | 5(2747) | 三(2499) | その(4278) | −5221 | 等(2270) | 音楽(4402) | 6(4529) | あっ(5305) | 教育(2902) |
| 20 | 3(3311) | 開発(2730) | ため(2467) | 後(4242) | )(5043) | また(2235) | 日本(4279) | として(4369) | られ(4876) | 区(2778) |
| 21 | 号(3278) | 使用(2671) | 5(2460) | ドイツ(3947) | られる(5040) | その(2235) | アルバム(3990 | 回(4238) | か(4662) | 3(2611) |

図 12　従来の lda.py

| | topic0 | topic1 | topic2 | topic3 | topic4 | topic5 | topic6 | topic7 | topic8 | topic9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 県(10335) | 軍(9791) | −7562 | れる(12319) | 的(13542) | 駅(14651) | 2(12569) | 氏(5458) | 』(15446) | 』(19144) |
| 3 | 日本(8742) | 第(6697) | 機(6056) | や(10534) | 者(12624) | 市(9472) | 1(11175) | として(5069) | 『(13606) | 『(17733) |
| 4 | 市(8566) | 人(6618) | 2(6002) | ″(10230) | や(10115) | 線(7716) | 3(9057) | 時代(4730) | 放送(9892) | −9480 |
| 5 | 学校(6132) | 2(5767) | ため(5958) | 的(8863) | ない(8979) | 1(7271) | 戦(8010) | 』(4611) | 番組(7329) | として(6852) |
| 6 | 大学(4954) | 1(5359) | 1(5769) | ない(8606) | れる(7735) | 県(7191) | −7932 | なっ(4391) | 1(6739) | 映画(6554) |
| 7 | 昭和(4509) | この(4961) | や(5749) | など(8556) | など(7676) | 町(6025) | 4(6487) | られ(4352) | 3(5929) | ((6400) |
| 8 | −4358 | 州(4893) | として(5521) | もの(6909) | として(7462) | なっ(5858) | 選手(6180) | れる(4162) | 2(5923) | 作品(6029) |
| 9 | 会(4256) | なっ(4615) | 車(5106) | よう(6824) | 人(7422) | 2(4963) | 出場(5621) | 『(4032) | として(5695) | 人(5976) |
| 10 | 東京(4029) | あっ(4200) | なっ(4852) | として(6293) | ため(5908) | 鉄道(4695) | 試合(5502) | あっ(4001) | 発売(5278) | や(5917) |
| 11 | として(3788) | へ(4178) | など(4819) | この(6081) | その(5356) | 郡(4566) | 位(5458) | 国(3499) | 4(4875) | なっ(5329) |
| 12 | 町(3723) | 世(4140) | れる(4350) | 性(5358) | なっ(5089) | や(4343) | チーム(5152) | 家(3353) | −4625 | )(5017) |
| 13 | 第(3690) | 3(4046) | .(4215) | ため(5229) | という(4940) | −4047 | なっ(5043) | 後(3336) | や(4044) | その(4865) |
| 14 | 研究(3461) | 後(4028) | 使用(4081) | なる(5213) | よう(4891) | として(3921) | 大会(5009) | その(3274) | なっ(3989) | ない(4743) |
| 15 | 区(3450) | として(3664) | 開発(4076) | その(5025) | この(4864) | 3(3838) | 5(4872) | 2(3171) | など(3818) | だ(4635) |
| 16 | 高等(3257) | 艦(3415) | ない(4051) | あり(4878) | もの(4386) | .(3826) | 6(4813) | 3(3110) | まで(3796) | 的(4555) |
| 17 | 1(3203) | ため(3259) | 型(4043) | ((4810) | あっ(4173) | まで(3794) | シーズン(4682) | 城(3082) | アルバム(3487 | 家(4386) |
| 18 | など(3096) | まで(3246) | 用(3879) | られる(4693) | において(3942) | など(3621) | リーグ(4651) | や(3045) | 10(3467) | この(4378) |
| 19 | 卒業(2955) | ドイツ(3189) | 3(3854) | −4315 | について(3862) | ,(3571) | 優勝(4593) | など(3002) | :(3416) | という(4367) |
| 20 | 3(2926) | 4(3108) | 形(3670) | 1(4252) | 日本(3803) | へ(3564) | として(4563) | ため(2982) | テレビ(3408) | よう(4030) |
| 21 | 部(2879) | その(2941) | 4(3408) | によって(4094) | られ(3774) | あり(3556) | 回(4393) | この(2741) | 曲(3294) | れる(3994) |

