

機械学習でEmbeddingは当たり前

単語、テキスト、グラフ、etc.

Doc2Vec → レビュー文の分散表現 → 感情分析, クラスタリング, 類似度比較

文書をベクトル空間に埋め込む Embedding

自然言語、画像、音声などの広い分野で既存のEmbedding手法でのベクトル化が一般的

分散表現の各次元は意味を持っていない！！

人間が解釈できない

Doc2Vec → 文書ベクトル

この文書の悲しい度合を表しているのはどこ？

ベクトルを見ただけで何もわからない

より意味的に踏み込んだベクトル演算ができない

文書ベクトル + 悲しい = ?

この文書よりも悲しさが強い文書が求められる？

ベクトルの各次元が意味を持って独立していない

ベクトルの各次元が重複のない意味を持つようにするには？

分散表現の... ベクトルの各次元に意味を持たせることで 映画を気になる観点で比較可能に!!

ちよい足し検索 類似度比較

「Disentangled」なエンコーダを作成！

- 現実空間の距離関係が反映
- ベクトルの各次元が独立
- ベクトルの各次元が意味を持つ

映画id ID:388069
 レビュー: 3.0 迫力満点! 想像よりも迫力が とてもよかったです!

従来のDoc2Vecによる映画ベクトル: 0.78, 0.26, 0.76, 0.34, 0.62, 0.11 (面白さ, ユーモアさ, 悲しさ)

Disentangledな映画ベクトル: 0.18, 0.42, 0.76, 0.01, 0.12, 0.04 (壮大さ, 悲しさ, ユーモアさ)

各次元が映画の特徴を重複なく説明!

工夫① ガイドタスク学習

Doc2Vecの周辺単語予測に加えて **メタデータ予測** も同時に予測

仮定: メタデータのような明確な情報はベクトルの一部の次元に固まる

「Disentangled」なエンコーダ

文書の内容 + 文書間の意味的関係性を学習!

工夫② β-factorをロスに適用

β-VAEを参考に... ロス計算で **MSND** との距離を考慮 (多変量標準正規分布)

映画ベクトルの分布

MSND

各次元を各変量と仮定 二つの分布の近さをKLD項で算出

各変量が独立

ベクトルの各次元の独立化を促す!

トイ・データを使用した評価 工夫②の効果検証

比較手法: ランダム, Doc2Vec, 工夫②のみ

知見① ベクトルの解釈可能性が向上

ベクトルの各次元の値の分散に注目

各次元が独立したベクトルは各次元の解釈可能性が高い

ベクトルの各次元の値の情報エントロピーに注目

各次元が独立したベクトルは各次元の情報量が制限

ベクトルの各次元の値上位文書のトピックに注目

各次元がそれぞれ特定のトピックについての情報を持つように

実データを使用した評価 「Disentangled」なベクトルが検証

比較手法: 提案手法, 工夫①のみ, 工夫②のみ, Doc2Vec

知見② ベクトルが現実空間の距離関係を反映

COS類似度算出

クエリ映画に似た映画ランキングを主観評価

どの手法でも類似映画が上位に!

知見③ ベクトルの各次元が独立

一つの次元を「ベクトル」として抽出

全次元の「ベクトル」のコーサイン類似度を総当たりで算出

映画ベクトルの各次元が直交する組み合わせの数を評価

提案手法のほとんどの次元同士が独立

知見④ ベクトルの各次元が人間の解釈できない意味を持つ

20~100個の次元を抽出

2~10位の映画をみて1位、最下位それぞれの共通度合を点数付け

自動評価

2~10位の映画ベクトルの平均と1位、最下位とのコーサイン類似度の差を算出

映画ベクトルの各次元上位映画に共通点があるかを被験者・自動評価

各次元の意味的なまとまりが被験者の回答と一致せず