Select number of topics for LDA model

Murzintcev Nikita

2015-09-08

Contents

Install package from github (only once for the first time). Remember that this is first public version of the package and it can contain errors.

```
devtools::install_github("nikita-moor/ldatuning")
```

Package `ldatuning` realize 4 metrics to select perfect number of topics for LDA model.

```
library("ldatuning")
```

Load "AssociatedPress" dataset from the `topicmodels` package.

```
library("topicmodels")
data("AssociatedPress", package="topicmodels")
dtm <- AssociatedPress[1:10, ]
```

The most easy way is to calculate all metrics at once. All existing methods require to train multiple LDA models to select one with the best performance. It is computation intensive procedure and `ldatuning` use parallelism, so do not forget to point correct number of CPU cores in `mc.core` parameter to archive the best performance. All standard LDA methods and parameters from `topimodels` package can be set with parameters `method` and `control`.

```
result <- FindTopicsNumber(
  dtm,
  topics = seq(from = 2, to = 15, by = 1),
  metrics = c("Griffiths2004", "CaoJuan2009", "Arun2010", "Deveaud2014"),
  method = "Gibbs",
  control = list(seed = 77),
  mc.cores = 2L,
  verbose = TRUE
)
```

```
## fit models... done.
## calculate metrics:
##   Griffiths2004... done.
##   CaoJuan2009... done.
##   Arun2010... done.
##   Deveaud2014... done.
```

Result is a number of topics and corresponding values of metrics

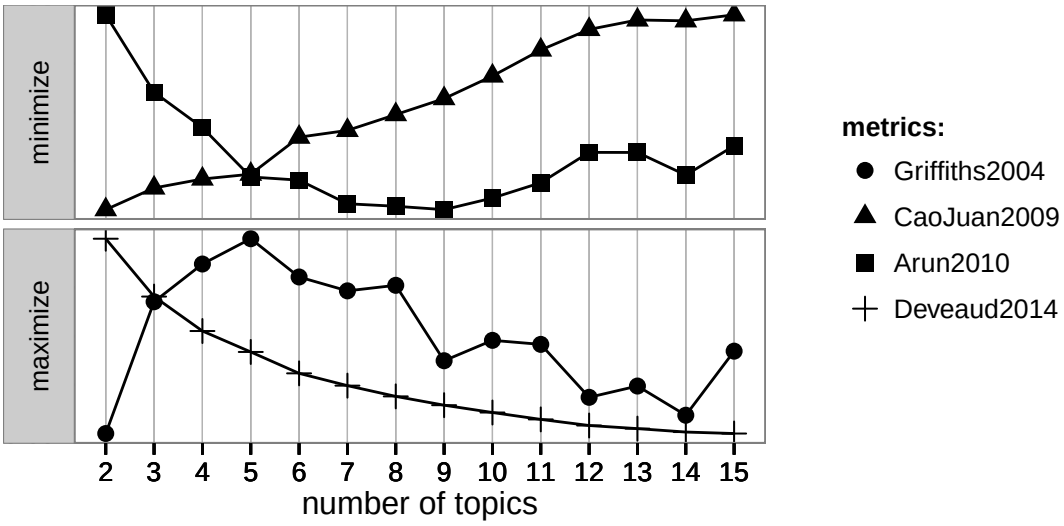
topics	Griffiths2004	CaoJuan2009	Arun2010	Deveaud2014
2	-15349.79	0.1169522	9.888687	0.6989189
3	-15266.66	0.1600736	8.959751	0.5318997
4	-15242.86	0.1779016	8.535274	0.4323482
5	-15226.91	0.1875260	7.942649	0.3718687
6	-15251.04	0.2612029	7.904418	0.3101625
7	-15259.80	0.2746812	7.621045	0.2746203
8	-15256.30	0.3061726	7.591150	0.2435689
9	-15303.87	0.3379840	7.550364	0.2181424
10	-15291.00	0.3829542	7.689896	0.1969989
11	-15293.55	0.4347111	7.873546	0.1770861
12	-15326.94	0.4756351	8.237908	0.1594651
13	-15319.82	0.4944709	8.236753	0.1504368
14	-15338.24	0.4927860	7.965343	0.1406462
15	-15297.82	0.5047240	8.316559	0.1362596

Simple approach in analyze of metrics is to find extremum, more complete description is in corresponding papers:

