Evolução da eTMLib - Embrapa's Text Mining Library para pré-processamento de dados textuais

Vinícius Fernandes Dias[1]
Maria Fernanda Moura[2]
Sérgio Aparecido Braga da Cruz[2]
Roberto Hiroshi Higa[2]

Para realizar o pré-processamento de coleções de textos existem algumas ferramentas de domínio público que têm sido utilizadas por grupos de Pesquisa, Desenvolvimento e Inovação (PD&I) com bastante sucesso. A maioria dessas ferramentas permite trabalhar com textos em várias línguas, considerando-os *bag of words*; e aplicando filtros tais como: limitar os atributos àqueles entre uma frequência mínima e uma máxima, nos textos (CORTES, 1958) e/ou na coleção de textos (SALTON, 1975); eliminar uma lista pré-determinada de palavras (*stopwords*); remover inflexões de palavras; permitir desfazer-se de caracteres especiais, números, ou *tags* de linguagens de marcação. Porém, nem todas possuem todas essa funcionalidades integradas, alguns possuem formato próprio de saída, deixam a desejar quanto à presença de filtros específicos e, muitas vezes, não são facilmente integráveis a outros sistemas. Além disso, várias ferramentas são cedidas por universidades, como resultado de trabalhos de pesquisa, que, pela sua natureza, costumam não ser tão robustas e têm seu desenvolvimento descontinuado.

Devido a essas limitações, foi criada a Embrapa's *Text Mining Library* (eTMLib) (YAMADA et al., 2012), com o objetivo de ser uma solução flexível, facilmente expansível, facilmente integrável a outros sistemas, que

[1] Universidade Estadual de Campinas; Bolsista PIBIC/CNPq, v.fernandesdias@gmail.com

[2] Embrapa Informática Agropecuária, {maria-fernanda.moura, sergio.cruz, roberto.higa}@embrapa.br

permita cobrir as necessidades de pré-processamento de textos em várias línguas e, que mantenha uma interface com outras ferramentas (a princípio, de domínio público) capazes de realizar uma análise sintática em um texto e gerar índices de frequência (escolheu-se, inicialmente, a Lucene); além disso, essa solução deve permitir exportar os resultados em vários formatos utilizados por ferramentas de inferência, que também possam ser expandidos. Embora o objetivo seja um pouco ambicioso, pretende-se lançar mão de desenvolvimento colaborativo, num futuro próximo, disponibilizando a solução atual como uma biblioteca de classes, sob licença *General Public License* (GLP), em um repositório de domínio público (http://www.agrolivre.gov.br/). Porém, a versão existente da eTMLib ainda não está disponibilizada, encontra-se em testes e evolução.

O objetivo deste trabalho é testar, validar e evoluir a eTMLib, de modo que ela possa ser colocada em um repositório e facilmente tratada por outros desenvolvedores. Assim, vêm sendo feitos:

- testes exaustivos de funcionalidades: com a construção de massa de testes de regressão;
- padronização dos parâmetros de execução: os parâmetros atuais não são mnemônicos e algumas de suas combinações não estão sendo corretamente executadas. Os parâmetros devem permitir a execução da eTMLib em linha de comando, para que possa ser facilmente integrável a outras ferramentas, por meio de chamadas, ou a um *workflow*;
- padronização do processo de tratamento de vocabulário controlado.
- evolução do processo de filtro: o processo está ineficiente, o algoritmo de tratamento será trocado e melhor documentado;
- inclusão da geração de n-gramas variados: hoje ela só gera até trigramas;
- criar arquivo de *log* de execução: para guardar os parâmetros e permitir a repetição de um experimento ou a geração de conjunto de atributos similares.
- documentar casos de uso: para teste e elaboração de manual de uso.

Como resultado dessas atividades, está sendo conduzida uma reestruturação do código da eTMLib, com reflexos em sua estrutura de classes (Figura 1). Nessa nova estrutura, as classes principais refletem o próprio

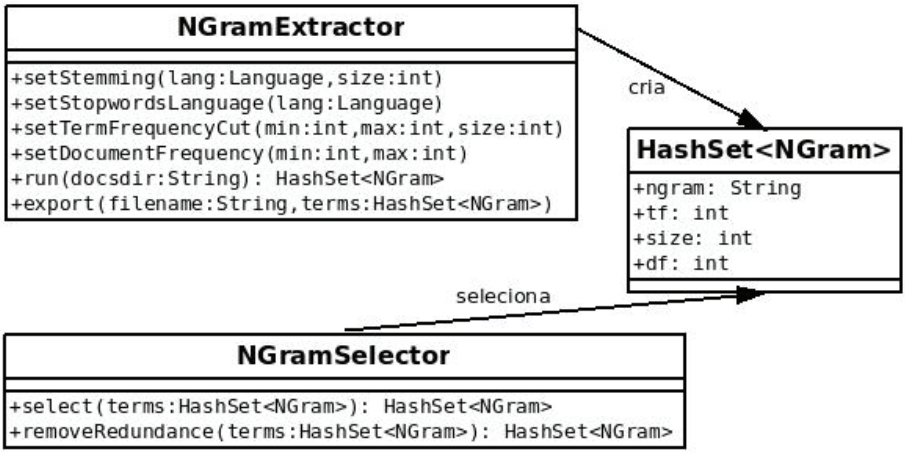

Figura 1. Classes principais da nova arquitetura da eTMLib.

processo de pré-processamento. A classe NgramGenerator gera um conjunto de n-gramas e suas correspondentes frequências; estes n--gramas, então, são selecionados por diferentes algoritmos pela classe NgramSelector. Ao final desse processo de reestrutação espera-se que a arquitetura da eTMLib seja muito mais simples e intuitiva. Além disso, para construção dos casos de testes, adotou-se o *framework* Junit, que auxilia na organização e documentação dos casos de teste, bem como permite que testes de regressão sejam rapidamente executados sempre que necessário.

Com essas evoluções espera-se tornar a eTMLib uma biblioteca mais robusta, mais simples de ser entendia e expandida.

Referências

LUHN, H. P. The automatic creation of literature abstracts. **IBM Journal of Research and Development**, Armonk, v. 2, n. 2, p. 159–165, 1958.

SALTON, G.; YANG, C. S.; YU, C. T. A theory of term importance in automatic text analysis. **Journal of the American Association Science**, Memphis, v. 1, n. 26, p. 33-44, 1975.

YAMADA, A. K.; MOURA, M. F.; CRUZ, S. A. B.; HIGA, R. H. Uma solução flexível para a etapa de pré-processamento em mineração de textos. In: CONGRESSO INTERINSTITUCIONAL DE INICIAÇÃO CIENTÍFICA, 6., 2012, Jaguariúna. **Anais**... Campinas: Embrapa: ITAL, 2012. p. 1-12. CIIC 2012. No 12611.