

Octubre 2014

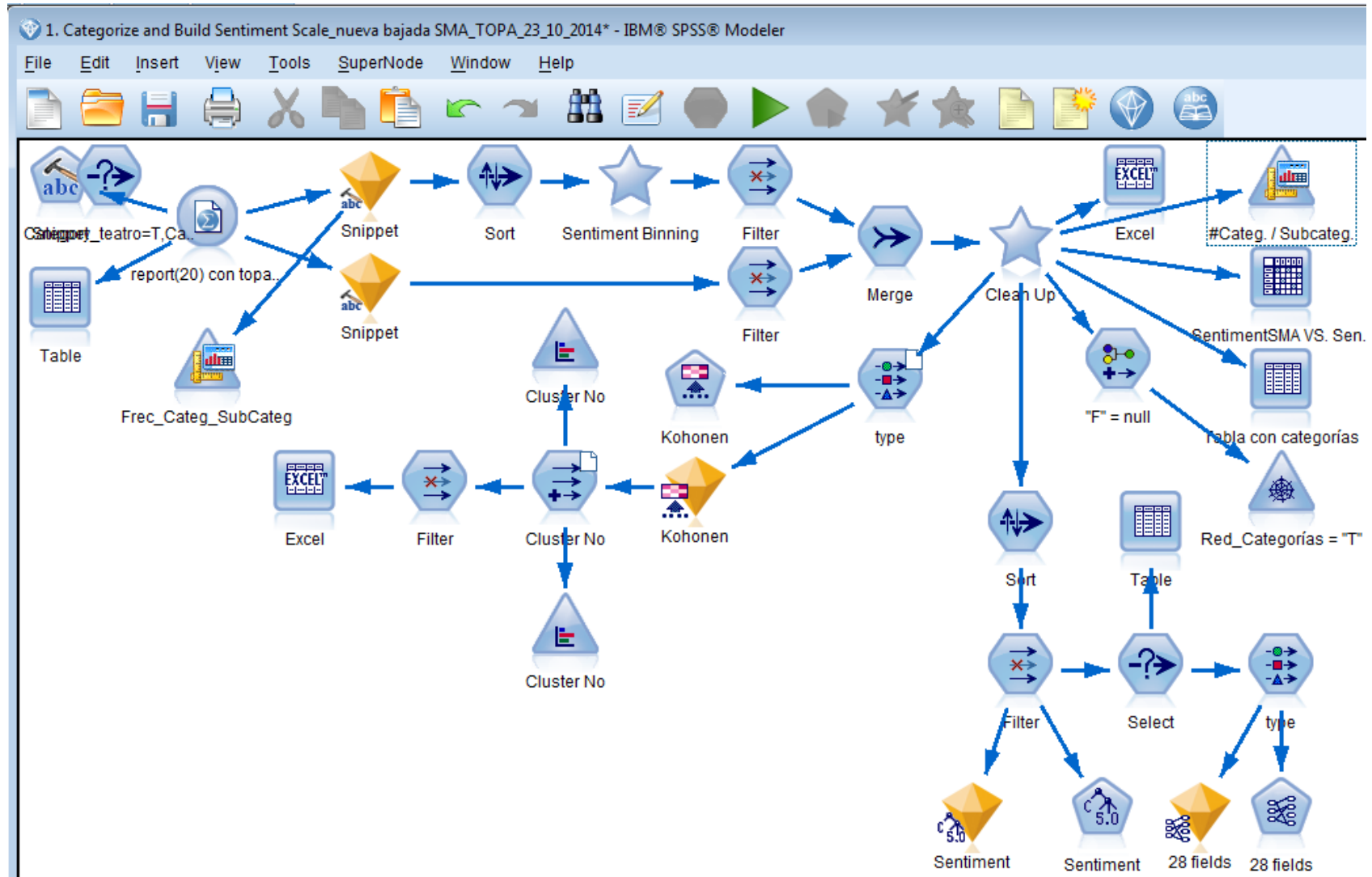
Disney - TOPA

Text Analytics: Lecturas Previas

Un abordaje estadístico

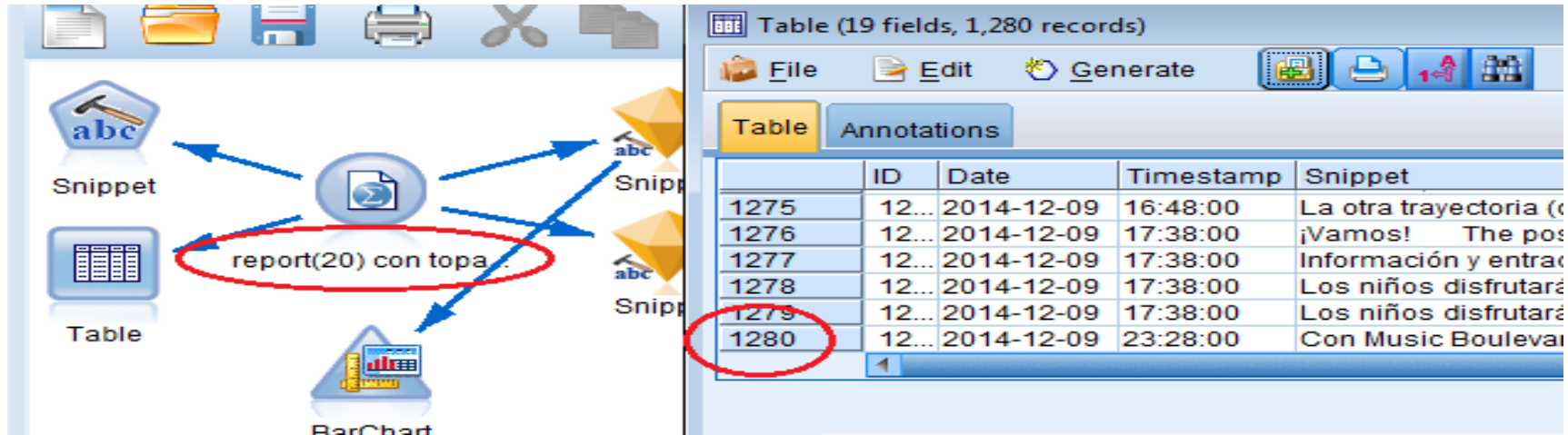


Modelo de text analytics con segmentación en SPSSSTA



BASE: Congruencia SMA vs. SPSSTA

