言語解析論

竹内孔一

内容

- 潜在意味解析(トピックモデルを意識して)

これらに対する初歩的な事例の理解

潜在意味解析とは

- 文書や単語に対して隠れた意味(潜在トピック)を計算で求めて，類似文書をまとめたり，取り出したりすることができる

- 参考図書 (1冊だけでは無理)
  - 言語処理のための機械学習入門 コロナ社(高村)
  - トピックモデル 講談社 (岩田)
  - トピックモデルによる統計的潜在意味解析 コロナ社(佐藤)

歴史

- LSI (latent semantic indexing/latent semantic analysis)1988
- pLSI (probabilistic LSI) 1998
- LDA (Latent Dirichlet Allocation) 2003

LSIを例に潜在意味解析

- アイデア
  - 文書と単語の共起行列を低ランク行列分解
- 利点
  - 直接，単語がでてなくても，共起関係から文書が取り出せる

文章を文書(D)と単語(W)の共起行列と考えてみよう

文書-単語行列

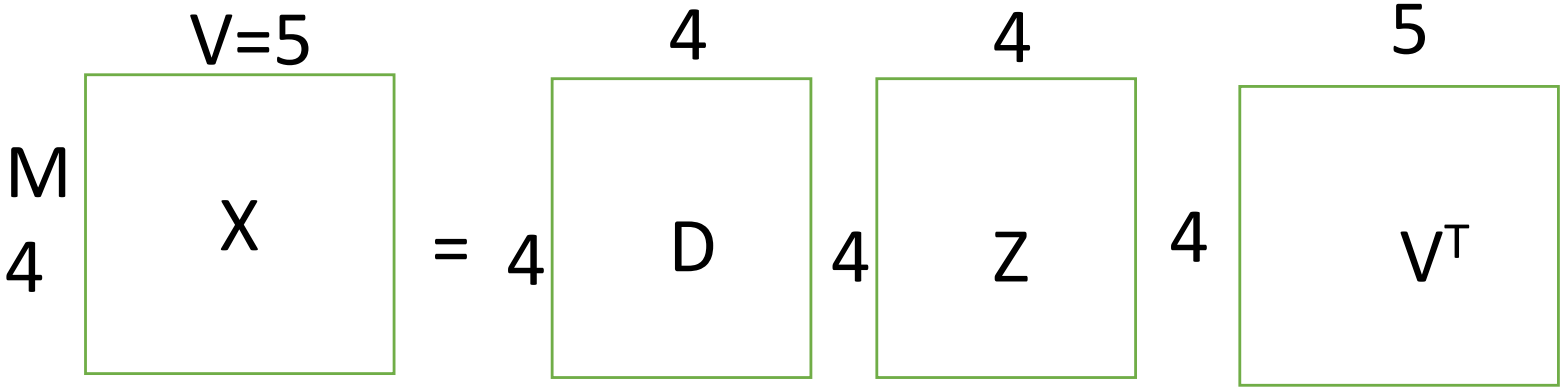
X

	ピアノ	楽譜	演奏	ゲーム	遊ぶ
D1	3	0	2	0	0
D2	0	2	3	0	0
D3	0	0	0	3	2
D4	0	0	0	2	3

「ピアノ」と「楽譜」は1つの文書で同時に出ていないので関連が見えない
➔ でも共通の単語「演奏」を通してとても関連している

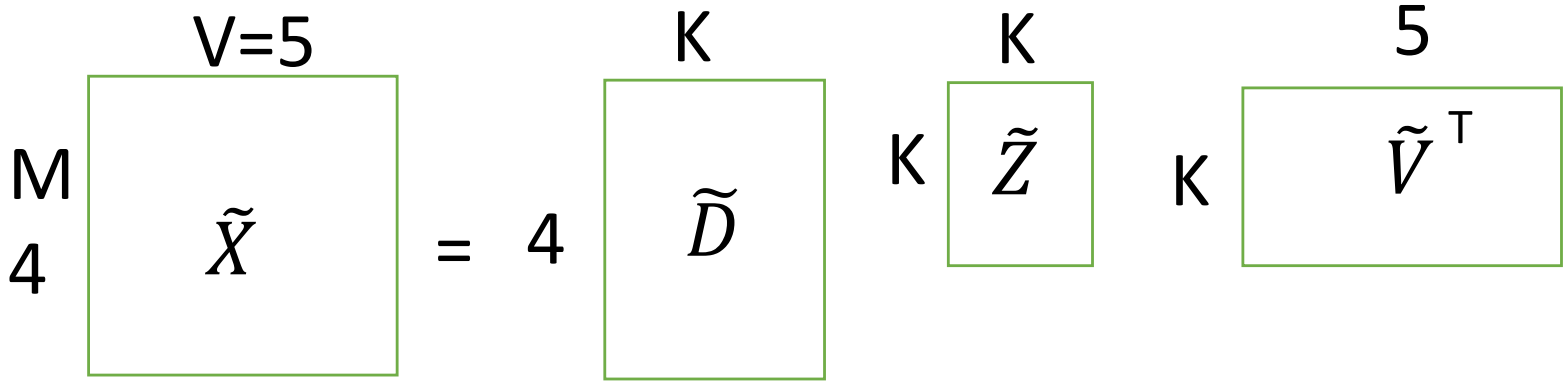
LSI

- 特異値分解し，低ランク行列に分解する
- 低ランクのK(トピック数)は人手で与える



Zを低ランクKxKにする

Zは対角行列



元の文書-単語行列Xも変わってしまう． Kがトピック数

LSI

- 低ランク行列分解後のXはトピックを考慮
- 元々相関の無かった「ピアノ」と「楽譜」が相関を持つようになる

$\tilde{D}$

	Topic1	Topic2
D1	0	値
D2	0	値
D3	値	0
D4	値	0

$\tilde{V}^{T}$　　ここはPython のgensimの計算結果

	ピアノ	楽譜	演奏	ゲーム	遊ぶ
topic1	0	0	0	0.763	0.646
