# LDAを用いた文書のトピック判別

2020/1/10　ロボット設計製作論実習5
千葉工業大学　先進工学部　未来ロボティクス学科
上田研究室　17C1112　能澤貴弥

## 背景

- 同一の単語でも違う意味を持っている場合や違う対象を指す言葉が増加
  例
  - リンゴのappleと社名のapple
  - 他分野での人名
  - 略称でのLDA
    - 線形判別分析
    - 局所密度近似
    - その他、社名や自動運転技術等にも

  →検索やソートの際の障害になりうる
- 単語の周囲との関係性等からその単語の意味の推定ができないか？
  →トピックモデル

## 目的

- トピック解析の手法の一つであるLDAを用いたプログラムを作成しトピックモデルによる文章の分類分けを行い、プログラムの書き方と文書の解析方法を学習する

## トピックモデル

- **トピックモデル**
  文書や単語ごとの潜在的なトピックを推論するモデル
- **トピック**
  トピックとは主題や題目という意味で会話や文書の中心となっているテーマのようなもの
  例
  - 松井秀樹選手がホームラン通算500本達成
    →この文のトピックは野球と推測
  - 増税の影響で消費が大きく落ち込んだ
    →この文のトピックは経済
  - お笑い芸人藤本とタレント木下優樹菜が離婚
    →この文のトピックは芸能界

  上記の例に「野球」「経済」「芸能界」の単語は出てこないがどんなトピックの分類は可能

- **言語処理以外でのトピックモデル**
  - 対象を置き換えた他分野への適用例
    - 文書を画像全体、単語を画像の一部に置き換え、画像全体のトピックを判別
    - 文章を楽譜、単語を一小節分の楽譜に置き換え、楽曲全体のトピックを判別

## 参考文献

[1] David M Blei, Andrew Y Ng, Michael I Jordan : “Latent Dirichlet Allocation”, in Journal of Machine Learning Research”
http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf
(last visit Jan. 9th, 2020)
[2] David Cournapeau:”scikit-learn”
https://scikit-learn.org/stable/user_guide/ (last visit Jan. 9th,2020)

## LDA(Latent Dirichlet Allocation)

- **LDA**
  - 潜在ディレクレ配分法とも呼ばれるトピック分析の手法の一つ
  - 判別したい文書内の単語のトピックを推定し、トピックの割合から文書全体のトピックを判断をする
- **LDAで用いる分布とパラメータ**
  - LDAでは以下の三つの分布とトピック選択関わるパラメータα、単語の出現に関するパラメータβを用いて計算
    - トピック分布θ：文書dについてトピックkの出現確率
      $\theta_{d,k} = \{\theta_{d,1}, \theta_{d,2}, \theta_{d,3}, \dots\}$
    - 潜在トピックz：文書dのi番目の単語$w_{d,i}$のトピックを仮定した分布
      $\{W_{d,1}, W_{d,2}, W_{d,3}, W_{d,4} \cdots\}$
      $\{Z_{d,1},\ Z_{d,2},\ Z_{d,3},\ Z_{d,4} \cdots\}$
      $\theta_d = \{\theta_{d,1}, \theta_{d,2}, \theta_{d,3}, \theta_{d,4} \cdots\}$
    - 単語の出現分布ϕ：各トピックkにおいて単語wの出現確率
      $\Phi_{w,k} = \{\Phi_{w,1}, \Phi_{w,2}, \Phi_{w,3}, \cdots\}$
    - トピックの選択確率を得る為のパラメータα
    - トピックに応じた単語の生成確率得る為のパラメータβ

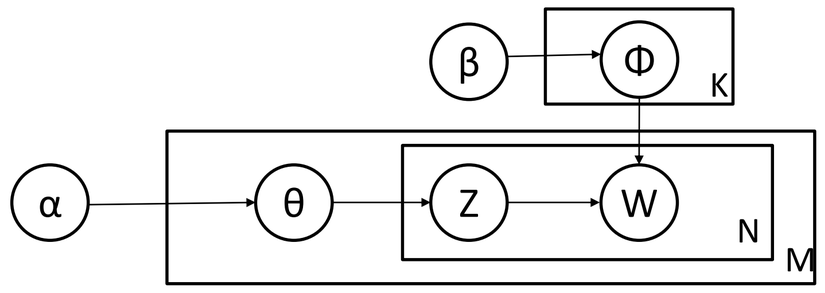


図 トピック数K,文書数M,単語数N としたLDAのグラフィカルモデル

- **LDAの推論と学習**
  - 推論は以下の単語に対するトピックとその確率分布$p(\theta,z|w,\alpha,\beta)$を求める事
    $$p(\theta,z|w,\alpha,\beta)=\frac{p(\theta,z,w|\alpha,\beta)}{p(w|\alpha,\beta)}$$
  - LDAの学習は既存のコーパスに対して尤度を最大化させる、つまり以下の式ををを最大化させるα,βを求める事
    $$l(\alpha,\beta) = \sum_{d=1}^{M} \log p(w_d|\alpha,\beta)$$

## 実行結果と課題

- **実験**
  - 教師データとしてApple製品について、Windows製品について、野球について、宇宙についての四つの記事を学習
  - 各トピックごとに上位20個の単語を表示
  - Appleについて、野球について、宇宙についての3つの文書のトピックの判別
- **成功条件**
  - 単語と対応した文章の正しい分類分けを確認する為以下の3つを条件とする
    - 各トピック上位10個の単語内にそれぞれの記事に関する単語(以下重要単語)が一つ以上ある
    - 1トピック内に複数の記事の重要単語がない(共通な単語を除く)
    - 用意した文書すべての正しい判別
- **結果**(試行回数：30回)

| | 条件１ | 条件2 | 条件3 |
|---|---|---|---|
| 成功率 | 100% | 60% | 93% |

- **課題**
  - 教師データのトピックを強く示す重要単語が上位に来ない記事がある
    →記事ごとに重要単語の出現頻度の違いがある事が原因と考察
    →出現頻度の多い単語を弱く、低い単語を強く重み付けが必要か