Se tienen 1280 Snippets según filtros y contexto en SMA de la base *report(20)con topa.txt*:

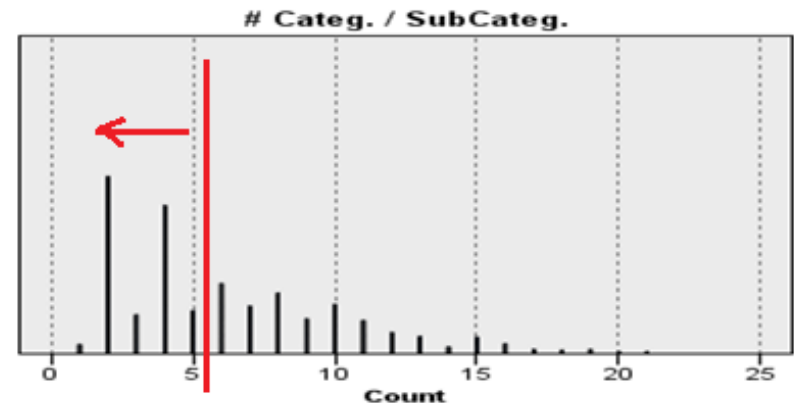


SentimentSMA	Dissatisfied	Neutral	Satisfied
	0	5	0
ambivalent	0	0	7
negative	0	21	11
neutral	4	530	208
positive	1	77	144

Cells contain: cross-tabulation of fields (including missing val...

