



- Search Submit Search

- by [LinkedIn Learning](#)

We use your LinkedIn profile and activity data to personalize ads and to show you more relevant ads. [You can change your ad preferences anytime.](#)

☐ ☒

Ldaのモデル選択におけるperplexityの評価

- Share
- Like
- Download
- ...

☐ ☐ ☐

1/14

モデル選択において、様々なトピックモデルの手法に対して汎用的であるという理由を、しばしPerplexityが用いられる。ところが、今回実際に

Published in: [Data & Analytics](#)

0 Comments

5 Likes

Statistics

Notes

LDAのモデル選択における
Perplexityの評価

Full Name
Comments
12 hours ago [Delete Reply Block](#)

Are you sure you want to delete this comment?
[Yes No](#)

東京農工大学工学部情報工学科2年 森尾 学
2016/02/01

Share your thoughts...

Post

• Be the first to comment



[Shin Haya](#)
4 months ago



[Koichiro Kondo](#), [Data Scientist at Dentsu Aegis Network at Dentsu Aegis Network](#)
9 months ago



動機と目的



[Masanao Ochi](#), [student at Univ.](#)
2 years ago
統計的潜在意味解析において、Perplexityは依然として汎用的かつ重要な評価指標となっている。
- 例えば、トピック数Kを決定するモデル選択など



[Chang, Jonathan](#), [et al.](#) 2009)によって、Perplexityが必ずしも人間の評価に即さない可能性が指摘される。



[nitorodays](#)
2 years ago
本実験では、同一手法（LDA）を用いてPerplexityと人間の評価を比較することにより、モデル選択での有用性の是非を議論する。

No Downloads

Views

Total Views

1,754

On SlideShare

0

From Embeds

0

Number of Embeds

41

Actions

Shares

0

Downloads

17

Comments

0

Likes

5

Embeds 0

No embeds

No notes for slide

Ldaのモデル選択におけるperplexityの評価

1. 1. LDAのモデル選択における Perplexityの評価 東京農工大学工学部情報工学科2年 森尾 学 2016/02/01
2. 2. 動機と目的 □ 統計的潜在意味解析において、Perplexityは依然として汎用的かつ重要な評価指標となっている。 - 例えば、トピック数Kを決定するモデル選択など □ (Chang, Jonathan, et al. 2009)によって、 Perplexityが必ずしも人間の評価に即さない可能性 が指摘される。 □ 本実験では、同一手法（LDA）を用いて Perplexityと人間の評価を比較することにより、モデル選択での有用性の是非を議論する。

3. 統計的潜在意味解析とは？ 「トランプ氏は保守派のほか健康派の共和党支持者 や社会に不満を持つ人々など幅広い層の支持を集めて」 - NHKニュース

より、同じトピックに属する単語は共起しやすい。 - 例) 共和党, 社会, 保守など

4. 統計的潜在意味解析の前提条件

1. 単語の出現頻度が高い単語は、文書集合と呼ぶ。単語の集まり (順序は無視) を単語集合と呼ぶ

2. 文書は複数の潜在トピックを持つ - 政治, 宗教など

3. 各文書は潜在トピックを複数持つ。それらをトピック分布と呼ぶ

5. LDA (Latent Dirichlet Allocation) - 具体的には、文書が d 、文書 d がトピック k を持つ確率を $\theta_{d,k}$ とする。文書 d でトピック k が出現する確率を $\theta_{d,k}$ とする。トピック k が出現する確率を $\phi_k = (\phi_{k,1}, \phi_{k,2}, \dots, \phi_{k,V})$ とする。LDA は $\theta_{d,k}$ と ϕ_k をパラメータとする。LDA は Dirichlet 分布 (Dir · と表記) による生成を仮定する。

6. LDA (Latent Dirichlet Allocation) - 具体的には、文書が d 、文書 d がトピック k を持つ確率を $\theta_{d,k}$ とする。文書 d でトピック k が出現する確率を $\theta_{d,k}$ とする。トピック k が出現する確率を $\phi_k = (\phi_{k,1}, \phi_{k,2}, \dots, \phi_{k,V})$ とする。LDA は Dirichlet 分布 (Dir · と表記) による生成を仮定する。

7. LDA (Latent Dirichlet Allocation) - 具体的には、文書が d 、文書 d がトピック k を持つ確率を $\theta_{d,k}$ とする。文書 d でトピック k が出現する確率を $\theta_{d,k}$ とする。トピック k が出現する確率を $\phi_k = (\phi_{k,1}, \phi_{k,2}, \dots, \phi_{k,V})$ とする。LDA は Dirichlet 分布 (Dir · と表記) による生成を仮定する。

