# 1 Exercícios Práticos

T1) Cosidere o arquivo `estatistica-metricas.rar`, disponível no site (`http://www.inf.ufes.br/~elias/dataSets/estatistica-metricas.rar`). Ao descompactá-lo, você encontrará vários outros arquivos no formato: **nome**`-metrica.dat`. Por exemplo, `clark-metrica.dat` e `jensenshannon-metrica.dat` são dois exemplos. Utilize utilitários Linux para colocar os valores de cada um desses como colunas de uma matriz. Crie um *makefile* para execução desse processo e submeta esse *makefile* e os códigos *shell-scripts*, caso você se utilize de algum.

T2) Tendo como exemplo a transformação do texto-documento em vetores feita no artigo (OLIVEIRA et al., 2007), utilize o indexador `p-indexer` para transformar os documentos da base de dados `orthogonals*.txt`, disponível em (`http://www.inf.ufes.br/~elias/dataSets/basic-datasets.tar.gz`). Esta atividade terá duas partes: A) Escreva um relatório de como utilizar esse utilitário para realizar essa atividade, desde a entrada na conta de usuário no LCAD, parâmetros necessários, arquivos gerados por conta da execução do `p-indexer` e suas respectivas importâncias para o processo de indexação, use também a referência (BÜTTCHER; CLARKE; CORMACK, 2010, Cap. 1 e 2) para fundamentar seus argumentos. B) Crie um *script* para gerar um gráfico com os pontos-documentos em um espaço $\mathbb{R}^3$. Submeta o relatório como um arquivo PDF, os *scripts* e o gráfico gerado pelo *script*.

T3) Considere o enunciado da atividade (T1). Crie um makefile para processamento da matriz gerada naquela atividade e, agora, calcule a média de cada uma das linhas dessa matriz. Além disso, calcule também a média das colunas. O primeiro resultado você escreverá em um arquivo de nome `row-mean.dat`, no formato de matriz coluna. O segundo resultado no arquivo `column-mean.dat`, no formato matriz linha. Além dos utilitários Linux utilizados na atividade acima citada, use agora também a linguagem R para realização dessa atividade para o cálculo das médias.

T4) Considere o enunciado da atividade (T1 e T3). Crie um makefile para processamento da matriz gerada naquela atividade e, agora, estude e discuta os resultados de correlação para cada par de colunas dessa matriz. O que você teria a dizer sobre essas métricas? Submeta todos os utilitários e um relatório PDF com suas reflexões quando ao observado das correlações.

T5) Explicar o conceito de *berry-picking* e sua relação com a busca por informação.

T6) Explicar a *Information Faraging Theory* (*referência [1271, 1269]*) e responder: o que isso ajuda a um sistema de busca?

T7) Discuta as ideias de Bates (*referência [156]*).

T8) Como, em um *browser*, saber o caminho realizado por um usuário, por entre os *hyperlinks* visitados, e o tempo gasto em cada *link* visitado, ao entrar em uma página? Escreva um relatório que explique o procedimento e submeta os códigos necessários para isso.

T9) Como determinar que um documento é relevante a um tópico (*referência [1451, 1402, 1687]*)?

T10) Qual seria a proposta para um novo buscador diante da Seção 2.3.2?

T11) Como utilizar a correlação para a sugestão de resultados em um buscador?

T12) Leia o artigo escrito por Salton, Wong e Yang (1975) e prepare um seminário para explicar a metodologia descrita nele. Use a `aLine` e a base de dados de A Tribuna[1] para reproduzir aquele experimento apresentado na Tabela 1. Como no artigo (SALTON; WONG; YANG, 1975), você deverá imprimir no **arquivosaida** o fonte LaTeX para montagem da Tabela 1 com os dados processados da base A Tribuna (`http://www.inf.ufes.br/~elias/dataSets/aTribuna-21dir.tar.gz`). Além disso, um texto deverá ser acrescentado logo após o código da tabela descrevendo a tabela e discutindo os resultados. Pense no **arquivosaida** como um arquivo a ser incluido por um *template* de um artigo maior.

