**ML Notes**

**Preprocessing:**

String->numeric: LabelEncoder

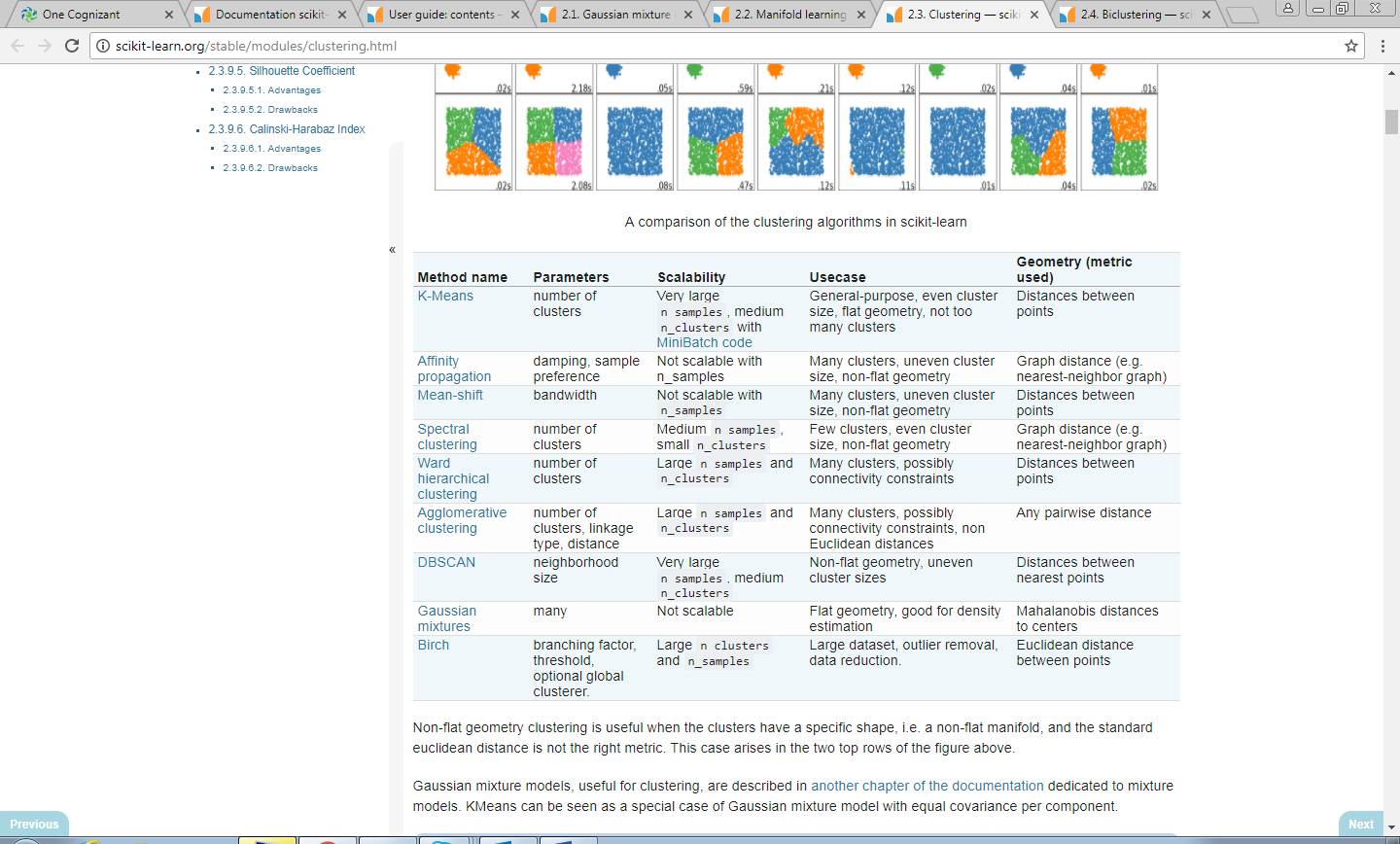
Dictionary: DictVectorizer

tokenization and occurrence counting: CountVectorizer

**Algorithms:**

**Unsupervised Learning:**

1. Gaussian mixture models: fastest for learning mixture models (BayesianGaussianMixture: regularized and less sensitive to number of parameters)
2. Manifold learning: non-linear dimensionality reduction
3. **Clustering:**



