## Ejercicios

## 1.- Creación de corpus (tm)

1. Importa los comentarios almacenados en el archivo disponible en Moodle.
2. Qué es un corpus? Crea un corpus con los datos que importaste.
3. Modifica tu corpus para utilizar letras minúsculas.
4. Remueve signos de puntuación.
5. Remueve números.
6. Remueve URLs.
7. Qué es una Stop Word? Remueve las stop words en inglés definidas en el paquete “tm” agregando stop words definidas por el grupo y quitando otras.
8. Aplica tres búsquedas de términos en tu corpus.
9. Reemplaza 3 casos de palabras que sean sinónimos.

## 2.- Matriz de terminos

1. Que es una matriz de términos? Genera una matriz de términos.
2. Cuales son los términos mas frecuentes?
3. Cual es la frecuencia de los primeros 5 términos que empiezan con p en las primeras 100 entradas?
4. Genera una grafica de barras con aquellos términos que aparecen mas de 50 veces.
5. Identifica asociaciones para las palabras mas frecuentes, que puedes observar?
6. Genera una nube de palabras con todos tus términos.
7. Aplica clustering jerárquico a tus términos. Remueve los términos que están muy dispersos (.95). Utiliza el método Ward.
8. Divide tu dendograma en 8 clusters. Describe cada uno de ellos.
9. Aplica k-means para hacer un clustering a las entradas. Traspón la matriz de términos en una matriz de documentos. Fija una semilla.
10. Imprime los centros de los clusters.
11. Imprime ejemplos de cada cluster para validar que son tópicos diferentes.