A mineração de textos como ferramenta de apoio a análise de artigos científicos

Lucas Dalla Lana Castoldi[1], Igor Yepes[2], Silvio Cesar Cazella[3]

[1]Bolsista de Iniciação Científica – Alunos do Curso Superior de Tecnologia em Sistemas para Internet - Instituto Federal Farroupilha – Frederico Westphalen.

[2]Curso Superior de Tecnologia em Sistemas para Internet - Instituto Federal Farroupilha – Campus de Frederico W estphalen.

[3]Faculdade de Educação – Universidade Federal do Rio Grande do Sul (CINTED/PPGIE/UFRGS) - Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

lucaskuia70@gmail.com, igor.yepes@iffarroupilha.edu.br, silvioc.ufcspa@gmail.com

Abstract. *This paper presents a use case of Voyant Tools text mining application on a corpus of scientific papers corpus. The initial objective is to verify the effectiveness of this tool for detecting relevant texts.*

Resumo. *Este trabalho apresenta um caso de uso da aplicação Voyant Tools para mineração de textos sobre um corpus composto por trabalhos científicos versando sobre uso de drones como ferramentas pedagógicas. O objetivo inicial é a verificação da eficácia dessa ferramenta para análise de relevância dos textos.*

1. Introdução

A vasta utilização de recursos tecnológicos em instituições de ensino vem despertando interesse crescente na busca por ferramentas que auxiliem no processo de aprendizado. Esse fato gerou um vasto campo multidisciplinar na área da computação, cujo foco está na pesquisa e no desenvolvimento de tais ferramentas.

Dentro desse paradigma surgem os drones, equipamentos robóticos (autônomos ou radiocontrolados) em evidência na atualidade, em geral, com uma divulgação negativa vinculada ao uso bélico e à invasão de privacidade. Contudo, pela sua versatilidade, muitas aplicações civis úteis têm sido desenvolvidas em paralelo, abordando primordialmente as áreas de segurança, indústria, agricultura de precisão, meio ambiente (monitoração e controle ambiental), fotografia aérea e filmagem (FERRI, 2010; VIEIRA, 2011).

Como é possível observar, em uma rápida busca na internet, uma área praticamente inexplorada, é a aplicação dos drones como ferramenta de cunho pedagógico. Os jovens vêm acompanhando o florescer dessa tecnologia com bastante entusiasmo, juntamente com os avanços em robótica e inteligência artificial. O fato de ter acesso a um equipamento desses em aula, por si só já torna a experiência do aprendizado muito mais interessante, o que facilita capturar a atenção desse público tão dinâmico e de fácil dispersão, características dos nativos digitais.

Contudo, apesar de não existir muito material versando sobre aplicação de drones na área da educação, há sim muito material disponível sobre uso de drones nas mais diversas áreas e muito mais ainda sobre robótica pedagógica (não envolvendo drones), o que torna a busca por material de apoio à pesquisa árdua e por vezes frustrante.

Assim, este trabalho visa verificar a possibilidade de utilizar técnicas de mineração de textos para selecionar material relevante sobre drones aplicados à educação, de forma direta ou indireta – por exemplo, são interessantes textos diretamente vinculados ao uso de drones em atividades pedagógicas, mas interessam também aspectos de hardware e plataformas de software utilizados para viabilizar esse uso.

1.1 Objetivo geral

Verificar a efetividade do uso da ferramenta de mineração de textos Voyant Tools para selecionar fontes de referência relevantes para trabalhos de pesquisa (neste caso, sobre aplicação de drones na educação).

1.2 Objetivos específicos

- Realizar uma breve pesquisa na Internet utilizando termos de busca relativos à robótica educativa e uso de drones, de forma a selecionar uma pequena amostra de textos (artigos, capítulos de livros etc.) com base no resumo;
- Estudar os recursos da aplicação web Voyant Tools para mineração de textos;
- Realizar o pré-processamento dos textos, de forma individual, de maneira a obter um corpus em formato padrão (txt), removendo partes desnecessárias;
- Realizar o processamento dos textos com a ferramenta de mineração;
- Ajustar a ferramenta na tentativa de obter uma boa visualização dos resultados;
- Analisar os resultados obtidos, efetuando ajustes adicionais para melhoria do processo.

2. Mineração de textos

A mineração de textos é uma tecnologia em ascensão que visa a extração de conhecimento em grandes conjuntos de documentos com textos não-estruturados ou semiestruturados. Consiste em descobrir, com base em grandes quantidades de texto, o conhecimento que pode não estar literalmente escrito nesses documentos, o que inclui a detecção de tendências, médias, desvios e dependências, entre outras tantas possibilidades.

