Actividad: Modelado de temas (LDA):

En mi caso para el problema de modelado de temas vamos a analizar y clasificar por diferentes tópicos, las reseñas de una tienda online. Esto le permitirá a la tienda identificar de manera automática (no supervisada) la temática principal de sus reseñas.
A la hora de generar los dataset de pruebas, he realizado dos variantes, la primera con 2 archivos .txt en los cuales hay cerca de 500 reseñas y luego una lista de registros implementada directamente en el código.
La elección de los parámetros está en el propio código comentado.
De la ejecución a partir de los ficheros de texto .txt he obtenido los siguientes resultados.

===== TOP PALABRAS POR TOPIC =====
Topic1: precio(0,077), mas(0,050), esperaba|mas(0,043), esperaba(0,043), decepciono(0,040), mas|decepciono(0,040), precio|esperaba(0,040), camara(0,030)
Topic2: calidadprecio(0,042), relacion(0,042), relacion|calidadprecio(0,042), recomendado(0,042), calidadprecio|recomendado(0,039), cafetera(0,033), zapatillas(0,029), silla(0,027)
Topic3: bien(0,073), rapido(0,062), llego(0,062), embalado(0,060), correcto(0,046), bien|embalado(0,045), rapido|bien(0,045), llego|rapido(0,045)
Topic4: calidad(0,066), funciona(0,058), nota(0,044), justita|nota(0,032), justita(0,032), calidad|justita(0,032), funciona|calidad(0,032), impresora(0,028)
Topic5: compre(0,063), coincide|descripcion(0,048), coincide(0,048), descripcion(0,048), descripcion|recomiendo(0,047), recomiendo(0,047), monitor(0,030), cumple(0,027)
Topic6: buena(0,058), buen(0,037), repetiria(0,035), compra(0,033), repetiria|compra(0,033), precio|repetiria(0,032), calidad|precio(0,032), buena|calidad(0,031)
==========

Topic1 Topic2 Topic3 Topic4 Topic5 Topic6
0,4545 0,0000 0,3636 0,0000 0,1818 0,0000 Las zapatillas aprietan en el empeine; mejor pedir medio número más.
0,0909 0,2727 0,4545 0,0000 0,1818 0,0000 El respaldo de la silla es duro; no aguanto 8 horas.
0,1111 0,4444 0,0000 0,4444 0,0000 0,0000 La luz del altavoz molesta de noche; debería apagarse.
0,1429 0,0000 0,0000 0,0000 0,0000 0,8571 La carcasa del SSD se calienta; pero aguanta.
0,0000 0,0000 0,1765 0,0000 0,8235 0,0000 Compré robot aspirador y funciona perfecto; cumple lo prometido (mapea bien).
0,0000 0,0000 0,0000 0,0000 0,8182 0,1818 Compré portátil y funciona perfecto; cumple lo prometido.

==========

Topic1 Topic2 Topic3 Topic4 Topic5 Topic6

0,0000 0,0000 0,7778 0,0000 0,1111 0,1111 El paquete llegó tarde y la caja venía golpeada.

0,0000 0,0000 0,0000 1,0000 0,0000 0,0000 La calidad del producto es mala, se rompió en una semana.

0,7500 0,0000 0,0000 0,0000 0,2500 0,0000 Para el precio que tiene, esperaba mejores materiales.

0,4000 0,0000 0,0000 0,4000 0,2000 0,0000 Pedí devolución y el reembolso tardó más de una semana.

Lo primero que vemos en la salida es ver qué palabras ha elegido para cada uno de los tópicos.
En el primer tópico podemos ver como hace referencia al precio, expectativas o a la descripción del artículo, en el segundo habla sobre la relación calidad precio, en el tercero sobre la entrega del producto, en el cuarto acerca de la calidad, en el quito si el producto coincide con la descripción del mismo y en el sexto acera de si está satisfecho con la compra.

A la hora de clasificar si es cierto que a veces reparte mucho las puntuaciones de cada tópico lo cual hace difícil de interpretar el resultado, en cambio en otra asigna puntuaciones altas dejando claro de qué temática se está hablando en la reseña.