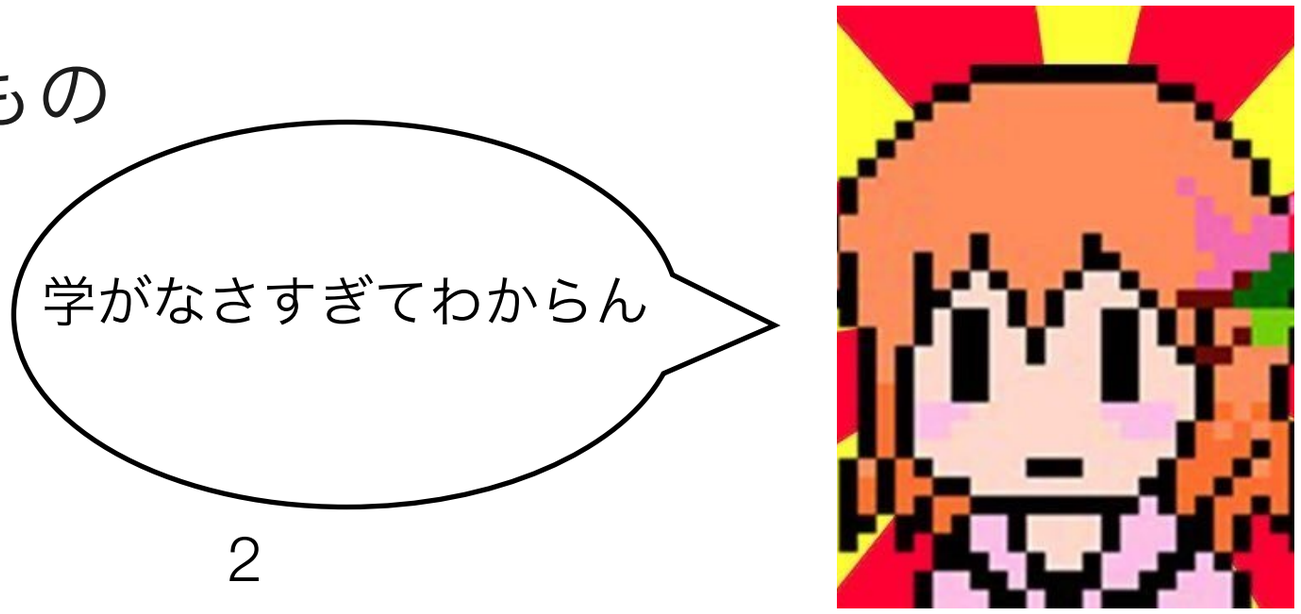
本人が一番わかっていないLDA説明スライド

神田拓実

LDAとは？

- Latent Dirichlet Allocation (潜在的ディリクレ配分法)
- トピックモデルと呼ばれる手法の１つ
- 文章解析に用いられる
- １つの文章が複数のトピックを持つものとして説明できる
- PLSIをベイズ化したもの

