Avaliação de métodos de detecção de tópicos em pré-processamento para classificação de textos agrícolas

Flavio M. M. Barros[1], *Stanley R. M. Oliveira*[2]

[1]Faculdade de Engenharia Agrícola, Universidade Estadual de Campinas
São Paulo, Brasil
flavio.barros@feagri.unicamp.br

[2]Embrapa Informática Agropecuária,
Campinas, São Paulo, Brasil
stanley.oliveira@embrapa.br

RESUMO

A classificação de documentos é uma tarefa útil para agrupamento, organização e também para recomendação de textos. Em particular, na agricultura em que boa parte dos usuários (e.g., agricultores e produtores rurais) de sistemas de informações não têm grande experiência com a tecnologia da informação, a classificação de documentos pode ser utilizada para gerar recomendações de leitura. Neste trabalho, buscou-se construir e comparar modelos capazes de diferenciar textos sobre a cultura da cana-de-açúcar de outros textos relacionados a outras culturas ou criações. Para criar modelos de classificação de textos, os dados são transformados em matrizes termos-documentos, de forma que os dados apresentam alta dimensionalidade. Para construir melhores modelos de classificação de textos agrícolas foram testados: a) métodos de redução de dimensionalidade utilizando LDA (Latent Dirichlet Allocation) e PCA (Principal Component Analysis); b) número de tópicos/componentes principais; c) unigrama/bigrama; e d) algoritmos Random Forest, Gradiente Boosting e SVM (Support Vector Machine), de forma a determinar os fatores que mais impactam o AUC (Area Under the Curve). Os resultados demonstraram que os fatores estatisticamente significativos são o tipo de pré-processamento, com vantagem para LDA, e o tipo de algoritmo utilizado, com destaque para o SVM. O número de tópicos e de componentes principais e o uso de unigrama e bigrama não tiveram efeito estatisticamente significativo na performance dos modelos em termos de AUC.

PALAVRAS-CHAVE: Mineração de textos, Aprendizado de máquina, Redução de dimensionalidade, Sistema de informação agrícola.

ABSTRACT

Document classification is a useful task for grouping, organizing and also for text recommendation. In particular, in agriculture in which a large proportion of users (E.g., farmers and rural producers) of information systems do not have much experience with information technology, document classification can be used to generate reading recommendations. In this work, we tried to construct and compare models capable of differentiate texts about the culture of sugarcane from other texts related to other cultures or creations. To create text classification models, the data is transformed into term-documents matrices, so that the data present high dimensionality. For building better models for classification of agricultural texts, we have tested: a) methods for dimensionality reduction using LDA (Latent Dirichlet Allocation) and PCA (Principal Component Analysis); b) number of topics / main components; c) unigram / bigram; and d) algorithms Random Forest, Gradient Boosting, and Support Vector Machine (SVM) to determine the factors that most impact the AUC (Area Under the Curve). The results demonstrated that the statistically significant factors are the type of pre-processing, with advantage for LDA, and the type of algorithm used, highlighting the SVM one. The number of topics and principal components, the use of unigram and bigram did not have a statistically significant effect in the performance of the models in terms of AUC.

KEYWORDS: Text mining, Machine learning, Dimensionality reduction, Agricultural information systems.

INTRODUÇÃO

A agricultura é um dos setores protagonistas da economia brasileira cada vez mais dependente de informação. Essa dependência se dá pelas necessidades advindas das constantes mudanças tecnológicas, pela operação em mercados globais e pela extrema sensibilidade às mudanças ambientais (climática, hidrológica, solo e etc). Assim o processo de tomada de decisões de longo e médio prazos, em resposta a esses e outros fatores, requer um entendimento e o acesso rápido e fácil a uma grande quantidade de informações técnicas agrícolas (CASH, 2001). O acesso facilitado à informação de qualidade é particularmente crítico para pequenos produtores, que em geral têm maiores dificuldades de acesso à informação sobre as melhores práticas na agricultura (PARIKH; PATEL; SCHWARTZMAN, 2007).

No Brasil e no mundo há diversas iniciativas com o objetivo de ofertar informações técnicas agrícolas de qualidade. E especialmente no Brasil, destaca-se um projeto da EMBRAPA (Empresa Brasileira de Pesquisa Agropecuária) chamado Agência Embrapa de Informação, um dos portais mais acessados da internet brasileira na área de agricultura e que compreende centenas de páginas web, artigos, planilhas e materiais multimídia.

