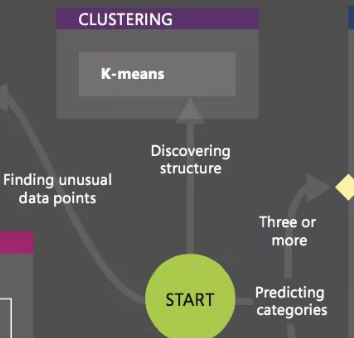


ALGORITHM CLUSTER

**In this session**

* Cluster Algorithms in Azure ML
* Model overview
* Dataset
* Feature Hashing module
* Train
* Edit Metadata

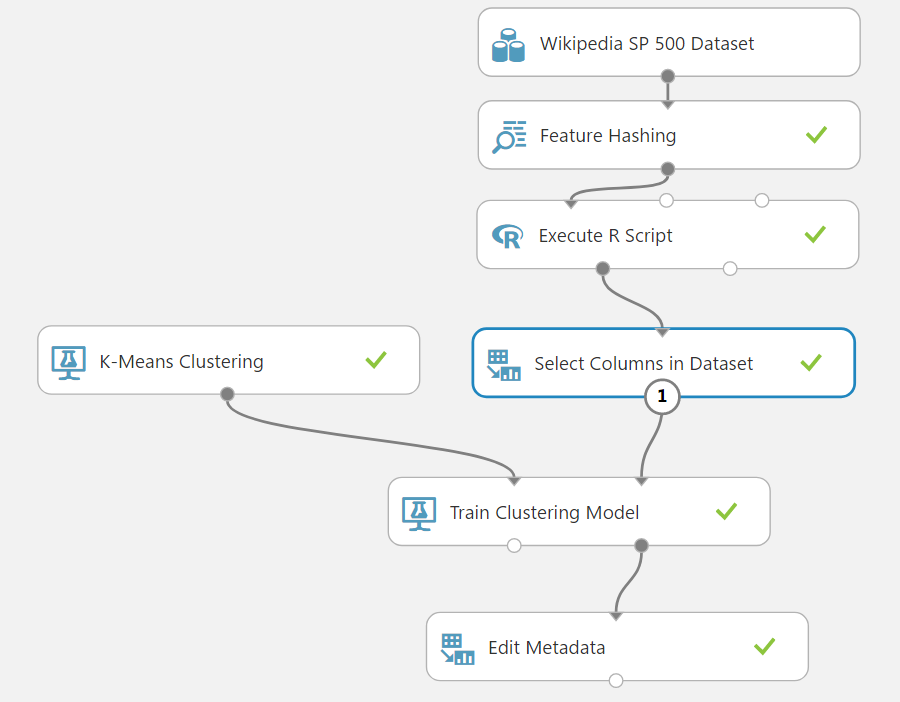
Cluster Algorithms in Azure ML



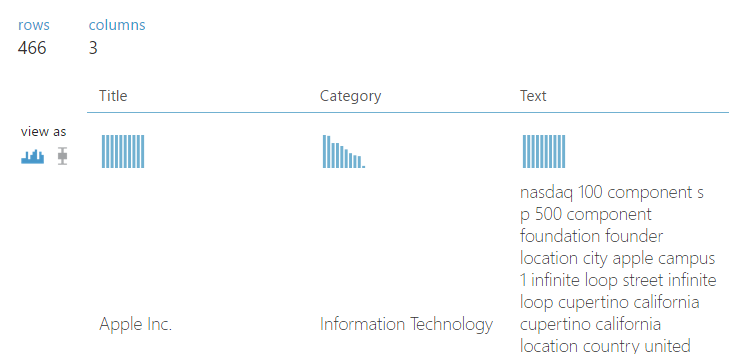
* Uses iterative techniques
* Group cases in a dataset into clusters
* Contain similar characteristics
* Useful for exploring data
* Identifying anomalies in the data
* Making predictions
* Identify relationships in a dataset
* Not logically derive by browsing or simple observation
* Used in the early phases of machine learning tasks
* Explore the data and discover unexpected correlations
* Only algorithm in AML that is Unsupervised

Model overview

* Dataset: Wikipedia SP 500 Dataset
* Feature Hashing: create feature from column Text
* R Script: reduce feature to 10 columns
* Select Columns: exclude column PC1
* K-Means: clustering algorithm
* Train: train model using only numeric column
* Edit Metadata: make category from Assignments column