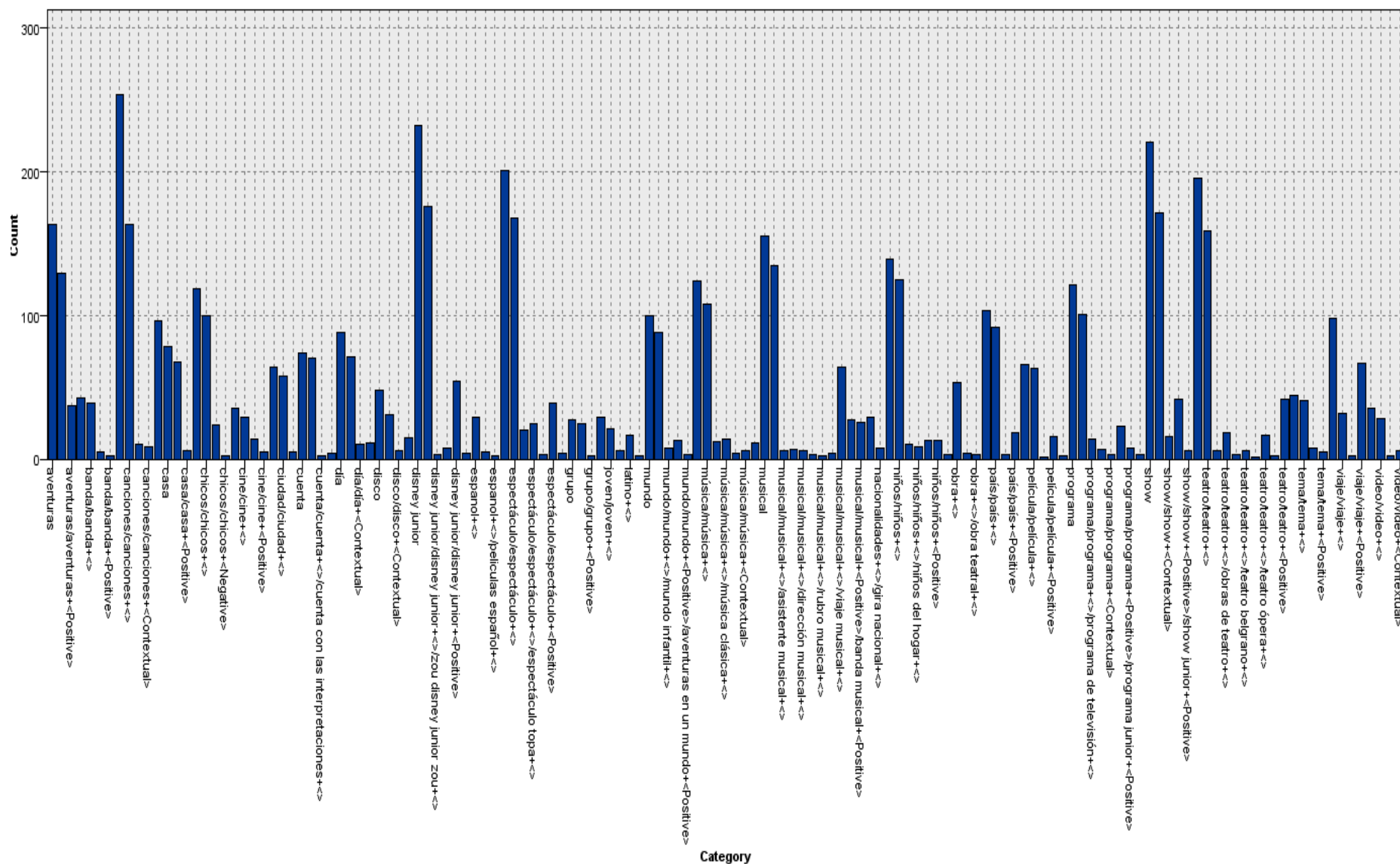
Chi-square = 115.222, df = 8, probability = 0

