PGR 210 - Natural Language Processing Part

Kristiania University College

By Huamin Ren

Huamin.ren@kristiania.no

Høyskolen Kristiania

Outline of week 43

- Text classification
- Topic models
  - LSA
  - SVD
  - LDiA

LSA

- Latent semantic analysis is based on the oldest and most commonly used technique for dimension reduction, singular value decomposition.
- A similar but simpler algorithm: Linear discriminant analysis (LDA)
- A similar but can break down documents into many topics: Latent Dirichlet allocation (LDiA)

Høyskolen Kristiania

Take one step further on LDA

- Re-think on the process

> - Compute the average position (centroid) of all the TF-IDF vectors within the class (such as spam SMS messages).
> - Compute the average position (centroid) of all the TF-IDF vectors not in the class (such as nonspam SMS messages).
> - Compute the vector difference between the centroids (the line that connects them).

Why it works? When it fails to work? How to evaluate? How to show performance?

Høyskolen Kristiania

- Framework: tf-idf + LDA
- Enlarge the vocabulary

LSA

- The algebra behind LSA called singular value decomposition.
- LSA uses SVD to find the combinations of words that are responsible, together, for the biggest variation in the data.
  - ✓Line up the axes (dimensions) in your new vectors with the greatest “spread” or variance in the word frequencies
- New basis vectors are generated; each dimension (axes) becomes a combination of word frequencies rather than a single word frequency.

SVD is an algorithm for decomposing any matrix into three “factors,” three matrices that can be multiplied together to recreate the original matrix.

Høyskolen Kristiania

- SVD was in widespread use long before the term “machine learning” even existed
- SVD decomposes a matrix into three square matrices, one of which is diagonal.

Application of SVD

- Matrix inversion

A matrix can be inverted by ***decomposing*** it into three simpler ***square matrices***, ***transposing matrices***, and then multiplying them back together.

Using SVD, LSA can break down your TF-IDF term-document matrix into three simpler matrices. And they can be multiplied back together to produce the original matrix, without any changes.

Even truncate!

- Truncate those matrices
  - Ignore some rows and columns before multiplying them back together, which reduces the number of dimensions in vector space model

The heart of LSA: SVD

$$W_{mxn} \Rightarrow U_{mxp}\, S_{pxp}\, V_{pxn}{}^{T}$$

m: number of terms in vocabulary
n: number of documents
p: number of topics (same as the number of words)

Truncating the topics

- Hint:

```
from sklearn.decomposition import TruncatedSVD

svd = TruncatedSVD(n_components=16, n_iter=100)
print(tfidf_docs.shape)
svd_topic_vectors = svd.fit_transform(tfidf_docs)
```

Implement LSA using the same dataset (sms-spam.csv)