A mineração de textos costuma ser confundida com a mineração de dados, contudo, a diferença básica é que na mineração de dados a busca é realizada em tabelas, planilhas ou bancos de dados, todos contendo informação estruturada, o que facilita a extração de informação.

Segundo Morais & Ambrósio (2007), a recuperação de informação, KDT (*Knowledge Discovery from Text*) e a mineração de textos apresentam alto grau de dependência com relação a processamento de linguagem natural (PLN). Nesse aspecto, a linguística computacional vem a ser o ramo que lida com a gramática e a linguística, no qual é desenvolvido o ferramental necessário para investigar textos e extrair deles informação sintática e gramaticalmente classificada.

Dessa forma, Sullivan (2001) define a mineração de textos como sendo o processo de compilar, organizar e analisar grandes conjuntos de documentos para auxiliar os analistas e tomadores de decisão na distribuição de informação, e para descobrir relações entre fatos relacionados que se dividem entre diferentes domínios de investigação.

Segundo Hearst (1999), a mineração de textos tem como objetivo a descoberta de informação e conhecimento previamente desconhecidos, que não constavam explicitamente em nenhum dos documentos analisados. De acordo com essa definição, a

mineração de textos seria um processo utilizado para descobrir novas informações ou conhecimento, e no qual a informação descoberta deveria ser desconhecida de antemão, inclusive pelos próprios autores dos documentos que tenham sido utilizados para análise.

Para poder realizar a descoberta de conhecimento em textos, devem ser seguidos alguns passos básicos. Assim, o corpus deve passar por um pré-processamento de forma a possibilitar que seja tratado computacionalmente, em seguida passa por um processo de mineração de textos e, finalmente, deve permitir a visualização dos resultados.

- **Pré-processamento:** ato de realizar operações ou transformações sobre o corpus textual de maneira a facilitar sua análise posterior por técnicas de mineração de textos. É um passo importante no processo, pois pode interferir no tipo de padrões que serão detectados. De forma geral, esta etapa requer bastante trabalho manual, removendo estruturas que o pesquisador considera irrelevantes ou que podem gerar padrões indesejáveis. Contudo, a maior parte desse trabalho é realizada de forma automatizada, incluindo tarefas como análise textual, classificação, técnicas de processamento de linguagem natural (*tokenization, lematization* etc.), técnicas de extração e de recuperação de informação (aquisição de padrões léxico-sintáticos, indexação, entre outros).
- **Mineração de textos:** etapa de descoberta. É a fase na qual as representações obtidas no pré-processamento são analisadas visando encontrar padrões de interesse ou novo conhecimento. Aqui entram técnicas de mineração de texto como a classificação, descoberta de associações e análise de tendências.
- **Visualização dos resultados:** etapa de exibição dos dados para o usuário da forma mais amigável possível. Em geral, os dados são apresentados não apenas de forma textual, mas com alguma representação gráfica para facilitar a compreensão de possíveis associações, visualização de padrões ou para destacar elementos de interesse detectados no processo de mineração.

2.1 Voyant Tools

Dentro do escopo das ferramentas existentes para trabalhar com mineração de textos, existe uma aplicação web denominada Voyant Tools (disponível em https://voyant-tools.org) cujas funcionalidades abrangem, entre outras tantas, contagem de palavras, criação de nuvem, concordâncias e detecção de tendências.

A Voyant Tools apresenta uma interface bastante amigável, apesar da ampla gama de ferramentas e opções de configuração que contém. Está incluído na aplicação um bom material explicativo (*help*), cobrindo de forma bem didática todas as funcionalidades disponibilizadas.

Essa ferramenta permite que o usuário extraia características de um corpus de forma rápida e precisa, auxiliando inclusive na compreensão dos processos e técnicas de mineração de textos.

Tal versatilidade e facilidade de uso, além da característica de ser uma aplicação web, não sendo necessária a instalação de nenhum aplicativo localmente, foram os principais motivos para a escolha da ferramenta.

Basta carregar o corpus (inserir os links, copiar o texto diretamente ou simplesmente realizar o upload dos arquivos desejados) que o Voyant Tools rapidamente processa o texto e já exibe uma interface inicial com um bom conjunto de informações sobre a análise realizada. Nesse ponto, cabe ao usuário refinar as configurações de cada ferramenta para obter o grau de visualização desejado.