topic2	0.487	0.324	0.811	0	0

(注)下記の値はイメージで正確ではない

$\tilde{X}$

文書-単語行列

	ピアノ	楽譜	演奏	ゲーム	遊ぶ
D1	2.1	0.9	1.5	0	0
D2	1.1	0.8	1.9	0	0
D3	0	0	0	2.5	1.4
D4	0	0	0	2.1	1.2

Topic1 が ゲームに関する単語分布
Topic2 が ピアノに関する単語分布であることがわかる

LSIを使う

- 未知の文書a が，既存の文書 D1, D2, D3, D4のどれに近いか，cosine類似度で求める

$$Xa = \widetilde{Da}\tilde{Z}\widetilde{V^T}$$
$$\widetilde{Da} = Xa\ \tilde{V}\tilde{Z}^{-1}$$

新しい文書aは 文書-単語行列の最後に1行加えたとする
そのときのDa (つまりtopic分解ベクトル)を求める

D1からD4についてのトピックは$\widetilde{D}$で求まっている
各$\widetilde{D1}$, $\widetilde{D2}$などと $\widetilde{Da}$とのcosineを求める

練習

- 「ピアノ 演奏 楽しい」という3単語の文書があったとする．下記の文書-単語行列の最後の行にXaとして書き加えよ
  - この時，「楽しい」は辞書に無いので無視する

	ピアノ	楽譜	演奏	ゲーム	遊ぶ
D1	3	0	2	0	0
D2	0	2	3	0	0
D3	0	0	0	3	2
D4	0	0	0	2	3
Xa					

練習

- 先のXaベクトルに対してtopic分解した新たなベクトル $\widetilde{Da}$ を求めよ

$$\widetilde{Da} = Xa \quad \tilde{V}\tilde{Z}$$

Xa

	ピアノ	楽譜	演奏	ゲーム	遊ぶ
Xa					

文書aの単語の頻度を入力

$\tilde{V}$

	topic1	topic2
ピアノ	0	0.4
楽譜	0	0.3
演奏	0	0.8
ゲーム	0.7	0
遊ぶ	0.6	0

$\tilde{Z}$

1.0	**0**
0	1.0

注意: 数値は簡単化している

答え Da = [0, 1.2]

文書aはtopic2に近い内容