

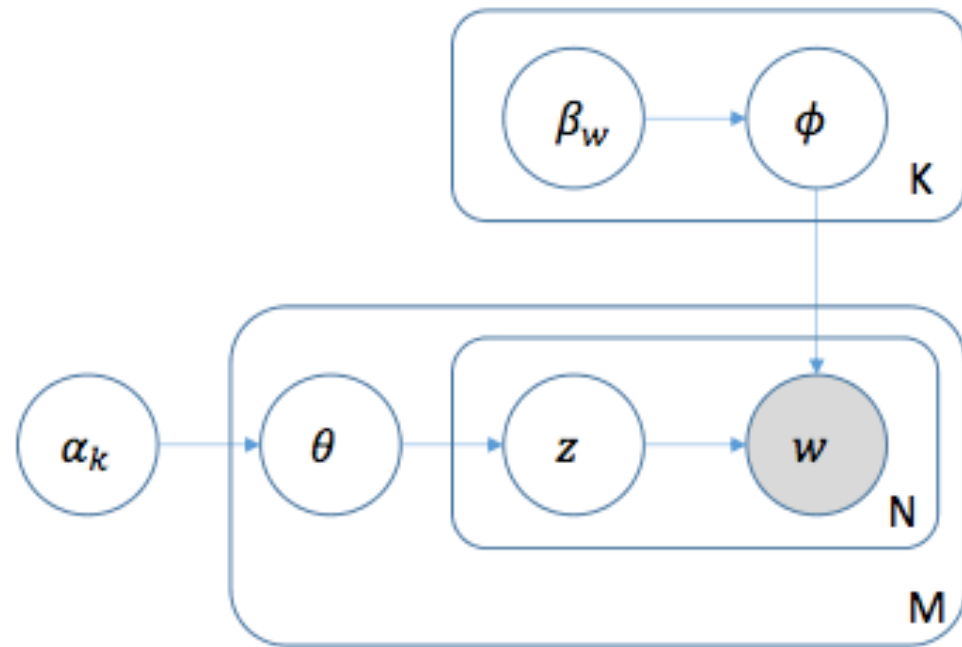
文書中の単語出現頻度を利用し たトピックモデル洗練化

高橋 仁 中川 博之 土屋 達弘

背景

- 開発者が自然言語を扱う機会が増えている
- 文章をトピックごとに分類するLDAが注目されている
- 膨大にある文章を分類しておくことで必要な文章を取り出しやすくなる
- しかし、そのままのLDAでは精度が低いため上手く分類することができない

LDA (Latent Dirichlet Allocation)