A fonte primária de informação do portal são textos técnicos agrícolas que condensam o conhecimento de anos de pesquisa em diversas unidades da Embrapa em todo o Brasil, compreendendo mais de 33 cultivos e criações, como cana-de-açúcar, soja, agronegócio do leite,

dentre outros. Todo o material está organizado em uma estrutura hierárquica, tal que textos relacionados a culturas específicas, como cana-de-açúcar, são agrupados em tópicos dentro do portal. A organização do material é feita pelos próprios pesquisadores da Embrapa e não é feita de forma automática.

De acordo com Barros et al. (2013), após uma investigação sistemática dos dados de uso da Agência, os pesquisadores concluíram que os os usuários podem não estar fazendo uso da totalidade de informações técnicas disponíveis. Muitos usuários visualizam somente uma parte muito pequena do material disponível por causa da dificuldade de encontrar a informação desejada em um grande volume de páginas Web. Ainda que as páginas estejam organizadas em uma estrutura de tópicos, há a possibilidade de que os usuários não tenham conhecimentos técnicos para encontrar rapidamente a informação desejada. Outro desafio é que a organização humana dos conteúdos pode deixar de lado relações importantes entre assuntos disponíveis em páginas diferentes, de forma que mesmo que um usuário utilize a estrutura hierárquica de tópicos ainda pode ter dificuldades em encontrar um texto correspondente às suas necessidades.

Em cenários como esse, um sistema automatizado, capaz de identificar o assunto relacionado a uma página sendo visitada por um usuário, poderia utilizar esta informação para gerar recomendações e/ou entender o perfil dos usuários. As informações do perfil da comunidade e as recomendações podem facilitar a navegação e exploração dos conteúdos da Agência. Entretanto, esta tarefa coloca diversos desafios técnicos: a) como obter a melhor representação estruturada do texto?; b) como escolher técnicas de classificação com bom desempenho no contexto?; c) como usar a representação estruturada para exploração dos hábitos de uso do portal?

Para abordar esses desafios, este trabalho teve como objetivo criar representações dos textos da Agência, utilizando ferramentas para detecção automática de tópicos e modelos de aprendizado de máquina capazes de classificar textos agrícolas sem intervenção humana. Especificamente, foram avaliados vários métodos de detecção de tópicos utilizados na etapa de pré-processamento para classificação de textos agrícolas.

MATERIAL E MÉTODOS

A metodologia deste trabalho é composta de algumas etapas relacionadas à detecção de tópicos de interesse para classificação de textos agrícolas.

Origem dos Dados

A coleção de textos utilizada foi extraída da Agência Embrapa de Informação [1], constituída de mais de 30 culturas agrícolas, com abrangência nacional. O usuário tem acesso a todo o conteúdo do site na forma de textos, artigos, arquivos de imagem, arquivos de som e planilhas eletrônicas. Todo o conteúdo foi organizado para atender pesquisadores, produtores rurais, profissionais de assistência técnica e extensionistas. De acordo com (BERTIN; LEITE; PEREIRA,

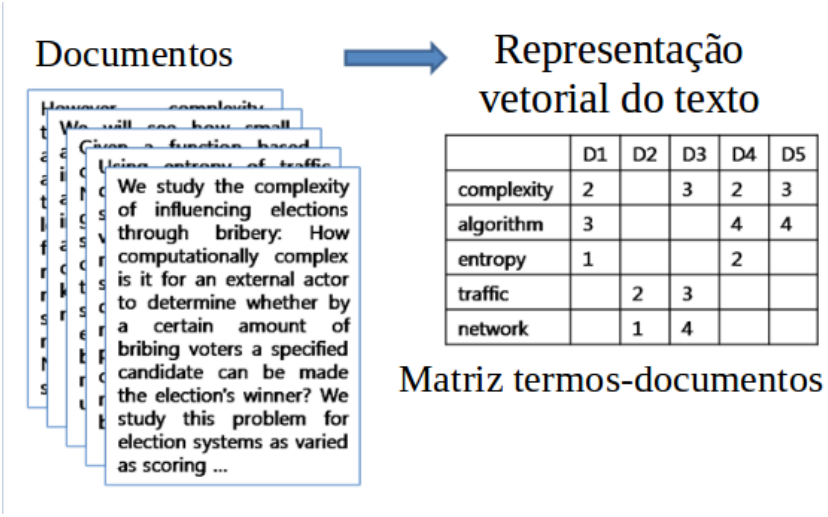
[1] http://www.agencia.cnptia.embrapa.br

2009), a EMBRAPA tem mantido uma política de oferta de informações técnicas para pesquisadores, produtores e a sociedade em geral, com a missão de fazer chegar aos agentes interessados os resultados da pesquisa científica, relativa a toda cadeia produtiva do agronegócio, com o objetivo de criar uma robusta infraestrutura social, técnica e econômica tão necessária ao processo de desenvolvimento. Todos os textos agrícolas utilizados neste artigo foram retirados de páginas web nesse portal.

