Busca e análise de notícias agrícolas sobre cana-de-açúcar

Rodrigo Bustamante Magalhães[1]
Maria Fernanda Moura[2]

Pretende-se fazer uma análise textual através do uso de mineração de texto, pelo método descrito por Moura et al. (2009), com o intuito de obter tópicos específicos, numa base de textos que consiste em notícias agrícolas sobre cana-de-açúcar. A base de textos é buscada por meio de um robô de busca em fontes previamente escolhidas (Jornal Cana: http://www.jornalcana.com.br). Após extração das notícias, é feito o tratamento dos textos e obtidas taxonomias de tópicos, isto é, são identificados assuntos e categorias cobertos pelo texto. Com os resultados da análise, pretende-se posteriormente localizar no espaço e no tempo os tópicos extraídos, para possibilitar uma inferência de fatores sócio-econômicos que possam influenciar na produção e safra da cana-de-açúcar.

Para a extração de notícias, vem sendo desenvolvida uma ferramenta de nome Newscap, que se utiliza da API HTTPUnit, em Java. Tal ferramenta faz recortes da página de *WebClipping* do JornalCana, extraindo as notícias para uma pasta local no disco rígido. A partir das notícias coletadas, o pré-processamento dos textos é realizado por meio da ferramenta PreText (SOARES et al., 2008) ou TextNSP (BANERJEE; PEDERSEN, 2003), guardando os resultados em uma

[1] *Universidade Estadual de Campinas; rodrigo@cnptia.embrapa.br*

[2] *Embrapa Informática Agropecuária; fernanda@cnptia.embrapa.br*

representação matricial. Nessa matriz, as linhas correspondem a cada documento, as colunas a cada atributo e cada célula a um valor de associação entre atributos e documentos (por exemplo, frequência). Um atributo pode ser uma palavra ou uma composição de palavras, aqui chamado n-grama. Com a ferramenta PreText II (SOARES et al., 2008) obtém-se n-gramas *stemmizados* ou não. A *stemmização* consiste na remoção de inflexões das palavras, o que permite reduzir o número de atributos. A Pretext II ainda permite que se executem filtros de atributos tais como: a retirada de stopwords, frequência mínima e máxima de cada atributo na coleção e/ou em documentos da coleção. A PreText gera a representação matricial de interesse; porém, para a TextNSP, foi desenvolvida a ferramenta N2Discover, em Java, que converte seu resultado na representação matricial. A TextNSP permite a identificação de n-gramas, especificando-se expressões regulares de formação destes, e ainda permite que se realizem testes de importância estatística dos n-gramas nos textos. De qualquer forma, a matriz resultante ainda é muito esparsa e contém n-gramas redundantes. Para minimizar esse problema está se desenvolvendo a ferramenta SelNGramas, com base no trabalho desenvolvido por Moura et al. (2009). À matriz filtrada aplicam-se os métodos da TaxTools (MORETTI et al., 2010), para criar um *cluster* hierárquico das notícias, e obter um resultado que torna fácil a análise e inferência das palavras-chaves e tópicos dos grupos de textos de notícias. A ferramenta TaxTools implementa agrupamentos hierárquicos de documentos (*bottom up*, utilizando similaridade de cosseno e o algoritmo "*average*"); cortes do agrupamento, com base em fusão e variância inter e intra grupo (MARCACINI et al., 2009); descritores estatisticamente mais significantes para cada grupo (MOURA, 2009); e, visualização dos resultados por uma *foldertree* associada a hiper-textos. O processo em desenvolvimento é ilustrado na Figura 1.