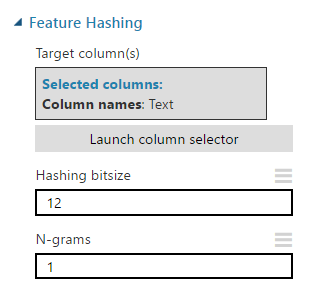
Dataset



Pre-processed outside Azure ML Studio

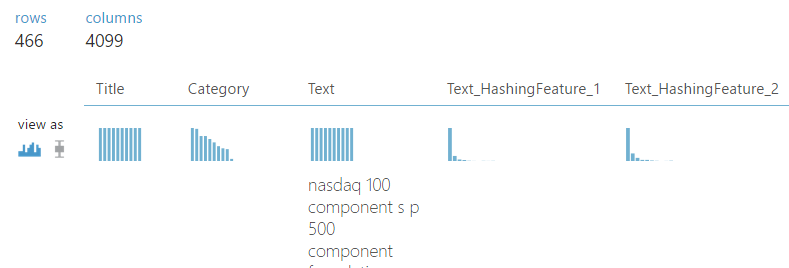
* Removing wiki formatting
* Removing non-alphanumeric characters
* Converting all text to lowercase
* Adding company categories, where known

Feature Hashing module

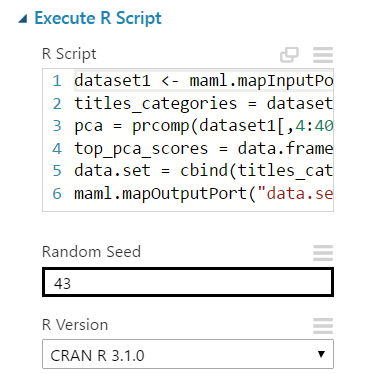
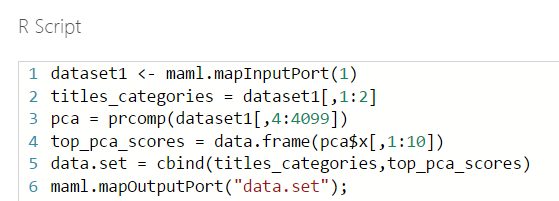


Feature Hashing module

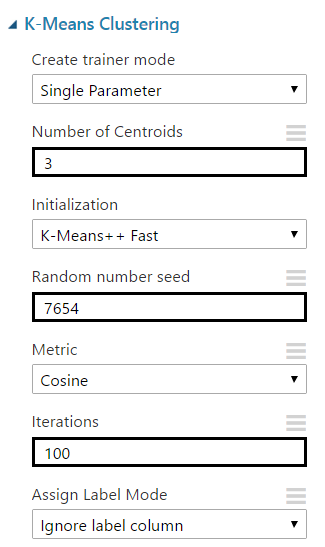
* Tokenizes the text string
* Transforms the data into a series of numbers
* Based on the hash value of each token



R Script

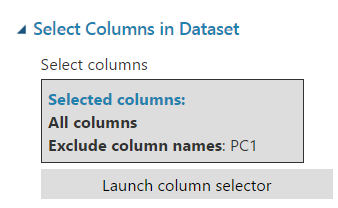


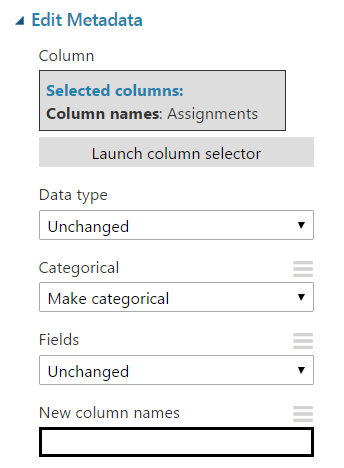
* Dimensionality of the data from hashing is too high (4K)
* Cannot be used by the K-Means clustering algorithm directly
* Principal Component Analysis (PCA) was applied using a custom R script
* Reduce the dimensionality to 10 variables
* View the result = double-clicking the right-hand output of the Execute R Script



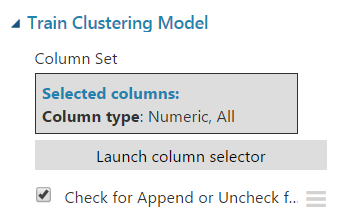
* Select Columns in Dataset
* K-Means Clustering

<https://msdn.microsoft.com/en-us/library/azure/dn905944.aspx>





* Train Clustering Model
* Edit Metadata



More information

K-Means Clustering

<https://msdn.microsoft.com/en-us/library/azure/dn905944.aspx>

This Experiment

<https://gallery.cortanaintelligence.com/Experiment/Clustering-K-Means-basic>