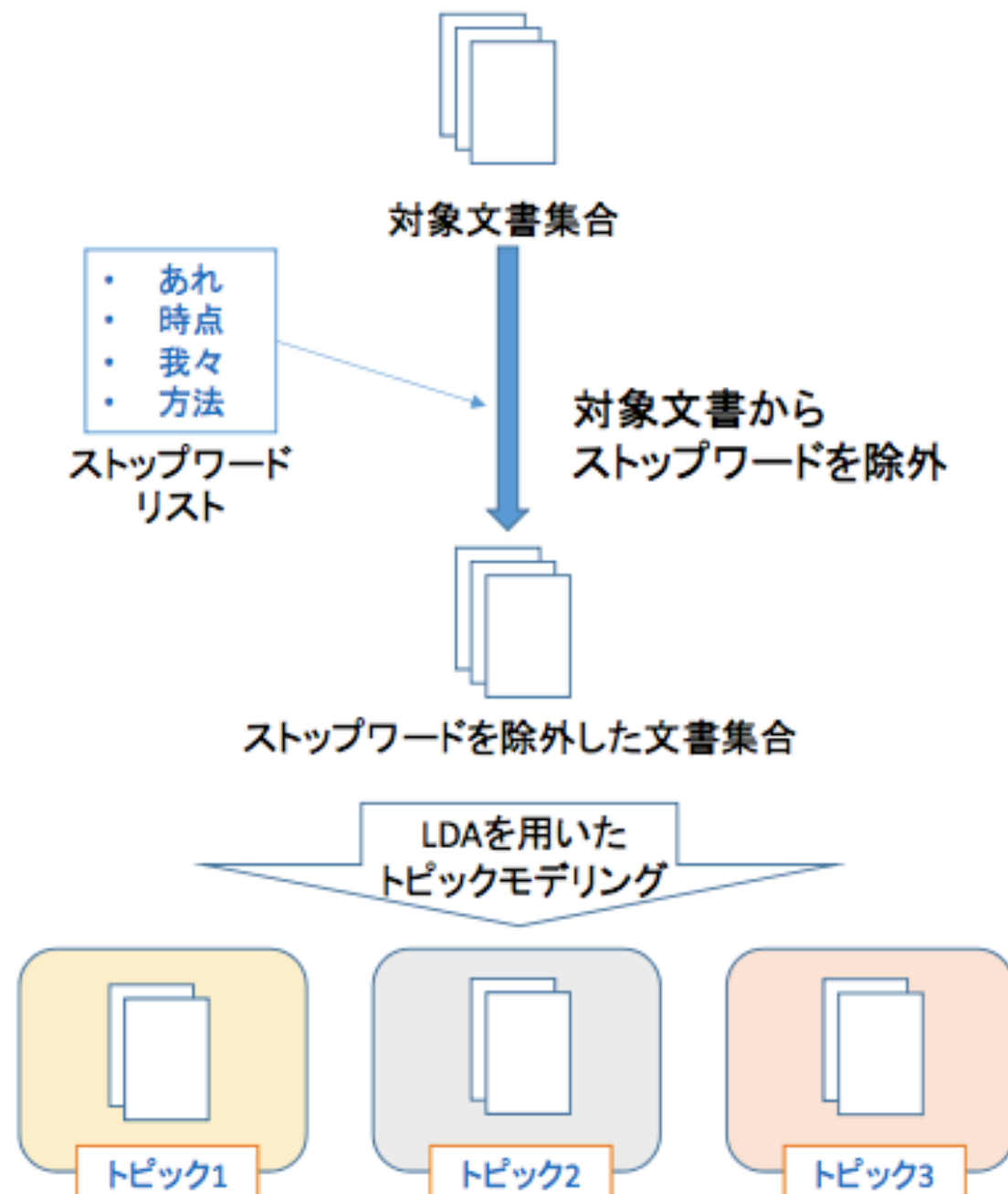


LDAのグラフィカルモデル

- 与えた文章集合の文章を特定のトピックに分類する手法
- トピックを構成している主要語を見ることがもできる

α_k : トピック混合率に関するハイパーパラメータ
 θ : 文書のトピック混合率
 z : 単語の背景トピック
 β_w : 単語の生成確率に関するハイパーパラメータ
 ϕ : 背景トピック k のときに単語 w が生成される確率
 w : 単語 w
 K : トピック数
 M : 文書数
 N : 1 文書中の単語数

ストップワード



● どの文章でも頻出する単語

- ストップワード[7]

● 正確なトピックモデリングをするためには前処理が必要

- 対象となる文章集合から分類の妨げになる単語を除外

● ストップワードはリストとしてまとめられる

- ストップワードリストは既に完成したものを用いる
 - Oracle Text[8] : 英語やドイツ語など12言語のリスト
 - SlothLib[9] : 日本語のリスト 310件のストップワード

[7] Rajaraman, A. and Ullman, J. D.: Mining of Massive Datasets, Cambridge University Press, New York, NY, USA (2011).