Os dados utilizados neste artigo foram textos agrícolas completos das páginas da Agência Embrapa de Informação sobre culturas como, por exemplo, a cana-de-açúcar. Devido ao grande volume de dados disponível no portal foi utilizada uma amostra aleatória de 1.500 páginas diferentes, contando com 1.086 páginas com tema sobre a cana-de-açúcar e 414 páginas relacionadas a outros cultivos, como arroz, açaí dentre outros. Destaca-se que textos relacionados à cultura da cana-de-açúcar são mais numerosos e somente esta cultura corresponde a mais de 45% dos acessos ao portal (BARROS et al., 2013).

Processamento de Dados

O processamento dos dados foi constituído basicamente de duas etapas: 1) a transformação dos textos em uma matriz de termos-documentos e 2) extração de tópicos/componentes principais. A primeira etapa é característica de qualquer análise de mineração de textos e consiste basicamente em transformar dados não estruturados, nesse caso textos, em dados estruturados em uma tabela, onde as colunas (ou linhas) são os termos e as linhas (ou colunas) são documentos, uma representação conhecida na literatura como *bag-of-words* (FELDMAN; SANGER, 2006). Salienta-se que no caso específico deste trabalho, optou-se por usar a medida TF-IDF (*term frequency-inverse document frequency*) na construção da matriz de termos-documentos. Na Figura 1 é mostrado um exemplo desse tipo de processamento dos textos.



	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

Figura 1: Transformação de textos em uma matriz termos-documentos. Os números representam a frequência com que cada termo aparece no respectivo documento, mas no trabalho foi utilizada a TF-IDF.

Utilizando a matriz de termos-documentos como entrada, foi utilizada a técnica de detecção de tópicos LDA (Latent Dirichlet Allocation) como descrita em Blei, Ng e Jordan (2003). Também, a partir da mesma matriz de termos-documentos foi utilizada a técnica PCA (Principal Component Analysis) para encontrar as componentes principais, como descrito em Hastie, Tibshirani e Friedman (2009). O número de tópicos/componentes principais é um parâmetro que deve ser escolhido pelo usuário, e optou-se neste trabalho em utilizar 30, 50 e 100 tópicos ou componente principais.

Modelos e Validação

A avaliação empírica neste artigo consistiu em um experimento que combinou dois tipos diferentes de técnicas de redução de dimensionalidade, LDA e PCA. Além disso também foram incluídos no experimento a avaliação do uso de unigramas e bigramas; a avaliação dos seguintes algoritmos (HAN, 2005): a) SVM b) Random Forest e c) Gradiente Boosting e o número de tópicos/componentes principais. Portanto, o experimento contou com a avaliação de quatro fatores diferentes: 1) a técnica de redução de dimensionalidade 2) bigramas ou unigramas 3) algoritmos e 4) número de tópicos/componentes principais. O objetivo principal do experimento foi comparar qual era o melhor método de redução de dimensionalidade. No entanto, apesar de se encontrar ampla literatura reportando performances de algoritmos, optou-se por incluir esses fatores no experimento deste artigo de forma a se obter a melhor configuração experimental para o contexto de textos agrícolas. Na avaliação foi utilizada a métrica Área sob a curva ROC (AUC), amplamente utilizada na literatura como métrica de comparação de classificadores (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Para os experimentos, foi utilizada validação cruzada com 10 folds já que ajustes de hiperparâmetros foram realizados para indicar a melhor configuração entre classificadores e métodos de redução de dimensionalidade.

Ferramentas de Suporte

Todas as etapas do trabalho foram realizadas com a linguagem R[2]. Foram utilizados os pacotes `tm` para criação do Corpus de textos, stemização, limpeza, tokenização; o pacote `rvest` para web scrape; o pacote `topicmodels` para modelagem de tópicos com o LDA; o pacote `caret` para modelagem e teste; as funções `lm()` e `prcomp()` do pacote base para o ANOVA e PCA, respectivamente.