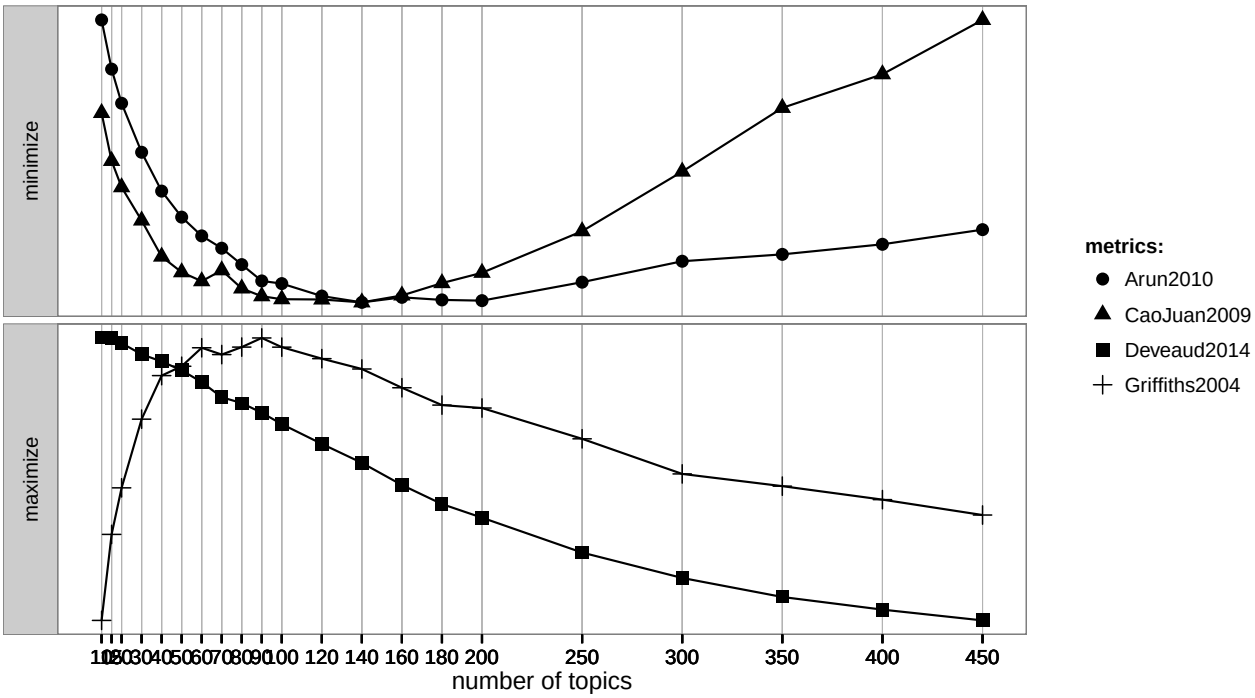
- minimization:
  - Arun2010 [1]
  - CaoJuan2009 [2]
- maximization:
  - Deveaud2014 [3]
  - Griffiths2004 [4,5]

For easy analyze of the results can be used support function `FindTopicsNumber_plot`

```
FindTopicsNumber_plot(result)
```



Results calculated on the whole dataset (about 10 hours on quad-core computer) looks like



From this plot can be made conclusion that optimal number of topics is in range 90-140. Metric Deveaud2014 is not informative in this situation.

References

1.Rajkumar Arun, V. Suresh, C. E. Veni Madhavan, and M. N. Narasimha Murthy. 2010. On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. In *Advances in Knowledge Discovery and Data Mining*, Mohammed J. Zaki, Jeffrey Xu Yu, Balaraman Ravindran and Vikram Pudi (eds.). Springer Berlin Heidelberg, 391–402. http://doi.org/10.1007/978-3-642-13657-3_43

2.Cao Juan, Xia Tian, Li Jintao, Zhang Yongdong, and Tang Sheng. 2009. A density-based method for adaptive LDA model selection. *Neurocomputing — 16th European Symposium on Artificial Neural Networks 2008* 72, 7–9: 1775–1781. http://doi.org/10.1016/j.neucom.2008.06.011

3.Romain Deveaud, Éric SanJuan, and Patrice Bellot. 2014. Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique* 17, 1: 61–84. http://doi.org/10.3166/dn.17.1.61-84

4.Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences* 101, suppl 1: 5228–5235. http://doi.org/10.1073/pnas.0307752101

5.Martin Ponweiser. 2012. Latent Dirichlet Allocation in R. Retrieved from http://epub.wu.ac.at/id/eprint/3558