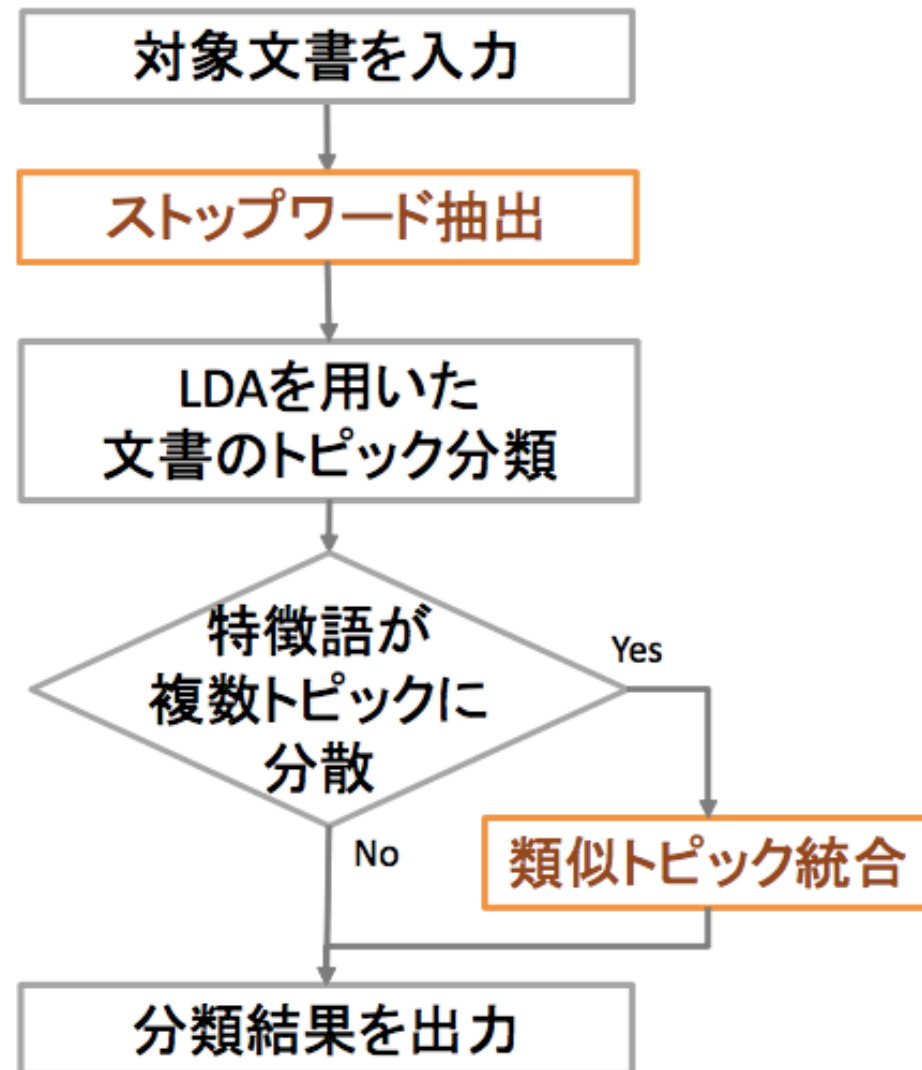
[8] Oracle Text で提供されるストップリスト, https://docs.oracle.com/cd/E16338_01/text.112/b61357/astopsup.htm.

[9] 大島裕明, 中村聡史, 田中克己: SlothLib: Web 検索研究のためのプログラミングライブラリ, 日本データベース 学会 Letters, Vol. 6, No. 1, pp. 113-116 (2007).

問題点

- 一般的なストップワードリストを分類に用いるには不十分
 - テスト記述：確認・押下などが頻出
- 単語の構成が類似したトピックが出現する場合
同じトピックが複数トピックにばらけてしまう

提案手法



1. 文章中の単語からストップワードを作成

- 予め設定した閾値よりも高いDF値をとる単語ストップワードとする

$$DF_i = \frac{|\{d : d \ni t_i\}|}{|D|}$$

$|D|$: 総文書数

$|\{d : d \ni t_i\}|$: 単語 i を含む文書数

2. 類似トピックを統合

- tf-idf コサイン類似度[11][12]を利用したクラスタリングを行い、類似したトピックを統合

[11] Tata, S. and Patel, J. M.: Estimating the Selectivity of Tf-idf Based Cosine Similarity Predicates, SIGMOD Rec., Vol. 36, No. 2, pp. 7–12 (2007).