1. Biclustering: simultaneously cluster rows and columns of a data matrix.
2. Nearest Neighbors: uniform interface to three different nearest neighbors algorithms: BallTree, KDTree, and a brute-force algorithm

**Supervised Learning:**

1. Linear and Quadratic Discriminant Analysis: These classifiers are attractive because they have closed-form solutions that can be easily computed, are inherently multiclass, have proven to work well in practice and have no hyper parameters to tune.
2. Support Vector Machines (SVM): set of supervised learning methods used for classification, regression and outliers detection.
   1. SVC, NuSVC and LinearSVC are classes capable of performing multi-class classification on a dataset.
   2. implementations of Support Vector Regression: SVR, NuSVR and LinearSVR.
   3. One-class SVM is used for novelty detection, that is, given a set of samples, it will detect the soft boundary of that set so as to classify new points as belonging to that set or not. The class that implements this is called OneClassSVM.
3. Nearest neighbors: Classification is computed from a simple majority vote of the nearest neighbors of each point
   1. KNeighborsClassifier implements learning based on the k nearest neighbors of each query point, where k is an integer value specified by the user.
   2. RadiusNeighborsClassifier implements learning based on the number of neighbors within a fixed radius r of each training point, where r is a floating-point value specified by the user.
   3. Neighbors-based regression can be used in cases where the data labels are continuous rather than discrete variables.
      1. KNeighborsRegressor implements learning based on the k nearest neighbors of each query point, where k is an integer value specified by the user.
      2. RadiusNeighborsRegressor implements learning based on the neighbors within a fixed radius r of the query point, where r is a floating-point value specified by the user.
4. Gaussian Processes (GP): are a generic supervised learning method designed to solve regression and probabilistic classification problems.
   1. The GaussianProcessRegressor implements Gaussian processes (GP) for regression purposes.
   2. The GaussianProcessClassifier implements Gaussian processes (GP) for classification purposes, more specifically for probabilistic classification, where test predictions take the form of class probabilities.
5. Cross decomposition: to find linear relations between two multivariate datasets: the X and Y arguments of the fit method are 2D arrays.
   1. Partial Least Squares (PLS)
   2. Canonical Correlation Analysis (CCA)
6. Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes’ theorem with the “naive” assumption of independence between every pair of features.
   1. GaussianNB implements the Gaussian Naive Bayes algorithm for classification.
   2. MultinomialNB implements the naive Bayes algorithm for multinomially distributed data, and is one of the two classic naive Bayes variants used in text classification
   3. BernoulliNB implements the naive Bayes training and classification algorithms for data that is distributed according to multivariate Bernoulli distributions.
7. Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.
   1. DecisionTreeClassifier
   2. DecisionTreeRegressor
8. Ensemble methods: to combine the predictions of several base estimators built with a given learning algorithm in order to improve generalizability / robustness over a single estimator.
   1. In averaging methods, the driving principle is to build several estimators independently and then to average their predictions.
      1. BaggingClassifier
      2. RandomForestAlgorithm
   2. By contrast, in boosting methods, base estimators are built sequentially and one tries to reduce the bias of the combined estimator.
      1. AdaBoostClassifier
      2. GradientBoostingClassifier
      3. GradientBoostingRegressor
9. <http://scikit-learn.org/stable/modules/multiclass.html>
10. Feature Selection: used for feature selection/dimensionality reduction on sample sets, either to improve estimators’ accuracy scores or to boost their performance on very high-dimensional datasets.
11. Semi-supervised Learning: Semi-supervised learning is a situation in which in your training data some of the samples are not labeled.
    1. Label propogation
    2. Label spreading