Terceiro Trabalho Prático

Luis Gustavo Lorgus Decker
209819
luisgustavo.decker@gmail.com

Luiz Antonio Falaguasta Barbosa
075882
lafbarbosa@gmail.com

I. Introdução

II. Soluções Propostas

III. Experimentos e Discussão

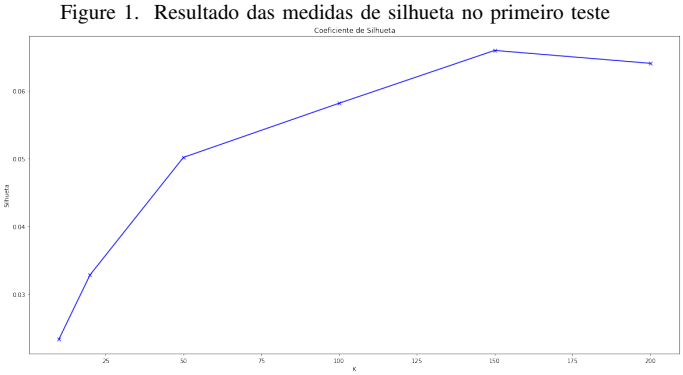
Primeiro, executamos o algorítmo de k-means nas informações provenientes do bag-of-words dos documentos. Como este algorítimo exige que o número de classes seja dado como parâmetro, testamos com vários valores visando verificar a eficácia. Para averiguar os resultados, utilizamos a informação do campo *Newsgroups* dos documentos. Estes metadados nos dão uma boa noção sobre o assunto do documento, e é esperado que documentos com assuntos semelhantes, de *Newsgroups* semelhantes ou iguais sejam agrupados juntos.

Para visualizar estes dados, na convergência do k-means analisamos os documentos pertinentes a cada cluster e então salvamos um arquivo contendo os *Newsgroups* do medóide de cada cluster e de seus 10 vizinhos mais próximos, obtidos através do algorítimo de KNN, e também a contagem de ocorrências de cada *Newsgroups* dos documentos do cluster.

Utilizamos também a medida de silhueta aplicada ao resultado de cada teste, visando por meio deste e da confirmação através do arquivo gerado com informações de *Newsgroups* encontrar o melhor K para nosso problema.

No nosso primeiro teste, utilizamos valores de k de 10, 20, 50, 100, 150 e 200 clusters, e então analisamos os resultados.

Figure 1. Resultado das medidas de silhueta no primeiro teste



IV. Conclusão e Trabalhos Futuros

References

[1] Christopher M. Bishop. “Pattern Recognition and Machine Learning”. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[2] Alex Krizhevsky et al. “The CIFAR-10 dataset”. https://www.cs.toronto.edu/ kriz/cifar.html.

[3] Bottou L. (2010) Large-Scale Machine Learning with Stochastic Gradient Descent. In: Lechevallier Y., Saporta G. (eds) Proceedings of COMPSTAT’2010. Physica-Verlag HD

[4] Mark Schmidt, Nicolas Le Roux, Francis Bach. "Minimizing Finite Sums with the Stochastic Average Gradient". Revision from January 2015 submission.

[5] Böhning, D. "Multinomial logistic regression algorithm". Ann Inst Stat Math (1992)

[6] Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors." Cognitive modeling 5.3 (1988): 1.