Se contrastaron los sentimientos original desde SMA y los resultantes (con Spanish Opinion Languages) en SPSSTA. No pareciera haber muchos casos incompatibles

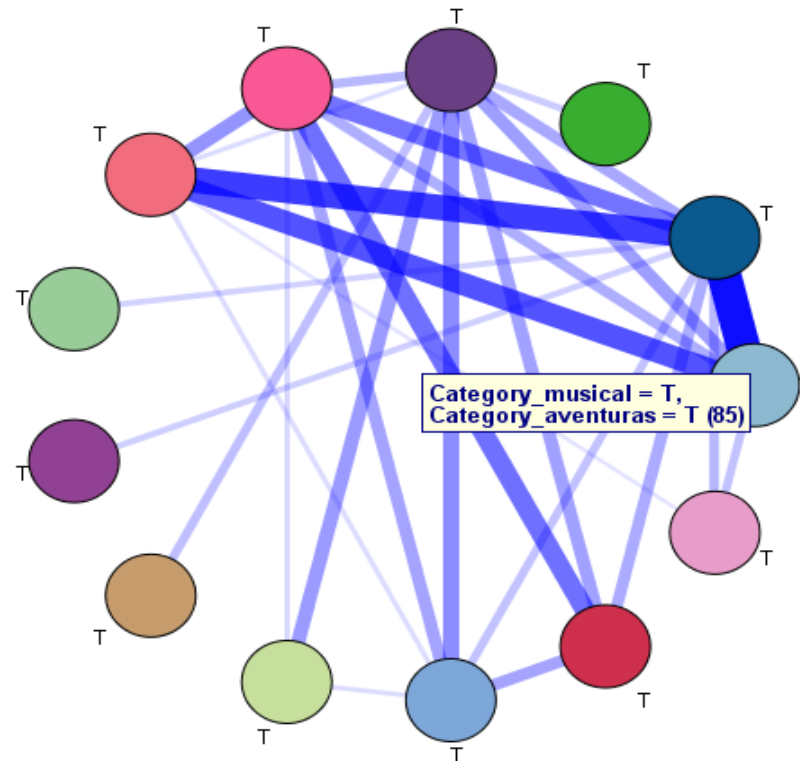


La mitad de los Snippets no produjeron más de 5 categorías simultáneamente. Se debería trabajar con la otra mitad de ellos

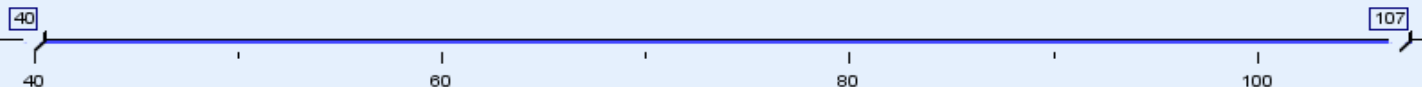
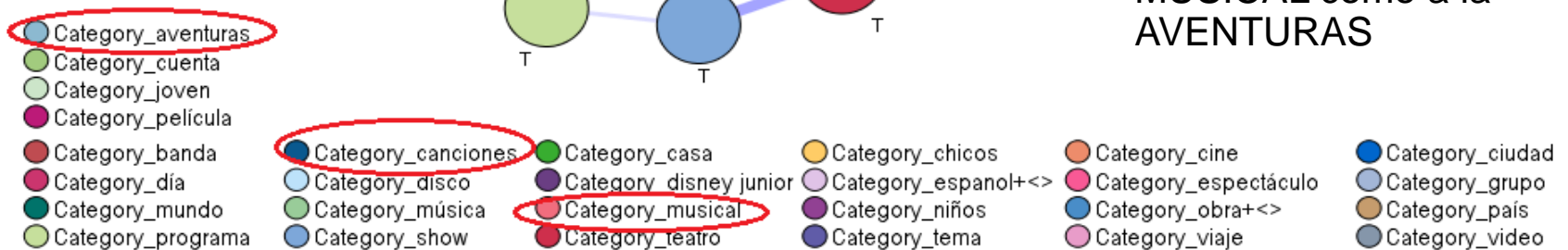
Frecuencias de todas las Categorías_Subcategorías con sentimientos