図 13　木村の lda.py

| | topic0 | topic1 | topic2 | topic3 | topic4 | topic5 | topic6 | topic7 | topic8 | topic9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 台帳(55) | ゴッタルド(54) | 魔物(54) | 復刊(54) | DT(55) | JNN(54) | 背髄(54) | ファシズム(55) | セネガル(55) | ウィーヴァー(55) |
| 3 | 宇喜多(54) | マンシュタイン( | 那賀(54) | ガメラ(54) | 外貨(54) | 聖体(54) | 北半球(53) | 見張り(54) | 大師(55) | 鶏肉(55) |
| 4 | ホイッグ(54) | 社債(53) | 栃(54) | モルダー(54) | 電極(54) | LSE(54) | 大隅(52) | ガリバルディ(5 | 小野田(55) | iPhone(55) |
| 5 | 永楽(53) | 清貧(53) | 保土ケ谷(54) | ゴダール(53) | ニーダーザクセ | AR(54) | ショスタコーヴ | アラゴン(53) | クリニック(55) | コントラバス(54) |
| 6 | とち(53) | アシ(53) | 関所(53) | 天馬(52) | エリトリア(53) | スーパーボウ | 。欧州(51) | 吉林(53) | ひろしま(54) | コネクタ(54) |
| 7 | 昇級(53) | 財前(53) | サンプソン(53) | 電離(52) | 納付(52) | インターセプト( | MRT(51) | ドリル(53) | マッケイ(54) | コンスタンティノス(54) |
| 8 | ラグビーフット | 渋沢(52) | 利根(52) | 坂田(51) | 調布(52) | グラフィックス( | ワキ(51) | 替え歌(53) | 重症(53) | ドロイド(54) |
| 9 | 利家(53) | 安井(52) | 宮津(52) | ガイ(51) | 埼京線(52) | 初号(53) | オートマトン(51 | ニューハンプシ | 失調(53) | Systems(53) |
| 10 | ゴキブリ(53) | -((52) | 渡船(52) | アスキー(51) | ガンマ(52) | ADHD(53) | ユリ(50) | ナント(52) | ジロ・デ・イタリ | ゼウス(53) |
| 11 | 葛城(52) | 猟犬(51) | 町奉行(52) | 。ジャケット(51) | アンドロメダ(52 | コーシー(52) | 江口(50) | アストラル(52) | エルサルバドル | ホビー(53) |
| 12 | 笹(52) | 箕面(51) | ボランチ(51) | 喜多(51) | RT(51) | 新進党(51) | 半球(49) | ウェスト(51) | カルボニル(52) | 興国寺(52) |
| 13 | AGC(52) | 石和(51) | 城陽(51) | 阿修羅(51) | サイパン(51) | アーロン(51) | ソフトバンクホ | プファルツ(51) | ロケッツ(52) | Apple(52) |
| 14 | サリン(51) | 寝技(51) | ヌビア(51) | 金堂(51) | HUB(51) | ジェノバ(51) | 。電子(47) | 倭人(51) | 京子(51) | ゲオルク(51) |
| 15 | 棋戦(51) | 飯能(51) | 橿原(51) | KADOKAWA(5 | セカイ(51) | ジョイ(51) | 死球(47) | 赤飯(51) | みなと(51) | 峡谷(50) |
| 16 | 平戸(51) | 犬塚(51) | 南陽(51) | 七夕(51) | クォーク(51) | 蘊(51) | DOHC(46) | 公館(51) | 新津(51) | OUT(49) |
| 17 | 宮内庁(46) | 小牧(50) | イノシシ(50) | エロ(50) | 潜航(50) | ブースター(50) | 火砲(46) | 光州(50) | シュトラウス(51 | 気流(49) |
| 18 | あめ(46) | 山車(49) | 清盛(50) | 椎名(49) | ガルベス(50) | エイト(49) | XX(44) | インドシナ(50) | 脂質(50) | 禁酒(47) |
| 19 | ゲン(46) | 十和田(49) | 帝室(49) | ヒグマ(49) | 地殻(49) | 岩屋(48) | グリル(44) | 佐原(50) | 経口(50) | 釜石(47) |
| 20 | 林道(46) | 内皮(49) | 戸山(49) | 任用(48) | アヘン(47) | カウントダウン( | 匡(42) | 竹島(49) | 光源(49) | ハリケーン(47) |
| 21 | 大月(45) | 長政(48) | ひずみ(49) | 氏康(48) | 鹿屋(47) | ライアン(47) | 茂雄(42) | ロワール(49) | 梗塞(49) | 参政(47) |