RESULTADOS E DISCUSSÃO

Foram realizados vários experimentos com o propósito de avaliar os melhores métodos para detecção e tópicos de interesse nos textos agrícolas. Quatro aspectos foram considerados:

[2]http://cran.r-project.org

redução de dimensionalidade com LDA/PCA, número de tópicos/componentes principais, uso de bigramas de unigramas e algoritmos. No total 36 configurações experimentais foram analisadas com a técnica ANOVA (DEVORE, 2008). De acordo com os resultados na Tabela 1, as técnicas de redução de dimensionalidade (Pré-processamento) e os algoritmos utilizados (Algoritmos) foram estatisticamente significativos (p-valor < 0.001); ambos, o número de tópicos/componentes principais (K) e bigrama/unigrama (ngrama), não apresentaram efeito significativo na métrica utilizada. Todos os modelos ficaram bem ajustados ao conjunto de dados ($R^2 > 0.6750$).

A Figura 2 mostra o histograma do AUC em relação aos algoritmos, usando todas as configurações experimentais consideradas nesse trabalho. Como pode ser observado, existe uma separação clara entre o desempenho dos algoritmos, com ligeira vantagem para o SVM. Percebe-se também que a distribuição do histograma para o SVM é bimodal, indicando que esse algoritmo tem efeito de interação com outro fator, no caso o uso do LDA ou do PCA.

Tabela 1: Tabela ANOVA do experimento.

	Mean Sq	F value	Pr(>F)
Pré-processamento	0.00155612	32.0432	$< 4 \times 10^{-6} ***$
K	0.00003526	0.7261	0.4009
ngrama	0.00005975	1.2304	0.2761
Algoritmos	0.00068758	14.1584	$< 5 \times 10^{-5} ***$
Residuals	0.00004856		
R^2	0.6750		
Adj. R^2	0.6209		
Num. obs.	36		

***p-value< 0.001

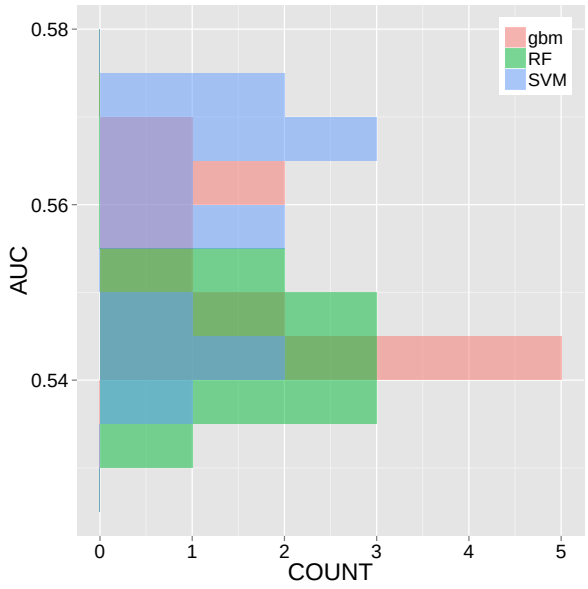


Figura 2: Histograma de AUC em relação aos algoritmos.

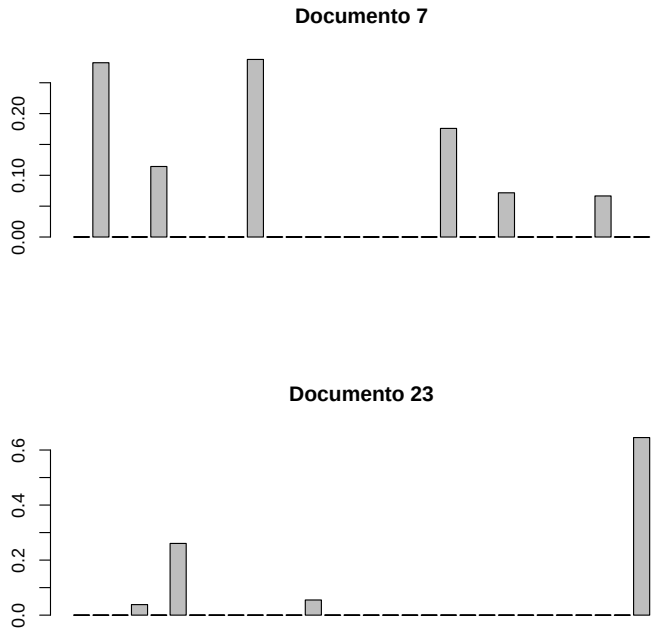


Figura 3: Exemplo de distribuição da probabilidade de tópicos para dois documentos da base de dados.

