Data Mining: Stage 2 Report

Yimeng Lu• 10.10.2017

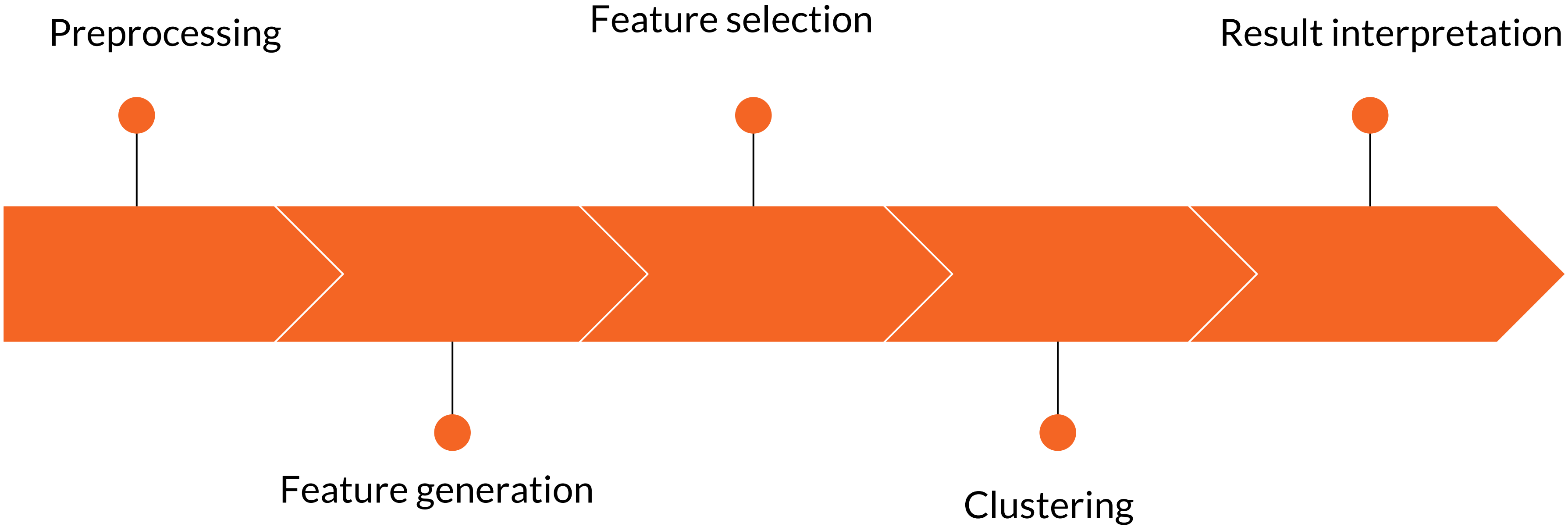
Clustering

Basic ideas

Use title, index and abstract to generate, select and cluster features

Tool boxes

Sklearn, pattern



Preprocessing

Feature generation

Feature selection

Clustering

Result interpretation

Preprocessing

Prepare for BoW

- Remove punctuations
- Tokenization
- Remove stopwords
- Lemmatize

ngrams

- Remove punctuations
- Tokenization
- Find “size-2” and “size-3” phrases

Feature Generation

Vector Generation

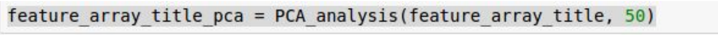
- CountVectorizer
- Different size of max_features for title, index and abstracts

Dimensionality Reduction

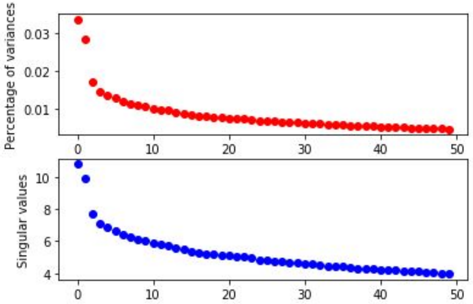
- PCA
- SVD
- Kernel PCA
- LLE

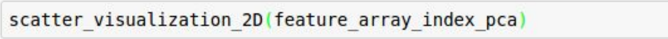
Feature Selection-1

Visualization

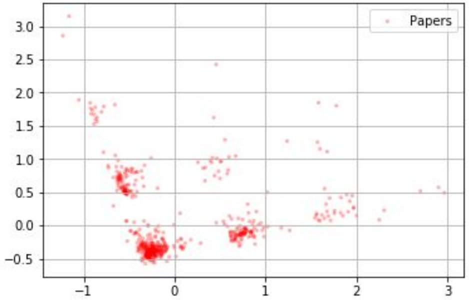


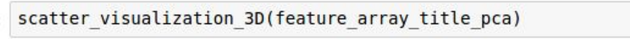
```
feature_array_title_pca = PCA_analysis(feature_array_title, 50)
```



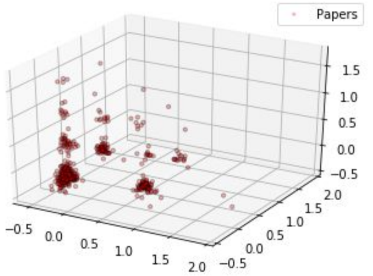


```
scatter_visualization_2D(feature_array_index_pca)
```





```
scatter_visualization_3D(feature_array_title_pca)
```