[12] Whissell, J. S. and Clarke, C. L.: Effective Measures for Inter-document Similarity, in Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management(CIKM'13), pp. 1361–1370, New York, NY, USA (2013), ACM.

実験の対象文章

◎ テストケース記述

- 結合テスト 日本語 142件

◎ ユーザレビュー

- Facebook for iOS (2015/8/5~9/8) 英語 3000件

◎ メーリングリスト

- Apache Commons User List 英語 1000件

テストケース記述

自然言語で書かれたテストケース記述

- A を入力として設定する
- B ウィンドウの表示が C であることを確認する
- C ボタンを押下する
- D を出力する

表 2 対象テストケース記述

機能	テストケース記述数
共通結合テスト	17
部品受注に関する結合テスト	18
部品引当に関する結合テスト	1
部品出荷に関する結合テスト	6
部品売上に関する結合テスト	5
部品在庫に関する結合テスト	2
部品発注に関する結合テスト	33
部品入荷に関する結合テスト	9
部品パッチに関する結合テスト	9
ラベル受注に関する結合テスト	4
ラベル引当に関する結合テスト	4
ラベル出荷に関する結合テスト	2
ラベル売上に関する結合テスト	2
ラベル在庫に関する結合テスト	2
ラベル発注に関する結合テスト	2
ラベル入荷に関する結合テスト	4
ラベルパッチに関する結合テスト	9
部品請求に関する結合テスト	13

実験1（テストケース記述）：評価基準

.....

$$Recall = \frac{\sum_g d_{gt_g}}{|D|} \quad (2)$$

$$Precision_g = \frac{d_{gt_g}}{d_{t_g}} \quad (3)$$

$$Precision = \frac{\sum_g d_{gt_g} \frac{d_{gt_g}}{d_{t_g}}}{|D|} \quad (4)$$

$$F - measure = \frac{2 * Recall * Precision}{Recall + Precision} \quad (5)$$

t_g : 機能 g のテストケース記述が最も多く集約されたトピック

d_g : 機能 g のテストケース記述数

d_{t_g} : トピック t_g に集約されたテストケース記述数

d_{gt_g} : トピック t_g に集約された機能 g のテストケース記述数

$|D|$: 総テストケース記述数

◎ 評価基準

▶ 分類した時の精度の比較

- ストップワードなし
- 通常のストップワード
- 提案手法

実験1（テストケース記述）：ストップワード抽出

.....

表 3 DF 上位の単語：テストケース記述

単語	DF	単語	DF
押下	1	更新	0.873239
こと	1	実行	0.866197
確認	1	処理	0.852113
する	1	結果	0.838028
遷移	1	ダンプ	0.838028
表示	1	入力	0.830986
画面	1	条件	0.816901
リンク	0.985915	登録	0.802817
ボタン	0.978873	前後	0.78169
情報	0.957746	明細	0.739437
照会	0.950704	後	0.725352
検索	0.922535	指示	0.711268
データ	0.915493	出力	0.711268
メニュー	0.915493	在庫	0.704225
選択	0.894366	修正	0.697183

- DF値の高い単語のうち
分類の指標となりにくいと
考えられる単語
- 0.7以上をストップワード
へ追加

実験1（テストケース記述）：類似トピックの統合

.....

