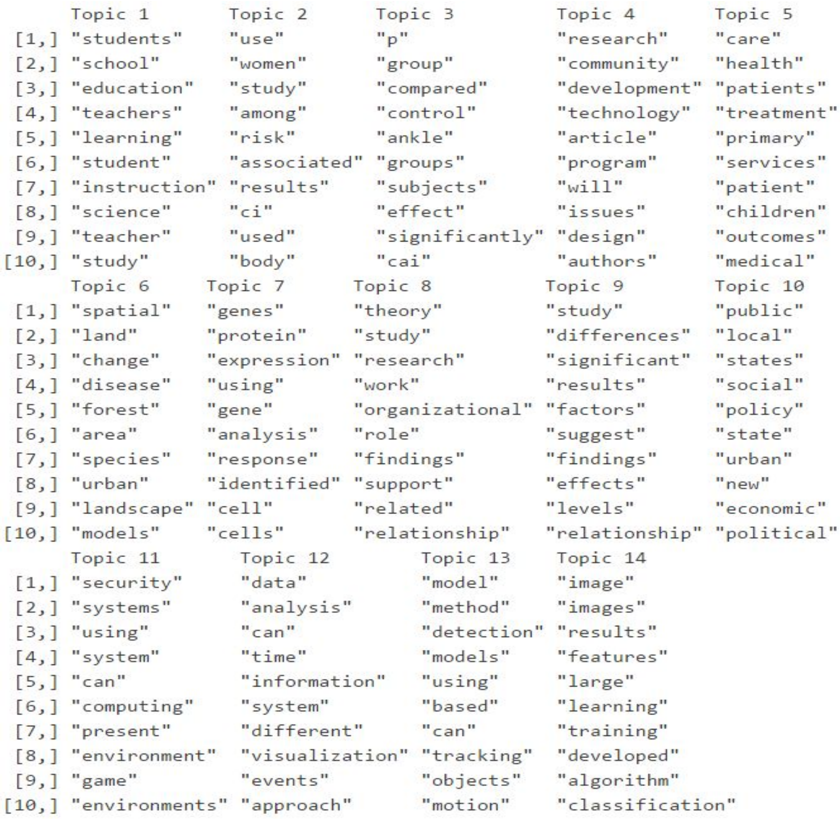
K-means is a general clustering algorithm which tries to cluster similar vectors together. Vector can be defined in any way. It can be numerical values of n dimensions or even text data. Depending on the type of vector similarity measure is defined. For numerical data, distance is generally used as similarity measure whereas for text data, cosine similarity or jaccard similarity. It is unsupervised machine learning algorithm that can be used to find outliers.
In this project, we made an attempt to do K-means clustering over streaming data. It is very easy when data is fixed or constant but it’s whole another story when we have to deal with live data streams. Though we were unsuccessful in attempting K-means but we are very close. Also, it is worth mentioning that there is no available code or similar type of analysis done with Spark streaming online.

Topic Modelling is another clustering algorithm but for text data. It identifies set of topics in different documents. Its input is set of documents and its output is word distribution in different topics and topic distribution in different documents.
This analysis will be really helpful in understanding what people are talking about in different regions of United States, How different these topics are, Many such questions. We managed to perform topic modelling on static data and below is the picture showing it.



```
      Topic 1       Topic 2      Topic 3         Topic 4       Topic 5
 [1,] "students"    "use"        "p"             "research"    "care"
 [2,] "school"      "women"      "group"         "community"   "health"
 [3,] "education"   "study"      "compared"      "development" "patients"
 [4,] "teachers"    "among"      "control"       "technology"  "treatment"
 [5,] "learning"    "risk"       "ankle"         "article"     "primary"
 [6,] "student"     "associated" "groups"        "program"     "services"
 [7,] "instruction" "results"    "subjects"      "will"        "patient"
 [8,] "science"     "ci"         "effect"        "issues"      "children"
 [9,] "teacher"     "used"       "significantly" "design"      "outcomes"
[10,] "study"       "body"       "cai"           "authors"     "medical"
      Topic 6     Topic 7      Topic 8          Topic 9        Topic 10
 [1,] "spatial"   "genes"      "theory"         "study"        "public"
 [2,] "land"      "protein"    "study"          "differences"  "local"
 [3,] "change"    "expression" "research"       "significant"  "states"
 [4,] "disease"   "using"      "work"           "results"      "social"
 [5,] "forest"    "gene"       "organizational" "factors"      "policy"
 [6,] "area"      "analysis"   "role"           "suggest"      "state"
 [7,] "species"   "response"   "findings"       "findings"     "urban"
 [8,] "urban"     "identified" "support"        "effects"      "new"
 [9,] "landscape" "cell"       "related"        "levels"       "economic"
[10,] "models"    "cells"      "relationship"   "relationship" "political"
      Topic 11       Topic 12        Topic 13    Topic 14
 [1,] "security"     "data"          "model"     "image"
 [2,] "systems"      "analysis"      "method"    "images"
 [3,] "using"        "can"           "detection" "results"
 [4,] "system"       "time"          "models"    "features"
 [5,] "can"          "information"   "using"     "large"
 [6,] "computing"    "system"        "based"     "learning"
 [7,] "present"      "different"     "can"       "training"
 [8,] "environment"  "visualization" "tracking"  "developed"
 [9,] "game"         "events"        "objects"   "algorithm"
[10,] "environments" "approach"      "motion"    "classification"
```

But we face the same problem that we had with K-means algorithm while preprocessing data. Again, there is no support for this on Spark Streaming but it is possible to do it.