[1] (`http://www.inf.ufes.br/~elias/dataSets/aTribuna-21dir.tar.gz`)

# Referências


AMATI, G.; RIJSBERGEN, C. J. V. Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. *ACM Transactions on Information Systems (TOIS)*, ACM, New York, NY, v. 20, n. 4, p. 357–389, out. 2002. ISSN 1046-8188.

AZEVEDO, L. L. et al. Recuperação de Informação Através do Processo de Aproximações Sucessivas. In: *XXI Congresso Brasileiro de Biblioteconomia, Documentação e Ciência da Informação*. Curitiba: CBBD, 2005.

BAEZA-YATES, R.; RIBEIRO-NETO, B. *Modern Information Retrieval*. 2. ed. New York: Addison-Wesley, 2011.

BRIN, S.; PAGE, L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, v. 30, n. 1-7, p. 107–117, 1998.

BÜTTCHER, S.; CLARKE, C. L. A.; CORMACK, G. V. *Information Retrieval – Implementing and Evaluating Search Engines*. New York: MIT Press, 2010. Disponível em: <http://www.ir.uwaterloo.ca/book/>.

Campana Filho, J. C. *Uma Proposta e Avaliação de Sistema de Detecção de Suspeita de Plágio em Códigos de Linguagem C para Apoio à Atuação Docente*. Dissertação (Mestrado) — Programa de Pós-Graduação em Informática, Universidade Federal do Espírito Santo, Vitória, ES, nov. 2016. Disponível em: <http://www.informatica.ufes.br/pos-graduacao/PPGI/detalhes-da-tese?id=10537>.

Campana Filho, J. C.; OLIVEIRA, M. G.; OLIVEIRA, E. Classificação de Códigos C Usando Medidas de Similaridade para Apoio ao Ensino em Programação. In: *XXVII Simpósio Brasileiro de Informática na Educação (SBIE)*. Ceará, CE: SBC, 2016. Disponível em: <http://www.br-ie.org/pub/index.php/sbie/article/view/6801>.

CHURCH, K. W.; HANKS, P. Word Association Norms, Mutual Information, and Lexicography. *Comput. Linguist.*, MIT Press, Cambridge, MA, USA, v. 16, n. 1, p. 22–29, mar. 1990. ISSN 0891-2017. Disponível em: <http://dl.acm.org/citation.cfm?id=89086.89095>.

CRESTANI, F. et al. &Ldquo;Is This Document Relevant?&Hellip;Probably&Rdquo;: A Survey of Probabilistic Models in Information Retrieval. *ACM Comput. Surv.*, ACM, New York, NY, v. 30, n. 4, p. 528–552, dez. 1998. ISSN 0360-0300. Disponível em: <http://doi.acm.org/10.1145/299917.299920>.

CUNHA, E.; FIGUEIRA, A.; MEALHA, O. Clustering and Classifying Text Documents – A Revisit to Tagging Integration Methods. In: *5th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. Vilamoura, Algarve, Portugal: SCITEPRESS(Science and Technology Publications, Lda.), 2013. p. 160–168.

DHANABHAKYAM, M.; PUNITHAVALLI, M. A survey on data mining algorithm for market basket analysis. *Global Journal of Computer Science and Technology*, v. 11, n. 11, 2011.

GLOVER, F.; LAGUNA, M. *Tabu Search*. [S.l.]: Kluwer Academic Publishers, 1997.

GOLDBERG, D. E. *Genetic Algorithms in Search, Optimization, and Machine Learning*. New York, USA: Addison-Wesley Publishing Company, 1989.