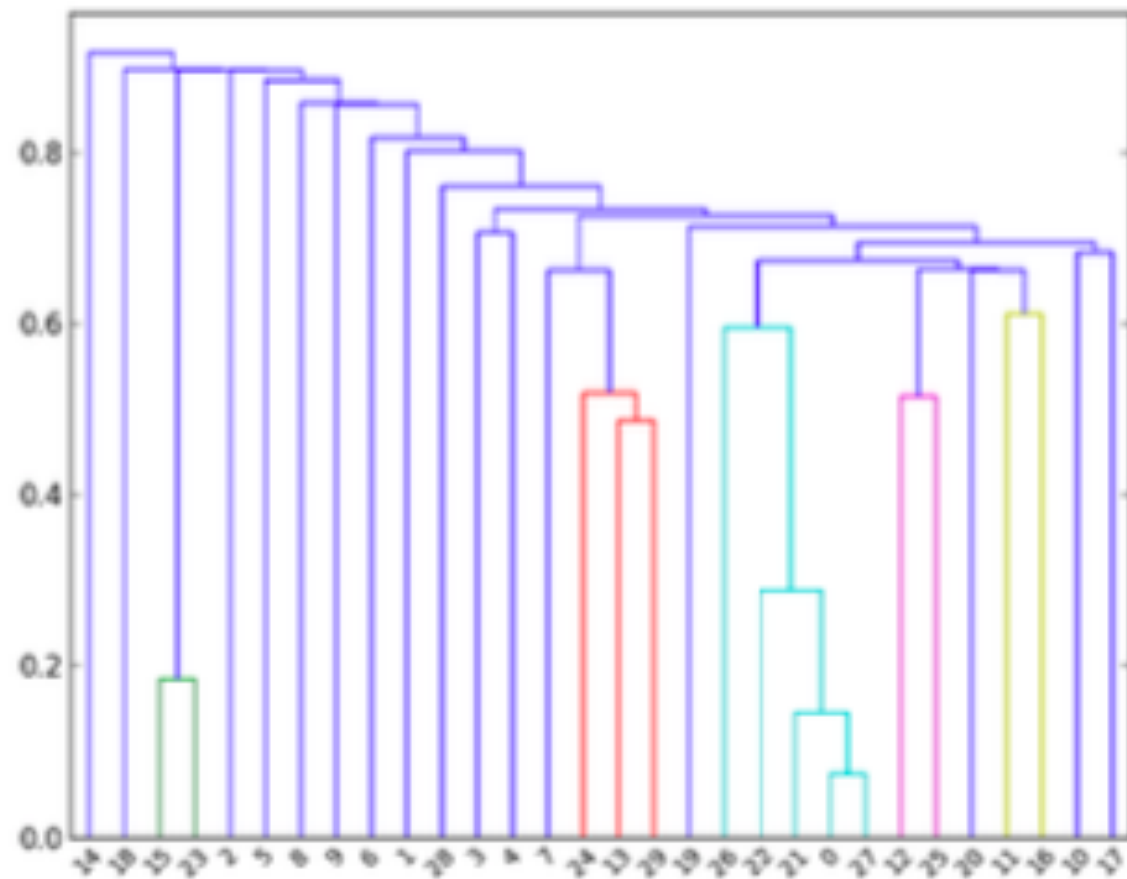


図 4 トピック間距離のデンドログラム：テストケース記述

- 全トピックの単語を入力として、トピック距離を計算（tfidfコサイン類似度）
- トピック間距離が0.6未満のトピックを統合する

実験1（テストケース記述）：結果

.....

表 4 テストケース記述分類の再現率，適合率，F 値

	ストップワード無し	通常	提案手法
再現率	0.634	0.641	0.718
適合率	0.454	0.469	0.548
F 値	0.529	0.542	0.622

- 再現率・適合率・F値が提案手法のほうが高い

実験2（ユーザーレビュー）:評価基準

.....

◎ 評価基準

- 通常と提案手法の比較
- トピックの主要語から各トピックで言及されている特定の機能が読み取りやすくなっているか
- 例：広告のサジェスト機能が自分に合っていないというレビュー

Title: Buggy, navigation is not good, too many ads

Body: ... And the ads that are "suggested" posts have no relevance to me.