Red de categorías (sin sentimiento) asociadas



Las asociaciones de categorías más frecuentes se muestran conectadas con líneas más gruesas cuanto más Snippets las presentan. Por ejemplo existen 85 Snippets que aplicaron simultaneamente a las categorías MUSICAL como a la AVENTURAS



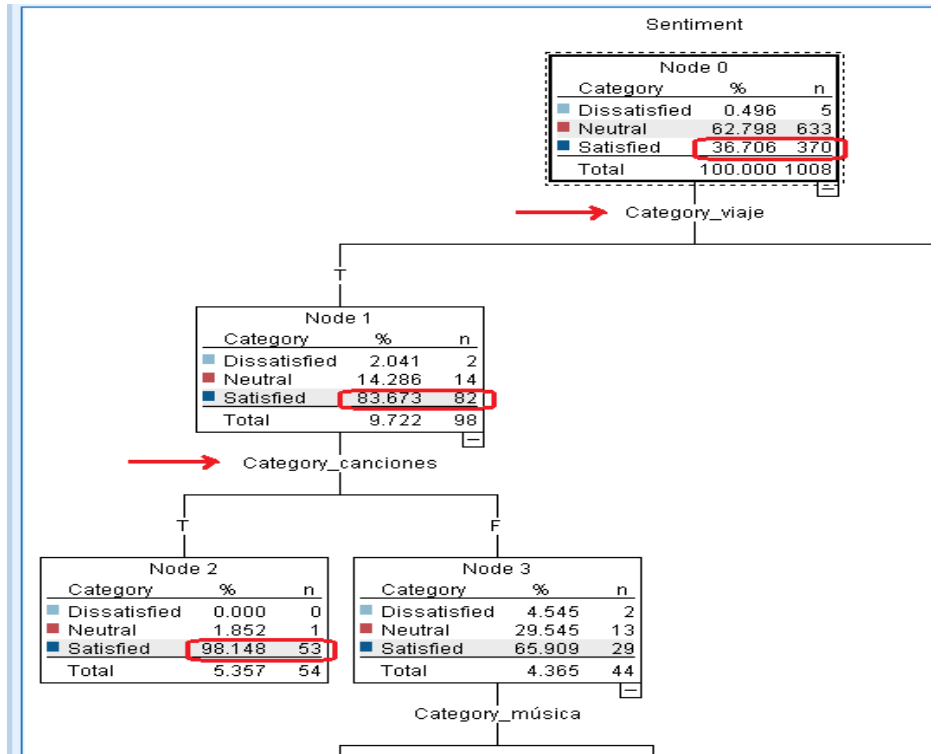
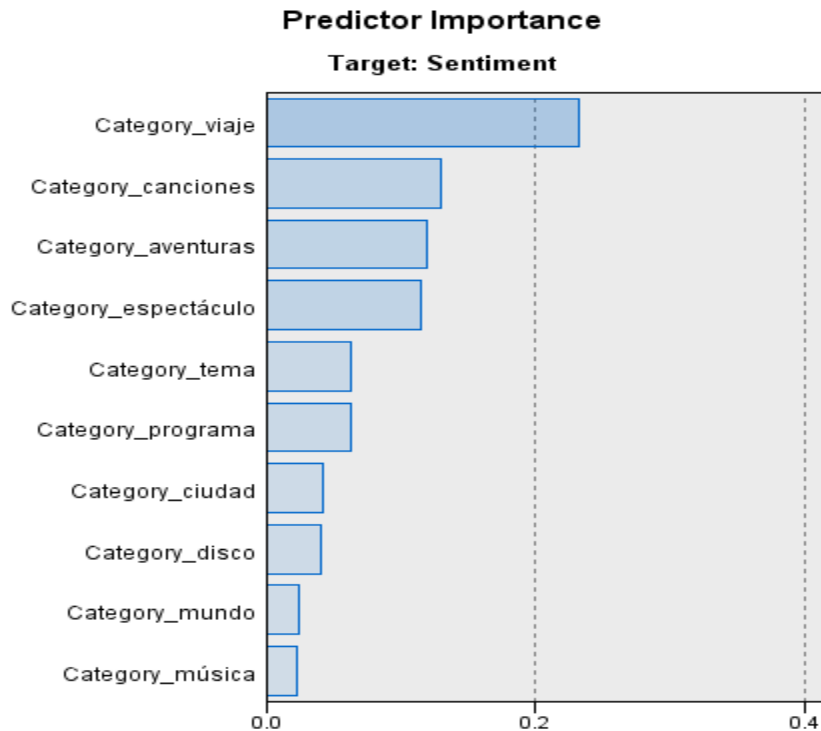
Lista de decisiones de las categorías (sin sentimiento)