図 14　従来の lda.py（coolcutter.py 済み）

| | topic0 | topic1 | topic2 | topic3 | topic4 | topic5 | topic6 | topic7 | topic8 | topic9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 台帳(55) | ウィーヴァー(5 | 小野田(55) | アリストテレス( | 永楽(54) | 栃(54) | マキ(54) | 全豪オープン( | 聖体(54) | シュトラウス(54) |
| 3 | LSE(54) | セネガル(55) | 渋沢(54) | ガメラ(54) | 復刊(54) | コンスタンティ | ハーブ(54) | 中郡(54) | 醗酵(53) | スーパーボウル(54) |
| 4 | 宇喜多(54) | ゴッタルド(54) | ホイッグ(54) | マッケイ(54) | 田宮(54) | 江口(54) | ガリバルディ(5 | 飯能(53) | ラグビーフットア | インターセプト(54) |
| 5 | AR(54) | 背鰭(54) | モルダー(54) | 啓示(53) | 清貧(53) | TS(53) | 那賀(52) | 昇級(53) | 財前(53) | 関所(53) |
| 6 | ドロイド(54) | ひずみ(54) | マンシュタイン( | 身元(53) | 廃位(53) | new(52) | シェリー(52) | サンプソン(53) | RT(52) | ADHD(53) |
| 7 | 保土ケ谷(54) | 林道(53) | 興国寺(53) | クリニック(53) | 天馬(53) | ADSL(52) | シマ(52) | エヴァンゲリオ | ギルバート(52) | 官制(52) |
| 8 | ゼウス(53) | アシ(52) | JNN(53) | ジロ・デ・イタリ | 埼京線(52) | 警察庁(51) | バドミントン(52 | 西武鉄道(52) | カルボニル(52) | 葛城(52) |
| 9 | -((53) | トランシルヴァ | レア(53) | マント(53) | 倭人(51) | オートマトン(51 | 電離(52) | ニューハンプシ | ロケッツ(52) | 社債(52) |
| 10 | 家光(53) | ガンマ(52) | 内皮(53) | ヒグマ(52) | KADOKAWA(5 | 外貨(50) | ガルベス(52) | エリトリア(52) | ウェスト(51) | 調布(52) |
| 11 | 佐原(52) | 舷側(52) | ショスタコーヴ | インドシナ(52) | 公館(51) | ホビー(50) | ガイ(51) | 犬塚(52) | つん(51) | 渡船(52) |
| 12 | 利根(52) | アンドロメダ(52 | ヌビア(51) | OUT(52) | 出国(50) | DT(47) | DOHC(51) | アストラル(52) | 阿修羅(51) | 鹿屋(52) |
| 13 | 戸山(52) | ユニバーシアー | 金堂(51) | 三枝(51) | アスキー(50) | 竹本(47) | 校正(51) | AGC(52) | 七夕(51) | 参政(52) |
| 14 | ニーダーザクセ | 傳(51) | 栄一(51) | ゲオルク(51) | 不条理(50) | 薬剤師(46) | コーティング(51 | ポトマック(51) | ドリル(50) | 石和(51) |
| 15 | ディズニーラン | MRT(51) | 氏康(51) | 新城(51) | 関西大学(49) | ハエ(46) | Σ(51) | 棋戦(51) | ゴキブリ(49) | 宮津(51) |
| 16 | 箕面(51) | ワキ(51) | 蘊(51) | クォーク(51) | 掛け声(48) | 菊花賞(45) | ジェノバ(51) | カーネル(50) | 赤飯(49) | ギネス(51) |
| 17 | 壬生(51) | 光州(50) | ブースター(50) | 喪(49) | コントラバス(47 | AU(45) | グリル(50) | 直江(50) | 城陽(48) | 橿原(51) |
| 18 | 寝技(51) | キリル(50) | 脂質(48) | 着水(48) | ドロシー(47) | ファンク(43) | サンシャイン(4 | 山車(49) | エルサルバドル | HUB(51) |
| 19 | 芳賀(51) | 。事故(50) | 魔物(47) | サイパン(48) | 杉田(47) | Baby(43) | シティー(49) | サリン(49) | MY(47) | ボランチ(50) |
| 20 | 新津(51) | 十和田(49) | 十郎(47) | ポー(45) | あらわす(46) | 飛行船(43) | 鶏肉(49) | アラゴン(49) | ときめき(47) | ジョイ(50) |
| 21 | セカイ(51) | 竹島(48) | 別館(47) | 火砲(44) | 連戦(46) | 発声(40) | ℃、(49) | 竜王(48) | Java(46) | 清盛(50) |

図 15　木村の lda.py（coolcutter.py 済み）