Na Figura 3 são apresentados dois exemplos da distribuição da probabilidade de tópicos em dois documentos da coleção, o Documento 7 e o Documento 23. Os valores representados nas barras são probabilidades tal que o valor total somado de todas as barras deve ser 1. Ele indica a probabilidade da presença de um certo tópico (ou conjunto de tópicos) em um certo documento. Assim, a estrutura de tópicos em um documento pode ser utilizada como um representação do texto, sendo uma alternativa a representação da matriz termos-documentos, onde cada termo em específico no texto é utilizado na representação. Utilizando tópicos, ao invés dos termos, como há menos tópicos, a representação do texto fica mais enxuta. Além disso, analisando as palavras que em conjunto formam cada tópico é possível fazer uma descrição qualitativa, mas sucinta, dos textos.

Na Figura 4 observa-se o histograma da métrica AUC em relação à técnica de redução de dimensionalidade utilizada. Oberva-se claramente que o LDA tem um desempenho superior na maioria das condições experimentais. Tanto na Figura 2 quanto na Figura 4, histogramas relativos aos dois fatores estatisticamente significativos da Tabela 1, observa-se que a configuração que apresenta melhores resultados é a configuração com LDA + SVM.

Como o teste estatístico da Tabela 1 rejeita a hipótese nula de que os fatores do experimento não tem efeito significativo, mas alguns fatores apresentam mais de um nível, como o fator algoritmos, para diferenciar em quais níveis há diferença estatisticamente significativa é necessário utilizar um teste *posthoc* como o Teste de Tukey (DEVORE, 2008). De acordo com os intervalos de confiança para as diferenças entre as médias para os algoritmos, no gráfico

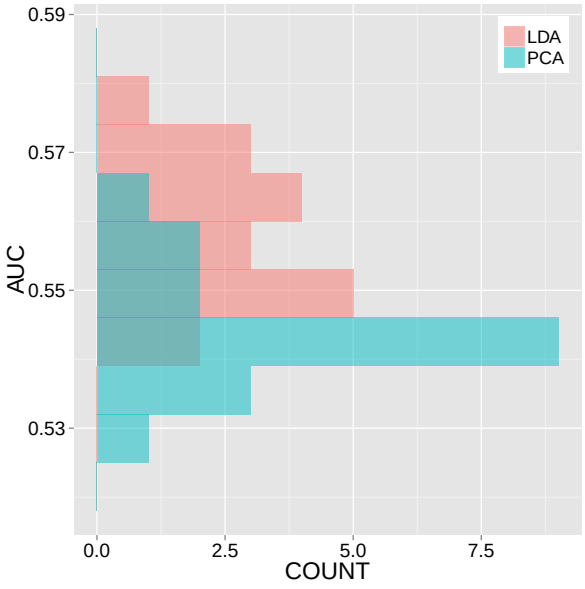


Figura 4: Histogram de AUC em relação à técnica de redução de dimensionalidade.

apresentado na Figura 5, o SVM apresenta desempenho superior aos dois outros algoritmos. Nota-se entretanto, que devido à bimodalidade da distribuição dos valores de AUC em relação ao algoritmo SVM, a diferença de desempenho entre os algoritmos SVM e Gradiente Boosting foi pequena e próxima do limite da margem de erro.

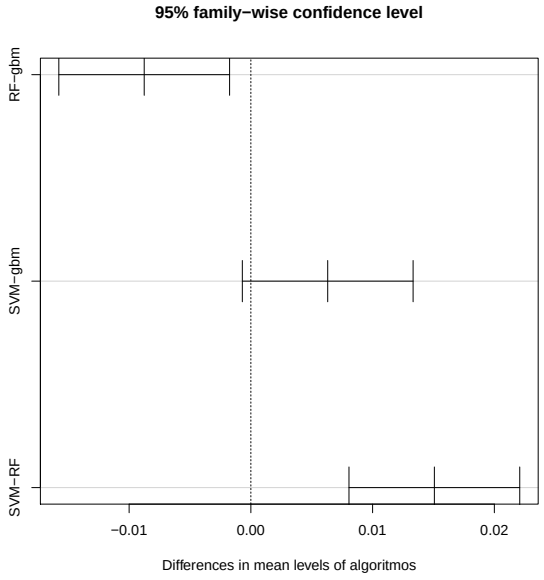


Figura 5: Intervalos de confiança para a diferença entre as médias do AUC para cada algoritmo.