8. LDA (Latent Dirichlet Allocation) - 具体的には、文書が d 、文書 d がトピック k を持つ確率を $\theta_{d,k}$ とする。文書 d でトピック k が出現する確率を $\theta_{d,k}$ とする。トピック k が出現する確率を $\phi_k = (\phi_{k,1}, \phi_{k,2}, \dots, \phi_{k,V})$ とする。LDA は Dirichlet 分布 (Dir · と表記) による生成を仮定する。

9. LDA (Latent Dirichlet Allocation) - 具体的には、文書が d 、文書 d がトピック k を持つ確率を $\theta_{d,k}$ とする。文書 d でトピック k が出現する確率を $\theta_{d,k}$ とする。トピック k が出現する確率を $\phi_k = (\phi_{k,1}, \phi_{k,2}, \dots, \phi_{k,V})$ とする。LDA は Dirichlet 分布 (Dir · と表記) による生成を仮定する。

10. LDA (Latent Dirichlet Allocation) - 具体的には、文書が d 、文書 d がトピック k を持つ確率を $\theta_{d,k}$ とする。文書 d でトピック k が出現する確率を $\theta_{d,k}$ とする。トピック k が出現する確率を $\phi_k = (\phi_{k,1}, \phi_{k,2}, \dots, \phi_{k,V})$ とする。LDA は Dirichlet 分布 (Dir · と表記) による生成を仮定する。

11. LDA (Latent Dirichlet Allocation) - 具体的には、文書が d 、文書 d がトピック k を持つ確率を $\theta_{d,k}$ とする。文書 d でトピック k が出現する確率を $\theta_{d,k}$ とする。トピック k が出現する確率を $\phi_k = (\phi_{k,1}, \phi_{k,2}, \dots, \phi_{k,V})$ とする。LDA は Dirichlet 分布 (Dir · と表記) による生成を仮定する。

12. LDA (Latent Dirichlet Allocation) - 具体的には、文書が d 、文書 d がトピック k を持つ確率を $\theta_{d,k}$ とする。文書 d でトピック k が出現する確率を $\theta_{d,k}$ とする。トピック k が出現する確率を $\phi_k = (\phi_{k,1}, \phi_{k,2}, \dots, \phi_{k,V})$ とする。LDA は Dirichlet 分布 (Dir · と表記) による生成を仮定する。

13. 実験方法

□ Group1 - ランダムに選んだ10000文書を用いる。 - (訓練データ: テストデータ) = (5: 5)とする

□ Group2 - ランダムに選んだ70000文書を用いる。 - (訓練データ: テストデータ) = (8: 2)とする

日本語版Wikipedia から得た文書集合において、出現回数が5未満又は全文書集合のうち出現頻度が0.01を超過する単語を除外して作成した辞書とコーパスを用いた。それぞれのグループによって得られたモデルに対し、トピック数5, 50, 150のPerplexityとMPを集計する。

14. 実験結果

□ Group1 トピック数50, 150のときのPerplexityはそれぞれ512.7, 603.7

15. 実験結果

□ Group2 トピック数50, 150のときのPerplexityはそれぞれ200.7, 195.4

16. 議論

□ 人間による評価ではモデルの良し悪しは - Group1とGroup2で「ほとんど変わらない」 - しかしPerplexityの示す性質は劇的に変化

17. 議論

東京大学の佐藤一誠の著書 [佐藤一誠, 2015]によると…モデル選択について「Perplexityは、汎用能力を表す指標なので、Perplexityを用いてトピック数を決定する方法が考えられる。」と述べており、その過程で「複数のトピック数で比べて、Perplexityが最も低いものを選択する。」と手法について述べている。

18. 議論

ところが、今回得られた結果ではPerplexityは必ずしもモデル選択において人間的に正しい結果を示さないことを示唆していることがわかった。すなわち…

19. 結論

前に述べた「複数のトピック数で比べて、Perplexityが最も低いものを選択する。」という手法は人間にとって有益なモデルを選択するのに全く役に立たない。

統計的潜在意味解析とは？

20. 参考文献

Blei, David M., Ng, Andrew Y. and Jordan, Michael I. 2003. Latent dirichlet allocation. s.l. : The Journal of Machine Learning Research, 2003, pp. 993-1022.