Consequent	Antecedent	Support %	Confidence %
Category_canciones	Category_musical Category_espectáculo Category_aventuras	12.389	95.238
Category_aventuras	Category_musical Category_espectáculo Category_canciones	12.389	95.238
Category_canciones	Category_musical Category_aventuras	22.714	94.805
Category_canciones	Category_viaje Category_musical	10.914	94.595
Category_canciones	Category_espectáculo Category_aventuras	16.814	92.982
Category_aventuras	Category_musical Category_canciones	24.484	87.952
Category_disney junior	Category_casa	12.094	85.366
Category_disney junior	Category_programa Category_show	10.029	85.294
Category_canciones	Category_viaje	16.814	82.456
Category_canciones	Category_viaje Category_aventuras	10.029	82.353
Category_país	Category_casa Category_disney junior	10.324	80.0
Category_disney junior	Category_programa Category_musical	10.324	80.0

La primer fila de la tabla se lee:
 Los snippets que identificaron
 simultáneamente a las categorías
 MUSICAL – ESPECTACULO –
 AVENTURAS, también identifican
 a la categoría CANCIONES en un
 95.238% de ellos Snippets.

A su vez es el 12.389% de la
 población total de snippets los
 que se identificaron
simultaneamente a las categorías
 MUSICAL – ESPECTACULO –
 AVENTURAS.

Categorías relevantes al sentimiento asignado



PREDICTORES: Tomando como válido el sentimiento dado por SMA (al fin de cuentas no son muy distintos del de SPSSTA) resulta que la Cat. VIAJE es la más «separa» el valor 1: 'Dissatisfied' del 3: 'Satisfied' (i.e. que menos 2: 'Neutral' aplica). Asimismo Cat. CANCIONES es la que sigue en importancia, y así con todas las demás categorías.