GOLDSTEIN, J. et al. Summarizing text documents: Sentence selection and evaluation metrics. In: *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 1999. (SIGIR '99), p. 121–128. ISBN 1-58113-096-1. Disponível em: <http://doi.acm.org/10.1145/312624.312665>.

HERNÁNDEZ, R. A. G.; LEDENEVA, Y. Identification of Similar Source Codes Based on Longest Common Substrings. 2014.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. *Introduction to Information Retrieval*. New York: Cambridge University Press, 2008.

OLIVEIRA, E. et al. Combining Clustering and Classification Approaches for Reducing the Effort of Automatic Tweets Classification. In: *6th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. Rome, Italy: IC3K, 2014.

OLIVEIRA, E. et al. Um Modelo Algébrico para Representação, Indexação e Classificação Automática de Documentos Digitais. *Revista Brasileira de Biblioteconomia e Documentação*, Brasília, v. 33, n. 1, p. 75–98, 2007.

OLIVEIRA, E. et al. Using the Cluster-Based Tree Structure of k-Nearest Neighbor to Reduce the Effort Required to Classify Unlabeled Large Datasets. In: *7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. Lisbon, Portugal: IC3K, 2015.

PENG, H.; LONG, F.; DING, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE*

*Transactions on pattern analysis and machine intelligence*, IEEE, v. 27, n. 8, p. 1226–1238, 2005.

QUINLAN, J. R. *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

QUINLAN, J. R. Improved Use of Continuous Attributes in C4.5. *Journal of Artificial Intelligence Research*, v. 4, p. 77–90, 1996.

ROBERTSON, S. E.; SPARCK-JONES, K. Relevance Weighting of Search Terms. *Journal of the American Society for Information Science*, Wiley Online Library, New York, NY, v. 27, n. 3, p. 129–146, 1976.

SÁEZ, E. F. et al. PAN FIRE: Overview of SOCO Track on the Detection of SOurce COde Re-Use. ACM, Cham, 2014.

SALTON, G.; WONG, A.; YANG, C. S. A Vector Space Model for Automatic Indexing. *Commun. ACM*, ACM, New York, NY, USA, v. 18, n. 11, p. 613–620, 1975.

SAÚDE, M. R. et al. A Strategy for Automatic Moderation of a Large Data Set of Users Comments. In: *Computing Conference (CLEI), 2014 XL Latin American*. Montevideo, Uruguay: IEEE, 2014. p. 1–7.

SOUZA, F.; CIARELLI, P.; OLIVEIRA, E. Combinando Fatores de Ponderação para Melhorar a Classificação de Textos. In: *Computer on The Beach 2014*. Florianópolis, SC: SBC, 2014.

SOUZA, F. P. *Uma Combinação de Métodos de Ponderação para Melhoria da Classificação de Textos*. Dissertação (Mestrado) — Programa de Pós-Graduação em Informática, Universidade Federal do Espírito Santo, Vitória, ES, jan. 2014.

SPARCK-JONES, K.; WALKER, S.; ROBERTSON, S. E. A Probabilistic Model of Information Retrieval: Develepment and STatus. *Department of Information Science, City University, London*, Citeseer, v. 74, 1998.

SPARCK-JONES, K.; WALKER, S.; ROBERTSON, S. E. A Probabilistic Model of Information Retrieval: Development and Comparative Experiments: Part 2. *Information Processing & Management*, ACM, v. 36, n. 6, p. 809 – 840, 2000. ISSN 0306-4573. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0306457300000169>.

TEIXEIRA, J. F.; COUTO, M. Automatic Distinction of Fernando Pessoas' Heteronyms. In: ____. *Progress in Artificial Intelligence: 17th*

*Portuguese Conference on Artificial Intelligence, EPIA 2015, Coimbra, Portugal, September 8-11, 2015. Proceedings*. Coimbra, Portugal: Springer International Publishing, 2015. p. 783–788. Disponível em: <http://dx.doi.org/10.1007/978-3-319-23485-4_78>.