Chang, Jonathan, et al. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. s.l. : Advances in Neural Information Processing Systems, 2009, pp. 286-294.

佐藤一誠. 2015. トピックモデルによる統計的潜在意味解析. 東京: コロナ社, 2015. 4339027588.

文書の集まりを文書集合と呼ぶ

単語の集まり (順序は無視) を単語集合と呼ぶ

JOIN BROADRIDGE FOR:

Modernizing legacy systems in Japanese financial institutions - Should we fix it, if it's not broken?

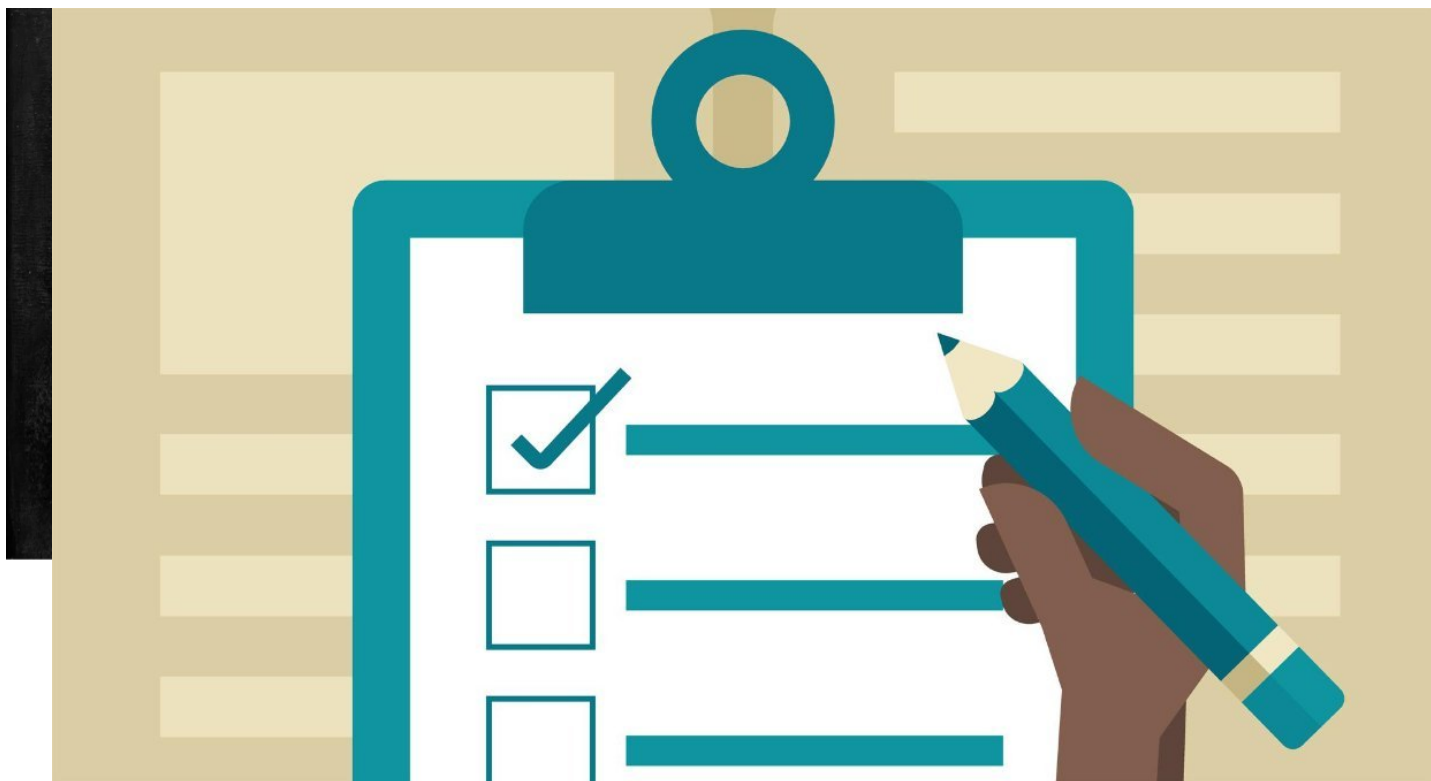
11:50am | Thursday 4 July

Ready for Next

waterstechnology

Tokyo
Financial
Information
& Technology
Summit





[Teaching Techniques: Writing Effective Learning Objectives](#)

[Online Course - LinkedIn Learning](#)



[Learning Online Marketing](#)

[Online Course - LinkedIn Learning](#)

