Felügyelet nélküli tanulás: Topik modellezés magyar törvényszövegeken

A klaszterezés egy adathalmaz pontjainak, rekordjainak hasonlóság alapján való csoportosítása, ami szinte minden nagyméretű adathalmaz leíró modellezésére alkalmas. A klaszterezés során az adatpontokat diszjunkt halmazokba, azaz klaszterekbe soroljuk, hogy az elemeknek egy olyan partíciója jöjjön létre, amelyben a közös csoportokba kerülő elempárok lényegesen hasonlóbbak egymáshoz, mint azok a pontpárok, melyek két különböző csoportba sorolódtak. Klaszterezés során a megfelelő csoportok kialakítása nem egyértelmű feladat, mivel a különböző adatok eltérő jelentése és felhasználása miatt adathalmazonként más szempontokat kell figyelembe vennünk.

install.packages("tm")  
install.packages("quanteda")  
install.packages("topicmodels")  
install.packages("tidytext")  
install.packages("ggplot2")  
install.packages("dplyr")  
install.packages("tidyr")  
install.packages("RTextTools")   
  
library(tm)  
library(quanteda)  
library(topicmodels)  
library(tidytext)  
library(ggplot2)  
library(dplyr)  
library(tidyr)  
library(RTextTools)

Az alábbiakban 1998-2002 és a 2002-2006-os parlamenti ciklus 1032 törvényszövegének topik modellezését és a szükséges előkészítő, korpusztisztító lépéseket mutatjuk be.

## Including Plots

You can also embed plots, for example:



Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.