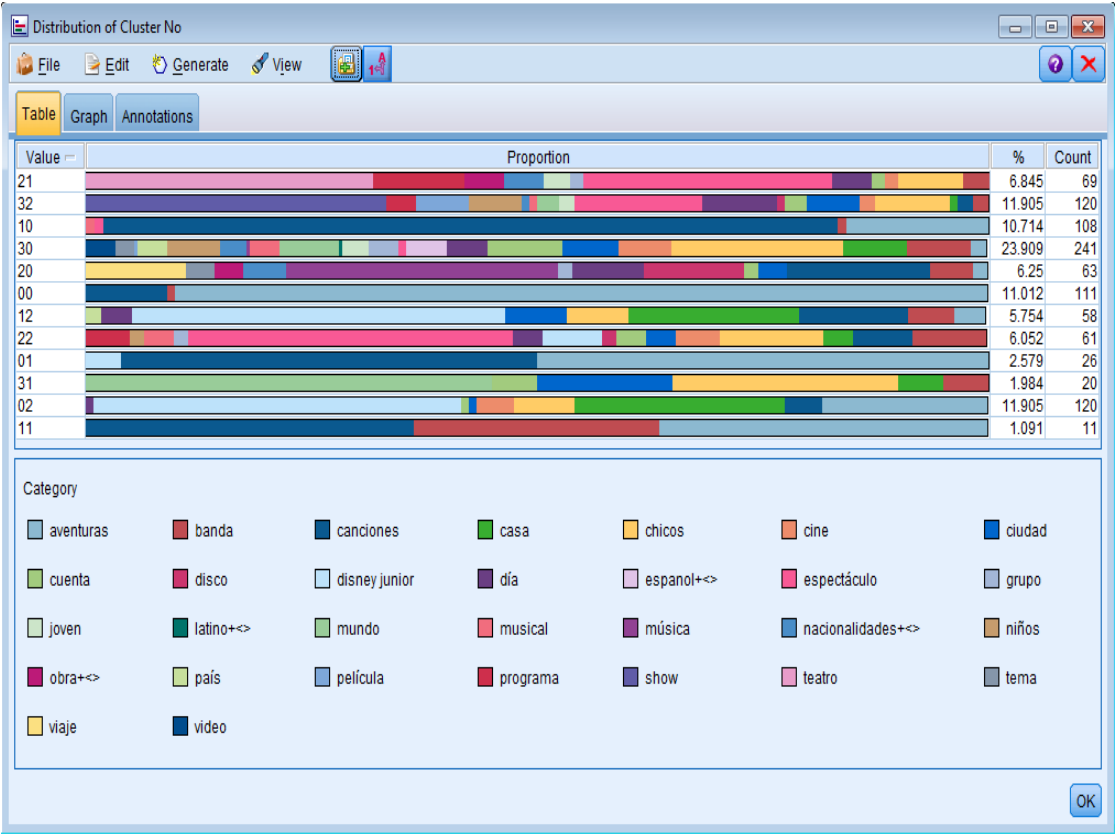
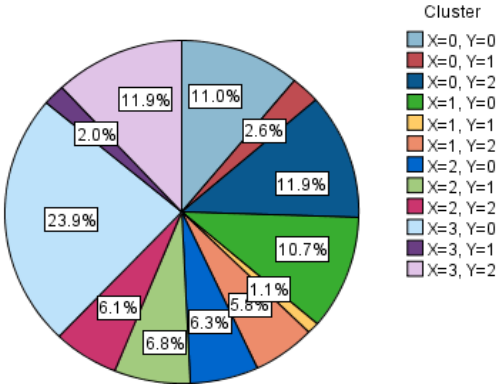
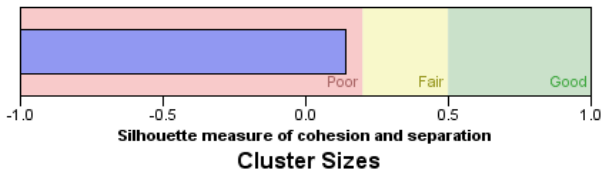
ARBOL DE DECISION: El 36.706% de la población de snippets posee conceptos de satisfacción (o sea valoraciones positivas). Además cuando los snippets aplican a la cat. VIAJE la satisfacción pasa al 83.672% Con esto podríamos afirmar que al tomar aleatoriamente una muestra, el 22% (= 82/370) de ellos poseerán alta satisfacción. Asimismo si además de la cat. VIAJE los snippets aplican a la cat. CANCIONES la satisfacción es casi completa con el 98.141%

Segmentación de Kohonen:

Model Summary

Algorithm	Kohonen
Inputs	28
Clusters	12

Cluster Quality



La calidad de la similaridad de Kohonen es un poco pobre y esto es porque de los 12 clústers determinados hay varios de ellos que poseen muchas categorías distintas (i.e. varios colores en un mismo clúster) entre ellos por ejemplo el mayor de ellos (con el 23% de los Snippets) que posee a TODAS las categorías. Hubiésemos preferido obtener, si bien muchos clústers (como tantas categorías), pero al menos los mayores con pocas categorías dentro como para saber qué categorías van juntas.

Preguntas?