

reporte_analisis

AFP

21 de enero de 2016

En este documento se justifican las decisiones tomadas a la hora de realizar el preprocesamiento

-Clasificación jerárquica:

Primero se debe usar el método sparse sobre el TDM de otro modo da error y mientras más cercano a 1 sea el sparse más palabras aparecen en la gráfica por ejemplo el encontrado en la carpeta gráficas

También cabe destacar que es el método ward.D2 es un método del 2014 y el único que necesita la distancia al cuadrado es el método de centroide fuente:

<https://stat.ethz.ch/R-manual/R-devel/library/stats/html/hclust.html>

-PCA: Uso el mismo valor de sparse que el anterior usamos el valor por default de ncp y luego simplemente graficamos el PCA,