- [Test Prep: PSAT](#)
- [Online Course - LinkedIn Learning](#)

【論文紹介】
3.7 評価方法 - 3.9 モデル選択 #トピ本
2018/06/26
@hoxo_m

[トピックモデルによる統計的潜在意味解析読書会 3.7 評価方法 - 3.9 モデル選択 #トピ本](#)
[hoxo_m](#)

【論文紹介】
トピックモデルの分散パラメータ
の事前分布について
2018/06/26
@hoxo_m

トピックモデルの分散パラメータの事前分布について
[hoxo_m](#)

latent Dirichlet Allocation)

Objectnessとその周辺技術
[Takao Yamane](#)

ポイント

を振るのが多項分布によるサンプリング

の形を決めるのがDirichlet分布

Visualizing Data Using t-SNE

Visualizing Data Using t-SNE

Tomoki Hayashi

トピックモデルの話

Shunsuke Takei

-

- [English](#)
- [Español](#)
- [Português](#)
- [Français](#)
- [Deutsch](#)
- [About](#)
- [Dev & API](#)
- [Blog](#)
- [Terms](#)
- [Privacy](#)
- [Copyright](#)
- [Support](#)

ポイント

実際にDirichlet分布のパラメータは**推定**する必要がある

マルコフ連鎖モンテカルロ法的一种である「**Collapsed Gibbs Sampling**」が一般的

導出過程は省略（かなり複雑）

Share Clipboard

X

- Facebook
- Twitter
- LinkedIn

Link

Public clipboards featuring this slide

x



No public clipboards found for this slide

Select another clipboard

X



Looks like you’ve clipped this slide to already.

Search for a clipboard
perplexity ☐ モデルの汎用性能を表す指標 – 具体的には選択肢の数（Perplexity1000は1000単語まで 候補を絞れたことを意味する） ☐ 少ないほど「性能が良い」と言
Create a clipboard

You just clipped your first slide!

Clipping is a handy way to collect important slides you want to go back to later. Now customize the name of a clipboard to store your clips.

Name*

Description

Visibility

Others can see my Clipboard ☐

Save this presentation

われている ☐ 学習用文書とテスト用文書に分け、各学習...



単語侵入タスク □ Changらが提案した手法 □ 被験者にトピックごとに5つの確率上位の単語と1つの ランダムに選ばれた単語を示し、仲間はずれの単語を 指摘させ

る。 □ 「悪いモデル」の場合、被験者はランダムに選ぶ傾向 があるため、値は1/6...



単語侵入タスクの例 □ 以下の単語の中から「仲間はずれ」を選んでください りんご なす ベッキー きゃべつ とまと



単語侵入タスク □ 結果をMPとして次のように定義 $MPs LDA = \sum_{k=1}^K \delta(i=k, s) \cdot \omega_{k,s}$ number of K ここで, K, ω, k, i はそれぞれ, 得られたトピックの集合 K , 被験者によっ

て選ばれた仲間はずれの単語のインデックス...



実験方法 □ Group1 - ランダムに選んだ10000文書を用いる。 - (訓練データ: テストデータ) = (5: 5)とする □ Group2 - ランダムに選んだ70000文書を用いる。 -

(訓練データ: テストデータ) = (8:...



実験結果 □ Group1 トピック数50, 150のときのPerplexityはそれぞれ512.7, 603.7



実験結果 □ Group2 トピック数50, 150のときのPerplexityはそれぞれ200.7, 195.4



議論 □ 人間による評価ではモデルの良し悪しは - Group1とGroup2で「ほとんど変わらない」 - しかしPerplexityの示す性質は劇的に変化



議論 東京大学の佐藤一誠の著書 [佐藤一誠, 2015]によると… モデル選択について「Perplexityは、汎用能力を表す指標なので、Perplexityを用いてトピック数 を決定

する方法が考えられる. 」と述べており, その過程で「複数...



議論 ところが, 今回得られた結果ではPerplexityは必ずしもモデル選択において人間的に正しい結果を 示さないことを示唆していることがわかった すなわち…



結論 前に述べた「複数のトピック数で比べて、 Perplexityが最も低いものを選択する。」という手法は人間にとって有益なモデルを選択するのになんて役に立たない可

能性がある。



参考文献 □ Blei, David M., Ng, Andrew Y. and Jordan, Michael I. 2003. Latent dirichlet allocation. s.l. : The Journal of Machi...

Upcoming SlideShare

Loading in ...5

×