Um exemplo de resultado do processo pode ser observado na Figura 2. Neste, ainda incompleto, observa-se que os nós hachurados podem corresponder a assuntos tais como: cultura agrícola, álcool e combustível. Deve-se notar que o processo é retro-alimentável, como

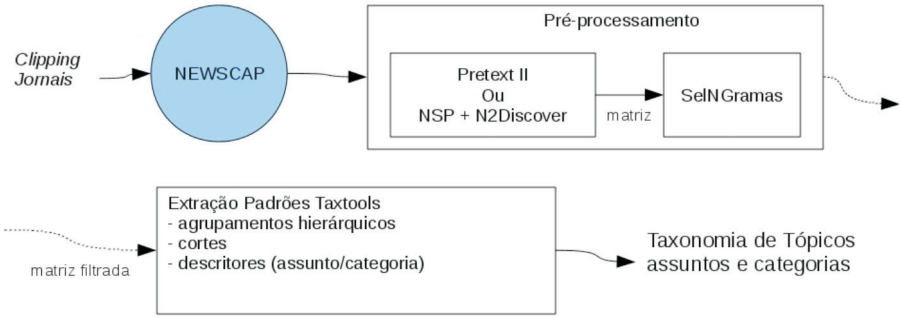

Figura 1. Processo em Desenvolvimento.

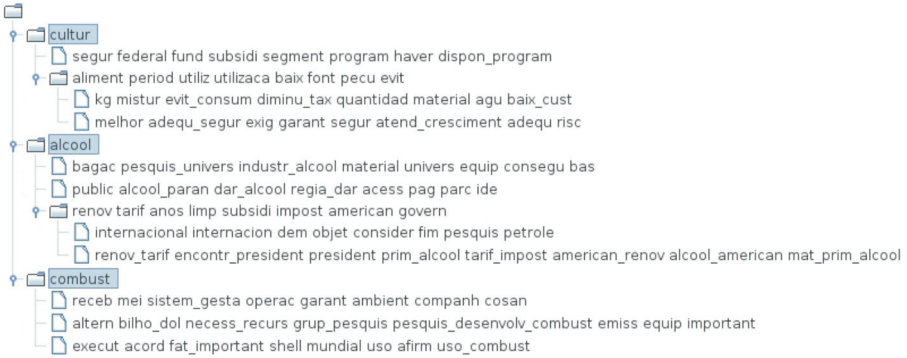

Figura 2. Exemplo de resultados com textos sobre cana-de-açúcar.

em qualquer análise de dados, logo, deverá ser repetido, sempre que sejam encontrados assuntos interessantes, até que o usuário esteja satisfeito com o resultado.

Concluímos que se está chegando ao primeiro objetivo do trabalho, a identificação dos tópicos em conjuntos de notícias. Os próximos passos consistem em tentar criar modelos para espacialização destas notícias para micro e macro regiões geográficas do Brasil.

Referências

BANERJEE, S.; PEDERSEN, T. The Design, implementation, and use of the Ngram statistics package. In: INTERNATIONAL CONFERENCE ON INTELLIGENT TEXT PROCESSING AND COMPUTATIONAL LINGUISTICS, 4., 2003, Mexico. **Proceedings**... Berlim: Springer-Verlag, 2003.

MARCACINI, R. M.; MOURA, M. F.; REZENDE, S. O. Uma abordagem para seleção de grupos significativos em agrupamento hierárquico de documentos. In: ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL, 2009, Bento Gonçalves. **Anais**... Porto Alegre: UFRGS, 2009. p. 1-16. 1 CD-ROM.

MORETTI, C. J. P.; PEIXOTO, B. M.; MOURA, M. F. **Tutorial da TaxTools**. Campinas, 2010. A ser editado pela Embrapa Informática Agropecuária. (Série. Documentos).

MOURA, M. F. **Contribuições para a construção de taxonomias de tópicos em domínios restritos utilizando aprendizado estatístico**. 2009. 137 f. Tese (Doutorado em Ciências de Computação e Matemática Computacional), Instituto de Ciências Matemáticas e de Computação, USP, São Carlos, SP, 2009.

SOARES, M. V. B.; PRATI, R. C.; MONARD, M. C. **PreText II**: descrição da reestruturação da ferramenta de pré-processamento de textos. São Carlos, SP: USP, ICMC, 2008. 45 p. (Relatório técnico, n. 333).