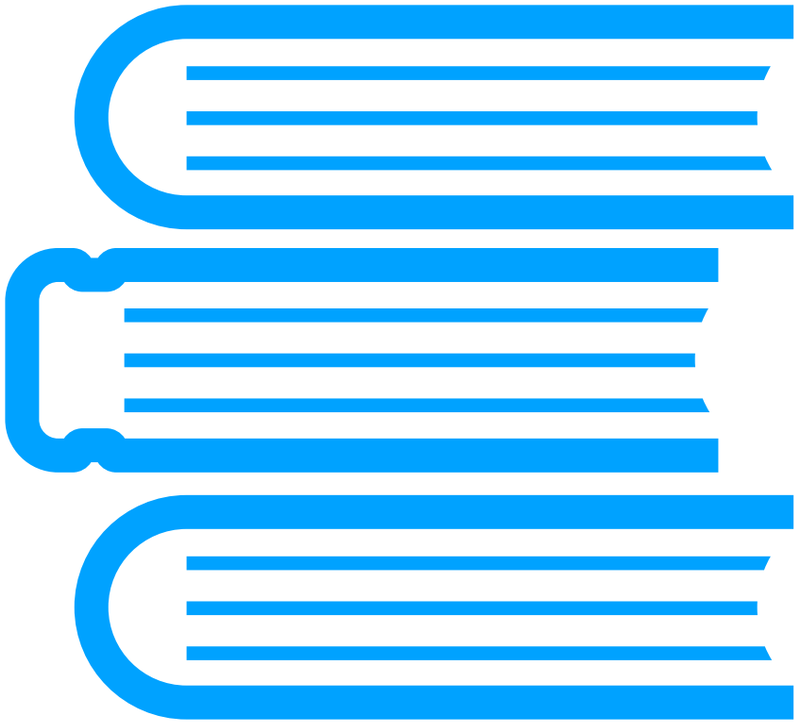


よくわかる解説

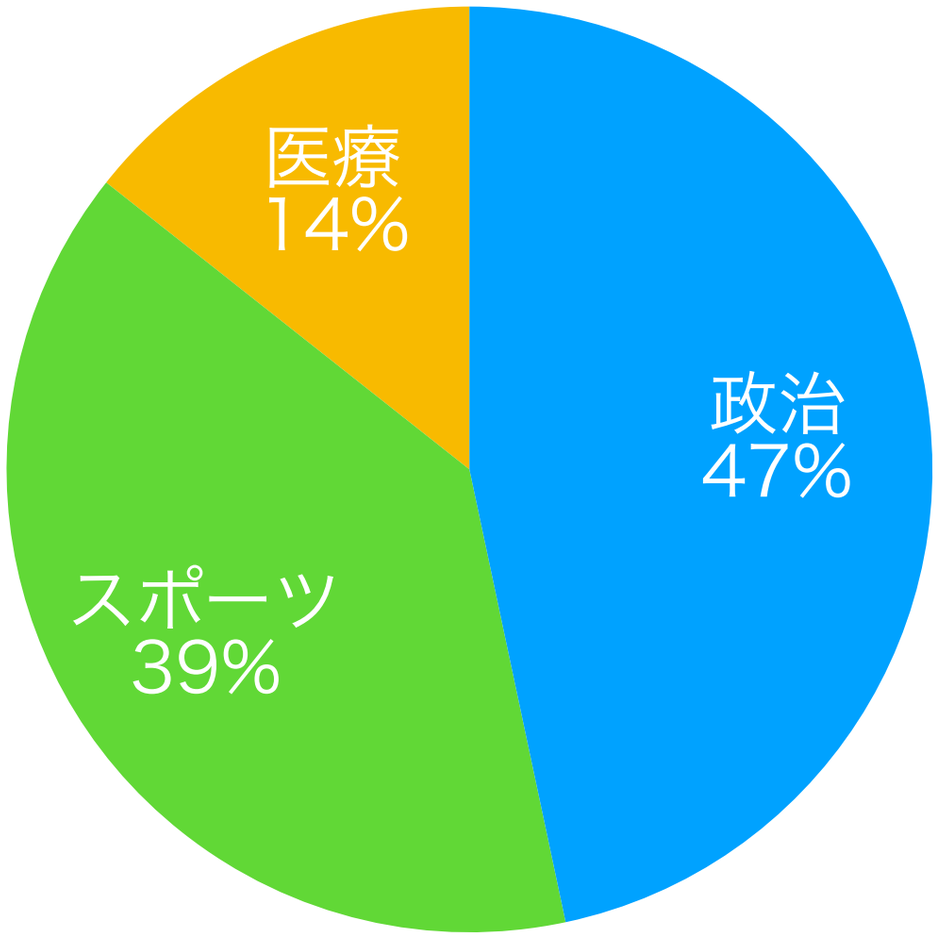
LDAは**「トピック分布」**と**「単語分布」**を用いて文章をモデル化する

トピック分布とは

- 文章が持っているトピックを割合で表したもの



文章A



単語分布とは

- トピックに関連する単語を割合の形で表したもの

政治 →

選挙	0.1%
首相	0.1%
採決	0.1%

医療 →

手術	0.5%
問診	0.1%
採血	0.1%

テキストコーパスを用いてLDAを「学習」することで、各文章のトピック分布と各トピックの単語分布を求めることができる

LDAの生成過程

1. 各文章の単語について、ランダムにトピックを割り当てる

2. 割り当てられたトピックから、文章ごとのトピック確率を計算する

3. 割り当てられたトピックから、トピックごとの単語確率を計算する

4. ２と３の積で計算される確率をもとに、各文章の単語にトピックを再び割り当てる

5. ２、３、４を収束条件まで繰り返す

LDAに関するあれこれ

- トピック分類は
  「トピック１の割合40%、トピック２の割合60%」のような形で得られるため、
  トピック1が実際にどのような話題であるかは人間が解釈する必要がある
  例：「トピック１はホームラン、盗塁などの単語が並んでいるから野球トピックだな」

- **LDAは教師なし学習である**ため、文章コーパスによっては「標準語」「関西弁」のようなトピックが出現することも考えられる