Feature Selection-2

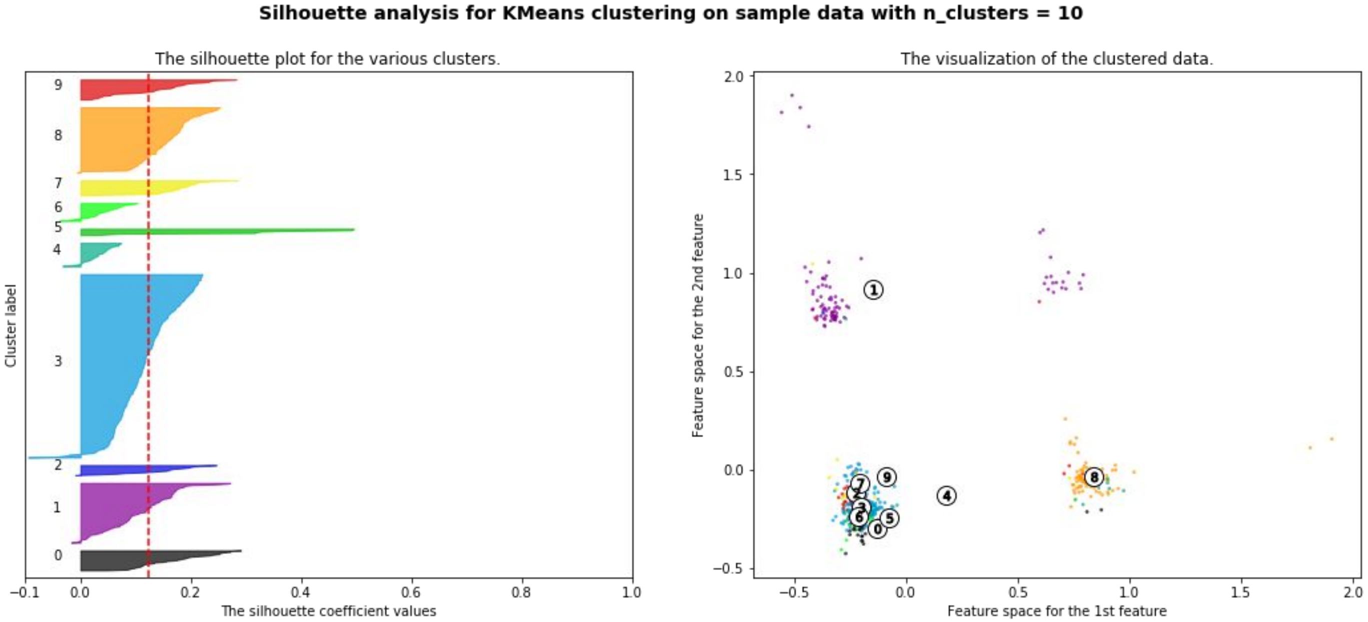
Decision

- 10 title features from PCA
- 20 abstract features from PCA

Observations

- titles are more representative than the words in abstracts

Clustering: k-means



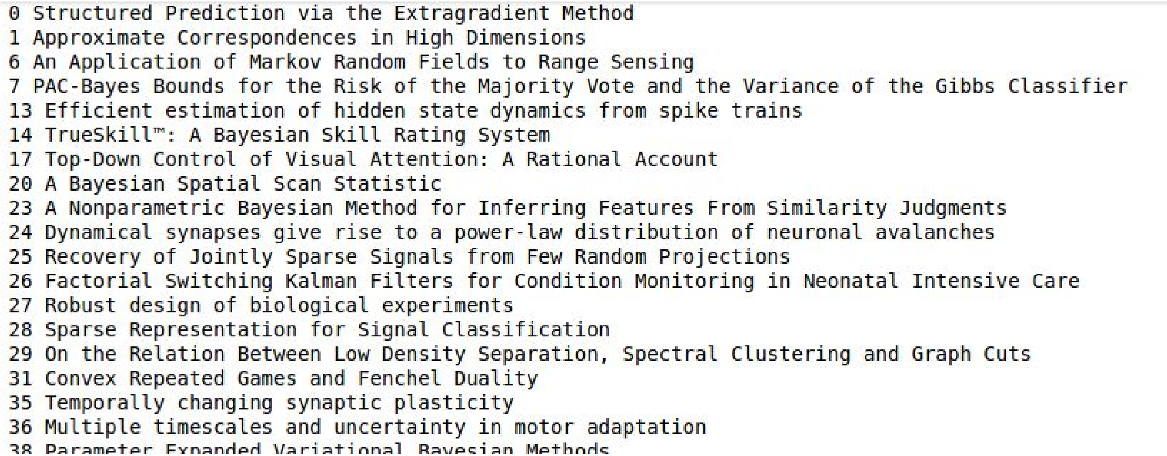
Silhouette analysis for KMeans clustering on sample data with n_clusters = 10

The silhouette plot for the various clusters.

The visualization of the clustered data.

Results

- 10 clusters of papers



```
0 Structured Prediction via the Extragradient Method
1 Approximate Correspondences in High Dimensions
6 An Application of Markov Random Fields to Range Sensing
7 PAC-Bayes Bounds for the Risk of the Majority Vote and the Variance of the Gibbs Classifier
13 Efficient estimation of hidden state dynamics from spike trains
14 TrueSkill™: A Bayesian Skill Rating System
17 Top-Down Control of Visual Attention: A Rational Account
20 A Bayesian Spatial Scan Statistic
23 A Nonparametric Bayesian Method for Inferring Features From Similarity Judgments
24 Dynamical synapses give rise to a power-law distribution of neuronal avalanches
25 Recovery of Jointly Sparse Signals from Few Random Projections
26 Factorial Switching Kalman Filters for Condition Monitoring in Neonatal Intensive Care
27 Robust design of biological experiments
28 Sparse Representation for Signal Classification
29 On the Relation Between Low Density Separation, Spectral Clustering and Graph Cuts
31 Convex Repeated Games and Fenchel Duality
35 Temporally changing synaptic plasticity
36 Multiple timescales and uncertainty in motor adaptation
38 Parameter Expanded Variational Bayesian Methods
```

Thank you