実験2（ユーザーレビュー）：結果

表 5 トピックの機能ラベル：ユーザレビュー

通常		提案手法	
トピック	機能	トピック	機能
0	写真	0	通知
1	要望	1	フリーズ
2	不明	2	アップデート
3	ボタン	3	アップデート
4	アップデート	4	メッセージャー
5	メッセージャー	5	アカウント
6	アップデート, フリーズ	6	写真, クラッシュ
7	スペイン語レビュー	7	不明
8	不明	8	要望
9	不明	9	ビデオ
10	ビデオ	10	ニュースフィード
11	画面	11	フレンド
12	ニュースフィード	12	バッテリー
13	フリーズ, クラッシュ	13	アップデート
14	ページ	14	スペイン語レビュー
15	バッテリー	15	クラッシュ, フリーズ
16	コメント	16	ページ
17	ビデオ, 写真	17	広告
18	アカウント	18	評価するレビュー
19	評価するレビュー	19	写真

- トピック数を20して各トピックの主要語を出力
- 手動でラベル付けを行う
- トピック2・8・9が改善

実験3（メーリングリスト）:評価基準

.....

◎ 評価基準

- 通常と提案手法の比較
- 言及されているソフトウェアごとに分類する
- 例：Jellyというスクリプトのタグ付けに質問している

Subject: [Jelly] Import and Include tags

What are the differences between the import and include core tags? When to use one vs the other?

実験3（メーリングリスト）：結果

表 6 トピックの主要語集合：メーリングリスト，提案手法

トピック	主要語
0	public void idx protected class bigdecimal return boolean bean timestamp values long java object attributes org double string apache
1	jelly paul tag org apache tags var set foo script output libbrecht run xmlns java impl class activemath xml
2	org java apache jar dbcp user yahoo cli mysql mail jdbc http class mailto garrett gmane smith command john
3	jxpath test string getvalue dmitri xforms pointer context label configuratorcontext ldap user jxpathcontext daniel xml return org book field
4	betwixt apache org junit java class bean test xml abstractbeanwriter user beanreader xmlintrospector thread debug lang vfs martin torque
5	digester object string properties public property bean xml rule user file connecturi org class element help apache mailto key
6	java org apache catalina core source unknown invoke error net postgresql logging yahoo lang connection trace io reflect logfactory
7	apache org map user mailto object mail collections integer java stephen int listutils dynabean maputils help tomap predicatedlist moritz
8	mailto user mail apache org help digester october original users pm scott mark comp parameter nbsp rule gmane friday
9	test jelly knut null expected user wannheden mail xmlunit isempty tag org stringutils parse isspace xu lang gmane comp
10	james yahoo strachan http uk jelly xml file need ll page user weblogs web radio music sport news charts

.....

11	user comp gmane will could craig ve time work don make need ll different release class bug implementation good
12	org apache user validator mailto file help mail form http ftp gmane comp fileupload field problem tan license release
13	error javac jelly org apache ant java class maven symbol sandbox tags src peter home apps location resolve cannot
14	digester java org apache jar xerces xml servlet parse lang xmldocumentfragmentscannerimpl region impl validator web parsers nosuchmethoderror parser error
15	beanutils string java public org apache bean propertyutils custom lang class util void object certification user seacor problem return
16	httpclient user option org apache id http string post options header column create system main robert application problem help
17	digester object person mail intended john please builder information stack confidential httpclient scioworks build organisation sender pop recipient will
18	log logging apache org struts thread tomcat logger user debug class file jar messages safety properties appender web classes
19	pool connection dbcp apache org jdbc user parameter tomcat driver datasource connections java null getconnection sql conn comp idle

- Jellyに関するトピック1・9・10・13がまとまっていると考えられる
- JDBC Utility Componentに関するトピックが2と19であると考えられる

実験3（メーリングリスト）：結果

.....

