Mineração de Dados - inf-0613
Trabalho Final

2018

Paulo Roberto de Almeida Costa

Renan Lordello de Aguiar

Questão 1

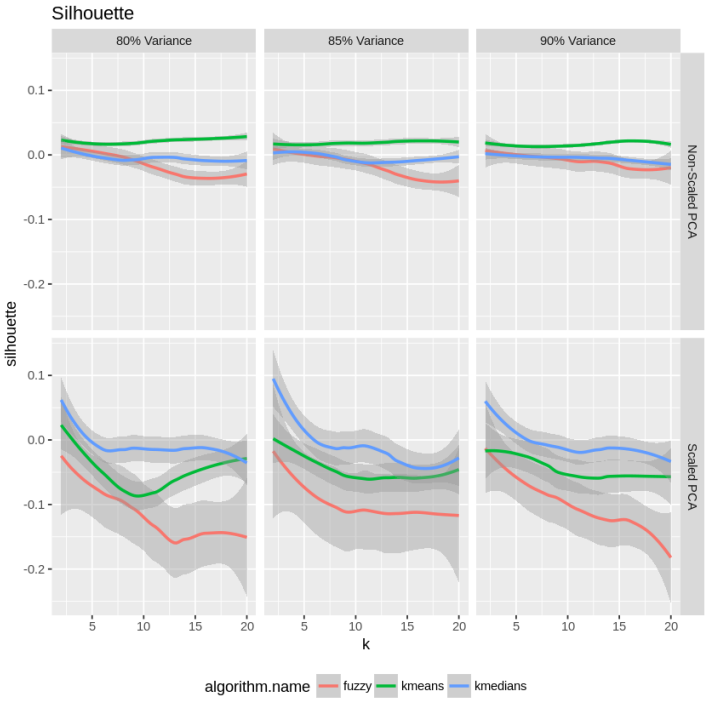
Essa parte do trabalho foi bem direta carregamos o arquivo de features em memória e rodamos o pca com e sem escala e analisando a variância acumulado geramos a tabela abaixo, para responder o questionamento de qual seria o número de componentes principais necessárias para se manter 85% e 90% da variância dos dados iniciais.

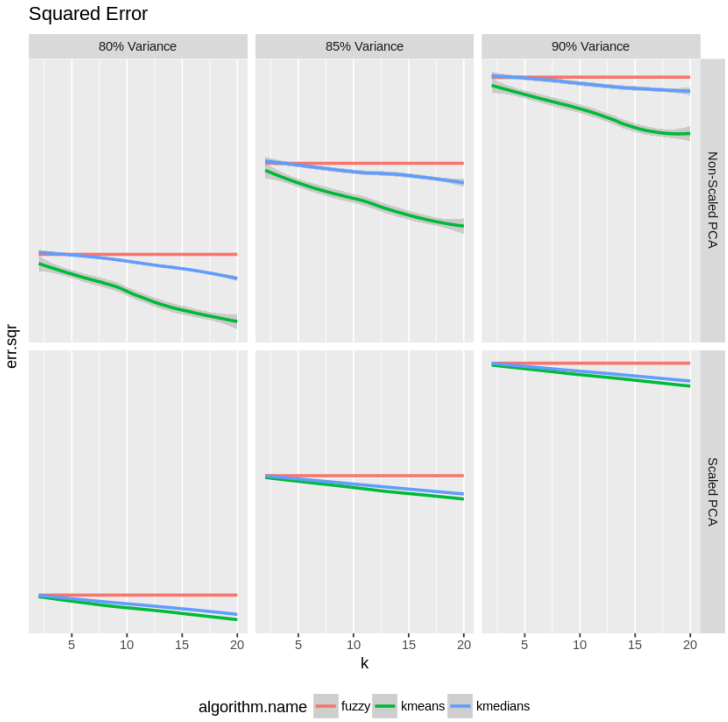
Resultados do PCA:

Variância Acumulada	Nº de Features s/ Normalizar	Nº de Features Normalizado
0.80	1210	1515
0.85	1390	1654
0.90	1598	1804
0.95	1845	1971

Questão 2

Rodamos um grid search variando a variância mantida pelo PCA, se os dados passaram por scalining, o número de clusters de 4 a 20 e também qual algoritmo fora usando, entre kmeans, kmedians, e fuzzy. O resultado dessa busca são os gráficos abaixo, que representam o coeficiente de silhouette e erro quadrático.
Analisando os gráficos abaixo podemos observar que neste dataset não normalizar os dados antes do PCA diminui o erro mas mantém a silhouette perto de zero ou seja há sobreposição entre os clusters, já a normalização tem o efeito contrário, temos uma melhora na separação mas um aumento no erro, indicando clusters mais esparsos, assim o melhor número de clusters é 17 para os dados não normalizados e de 5 para o dados normalizados.





Questão 3

Usando a biblioteca NLP geramos e contamos os bigramas para cada cluster (mostrado abaixo), analisando tais bigramas podemos ver que quando não normalizamos os dados e

usamos 17 clusters, cada cluster tmr notícias parecidas apesar da sobreposição entre os clusters, já com a normalização e 4 clusters as notícias parecem estar misturadas.

Bigramas para dados não normalizados (1 cluster por linha)

1º	2º	3º
13 years	12 13	election live
rural news	national rural	rural qld
death toll	over death	charged over
extended interview	interview ben	interview john
clarke dawe	michael clarke	ahead ashes
share market	abc business	business news
country hour	hour podcast	wa country
world cup	melbourne cup	cup final
mental health	health service	health minister
man charged	man dies	charged over
abc weather	abc sport	abc business
about future	rugby league	premier league
tony abbott	face court	people smuggler
pakistan court	killed pakistan	pakistan bombings
new zealand	new york	new england
gold coast	donald trump	charged over
talks about	nrl grand	nrl live

Bigramas para dados normalizados (1 cluster por linha)

1º	2º	3º
epa investigates	mango quality	water quality
shopping centre	released from	denied bail

cow corner	allergy free	cow produces
country hour	charged over	world cup
lleyton hewitt	into semis	australian open

Questão 4

Repetindo a análise para o subset das notícias de 2016, chegamos em resultados parecidos com o que já tínhamos, usamos assim a mesma configuração de 17 e 4 clusters. E analisando os bigrams notamos que temas como prisões, esportes e negócios repetem entre o todo e o subset de 2016.