Em resumo, a avaliação mostrou que o LDA usado como técnica de redução de dimensionalidade tem um desempenho superior ao PCA. Também mostrou-se que o SVM associado à técnica LDA tem um bom desempenho em termos de classificação de documentos. O uso de bigramas, pelo menos na tarefa de classificação de textos, não foi um fator significativo, tal

que pode-se optar por utilizar somente unigramas para reduzir a dimensão das matrizes termos-documento. Outra vantagem do LDA em relação ao PCA é que a técnica também permite analisar o significado de cada atributo utilizado após a redução de dimensionalidade, isto é, uma vez que cada documentos apresenta a prevalência de alguns tópicos, como apresentado na Figura 3, além da redução de dimensionalidade, esta informação pode ser utilizada de forma exploratória e para determinar os assuntos mais importantes para a comunidade de usuários.

CONCLUSÕES

Esse trabalho apresenta uma avaliação de métodos de detecção de tópicos, um procedimento utilizado no pré-processamento de classificação de textos. Vários experimentos foram realizados para avaliar os desempenhos das técnicas LDA e PCA para redução de dimensionalidade em matrizes de termos-documentos. Os resultados evidenciaram o potencial da técnica LDA para essa tarefa. O método proposto é capaz de melhorar o desempenho de classificadores em relação a técnicas tradicionais de redução de dimensionalidade como o PCA, especialmente quando se considera o algoritmo SVM na etapa principal do processo de mineração de dados.

A técnica LDA também tem mostrou grande potencial como ferramenta exploratória, podendo ser utilizada para caracterizar o interesse dos usuários em determinados assuntos. No futuro planeja-se avaliar outras técnicas de redução de dimensionalidade em comparação ao LDA, como *bag-of-related-words* e Lasso (least absolute shrinkage and selection operator).

REFERÊNCIAS

BARROS, F. d. F. M. M. de et al. Desenvolvimento e validação de um sistema de recomendação de informações tecnológicas sobre cana-de-açúcar. *Bragantia*, Instituto Agronômico, v. 72, n. 4, p. 1–9, 2013. ISSN 1678-4499. Disponível em: <http://www.scielo.br/pdf/brag/v72n4/aop_bragncea2061.pdf http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0006-87052013000400010&lng=pt&nrm=iso&tlng=en>.

BERTIN, P.; LEITE, F.; PEREIRA, F. Embrapa Technological Information: a bridge between research and society. *Agricultural Information Worldwide*, v. 2, n. 1, p. 10–18, 2009. Disponível em: <http://repositorio.bce.unb.br/handle/10482/4935>.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res.*, JMLR.org, v. 3, p. 993–1022, mar. 2003. ISSN 1532-4435. Disponível em: <http://dl.acm.org/citation.cfm?id=944919.944937>.

CASH, D. W. In order to aid in diffusing useful and practical information: Agricultural extension and boundary organizations. *Science, Technology & Human Values*, v. 26, n. 4, p. 431–453, out. 2001. ISSN 0162-2439, 1552-8251. Disponível em: <http://sth.sagepub.com/content/26/4/431>.

DEVORE, J. *Probability and Statistics for Engineering and the Sciences, Enhanced Review Edition*. Cengage Learning, 2008. (Available 2010 Titles Enhanced Web Assign Series). ISBN 9780495557449. Disponível em: <http://books.google.co.uk/books?id=Wbym40WgsXMC>.

FELDMAN, R.; SANGER, J. *Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. New York, NY, USA: Cambridge University Press, 2006. ISBN 0521836573, 9780521836579.

HAN, J. *Data Mining: Concepts and Techniques*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005. ISBN 1558609016.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning*. 2. ed. New York, NY: Springer New York, 2009. 745 p. (Springer Series in Statistics). ISBN 978-0-387-84857-0. Disponível em: <http://www.springerlink.com/index/10.1007/978-0-387-84858-7>.

PARIKH, T. S.; PATEL, N.; SCHWARTZMAN, Y. A survey of information systems reaching small producers in global agricultural value chains. In: *2007 International Conference on Information and Communication Technologies and Development*. IEEE, 2007. p. 1–11. ISBN 978-1-4244-1990-6. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4937421>.