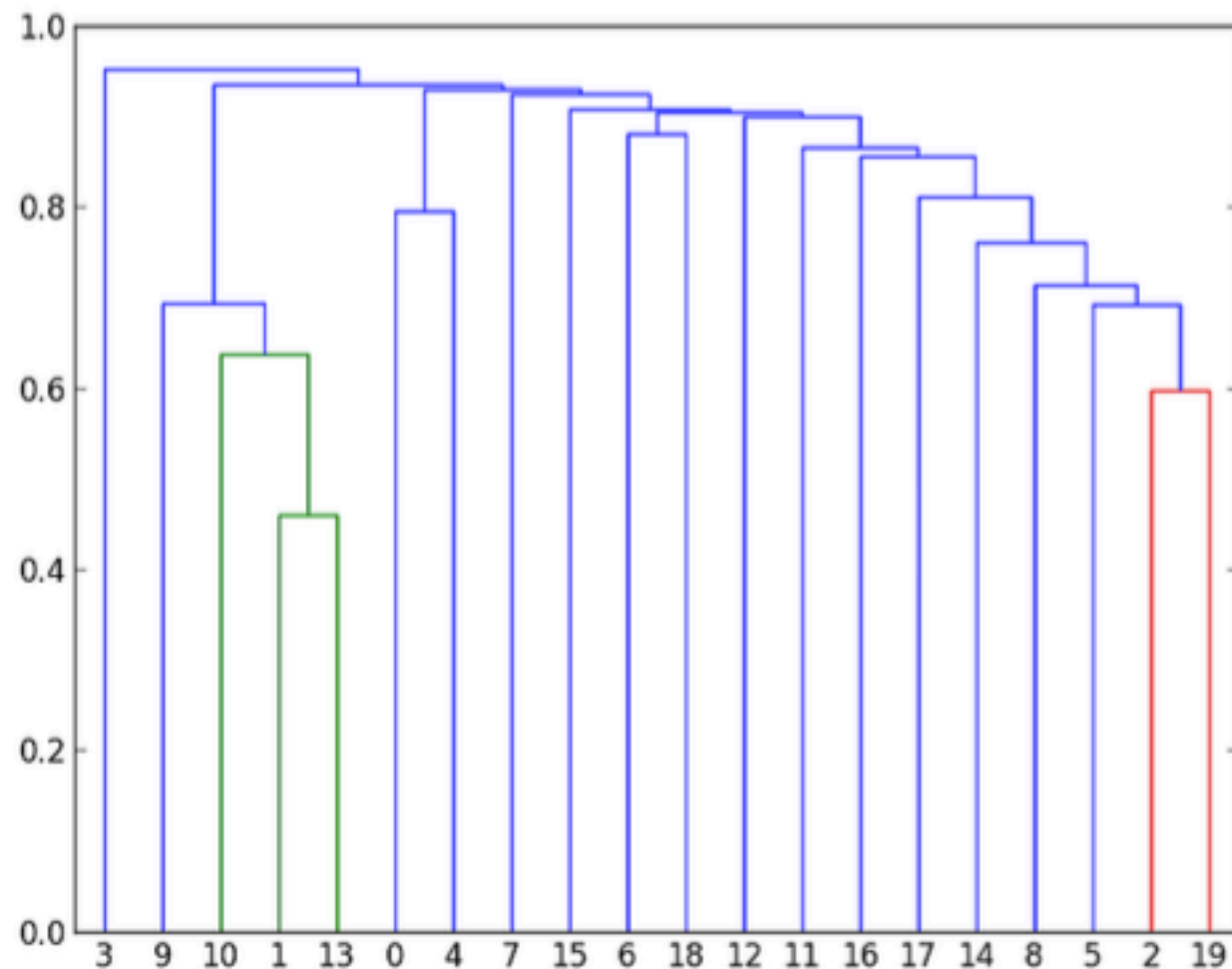


図 5 トピック間距離のデンドログラム：メーリングリスト

- トピック9・10・1・13の距離が近いことがわかる
 - Jellyのトピックが統合される
- トピック2・19の距離が近い
 - JDBCのトピックが統合される

まとめ

- LDAをを自然言語文章に適応する時に発生する問題を特定した
- ストップワードを追加することで適切なストップワードリストを作成する手法を提案した
- 分類したトピックの類似度を算出することで、トピック分類の正確性を向上させる手法を提案した

所感

- ◎ LDAを使う上で非常に重要な手法だと感じた
- ◎ 閾値を自動で決めれば・・・
- ◎ LDAの精度をあげる研究は他にも存在するのか？
- ◎ あれば、精度の違いはどうか