2.2 Tema do corpus: drones na educação

Não resta dúvida de que este é efetivamente o início da era dos drones, bem como sente-se cada vez mais a inserção dos avanços da robótica e da inteligência artificial no cotidiano das pessoas. Tal como muitas novas tecnologias com elevado potencial, os drones podem parecer tanto assustadores quanto instigantes, principalmente para crianças e adolescentes. Cabe a professores e pesquisadores da área de educação encontrar usos adequados, pacíficos, éticos e criativos para essa tecnologia. Contudo, pouca ou nenhuma aplicação desses equipamentos extremamente interessantes tem sido vislumbrada na área pedagógica, e quando existe, limita-se ao uso de drones para filmagem de atividades escolares ou somente uso dessas filmagens ou fotografias em determinados aspectos de uma disciplina.

Os drones estão invadindo a vida, e a privacidade das pessoas. Seja de forma positiva ou negativa, é um processo sem retorno e, por enquanto, carente de legislação específica. É necessário aproveitar essa tecnologia de forma realmente útil, gerando não apenas conhecimento na área STEM, mas também propiciando reflexão e consciência crítica quanto ao uso adequado de tecnologias emergentes.

No processo evolutivo da educação, um dos momentos de maior importância ocorreu no século XIX quando diferentes vertentes propuseram a mudança do paradigma da educação passiva, predominante na época. Já no século XX, destacam-se a teoria construtivista do psicólogo suíço Jean Piaget e a pedagogia do construcionismo desenvolvida pelo matemático Seymour Papert. Piaget afirmar que o conhecimento não se transmite, mas sim se constrói, ou seja, é criado ativamente na mente do aprendiz. O construcionismo segue a mesma linha, porém, acrescenta que para alcançar isso é preciso que o indivíduo construa algo tangível, um elemento fora da sua mente, que além de tudo, tenha significado pessoal para ele. Essa última teoria é na que se baseiam muitos dos principais desenvolvimentos de robótica pedagógica (GONZÁLEZ & JIMÉNEZ, 2009).

Assim, a robótica pedagógica pode ser definida como a área do conhecimento que utiliza os conceitos das engenharias e demais ciências no processo de concepção, construção, automação e controle de dispositivos robóticos com propósitos educacionais (ABREU & BASTOS, 2015).

Dessa forma, deve aqui ser analisado, com uso da aplicação Voyant Tools, um corpus formado por um conjunto de estudos sobre aplicação de drones como ferramenta pedagógica, os quais foram coletados após rápida pesquisa na Internet, e selecionados com base apenas no conteúdo dos respectivos resumos que, a primeira vista, atendiam todos os requisitos necessários para serem uma boa referência para a pesquisa em questão.

3. Materiais e métodos

Foram utilizados dez textos científicos coletados mediante pesquisa no Google Acadêmico, após uma rápida avaliação dos respectivos resumos.

Esses textos foram processados manualmente, removendo formulas, tabelas, cabeçalhos, rodapés, dados informativos dos autores e suas vinculações acadêmicas, editoras, agradecimentos e referências bibliográficas, gerando um corpus pré-processado com os dez textos em arquivos individuais no formato texto (.txt).

Foi então realizado o upload do corpus na ferramenta Voyant Tools, momento a partir do qual já foi gerada, em poucos segundos (menos de dez), uma visualização inicial padrão, servindo como base inicial para análise.

A seguir, foi realizada uma análise dos dados, tentando explorar os recursos disponibilizados pela aplicação de mineração de textos, na tentativa de identificar "peculiaridades" com relação ao corpus em estudo (informações pertinentes ou adversas

que tenham sido evidenciadas pela ferramenta), bem como verificar se a análise com base nos resumos foi eficiente ou não.

Como se trata de um corpus pequeno (dez artigos) a validação foi realizada mediante análise (leitura) de cada um dos textos utilizados, de forma a verificar os resultados obtidos com o uso da ferramenta.

Não foi utilizado nenhum equipamento especialmente configurado para realização do experimento. Toda a atividade foi executada em um MacBook Pro, com 8Gb de memória RAM e processador i5 de 2.4Ghz.

4. Experimento

Após pesquisa na Internet utilizando o Google Acadêmico, foi verificado o baixo índice de textos de cunho científico abordando especificamente o uso de drones como ferramenta pedagógica, o que já era esperado. A pesquisa foi realizada com o seguinte conjunto de termos:

- drone OR uav "STEM education" filetype:pdf (54 resultados)
- drone OR uav "educational robotics" filetype:pdf (21 resultados)
- drone OR uav "robotic kit" filetype:pdf (3 resultados)

Com base nos resultados obtidos na busca, foi realizada uma rápida análise do resumo dos textos "aparentemente" mais relevantes, o que permitiu selecionar uma amostra de 10 trabalhos incluindo artigos e capítulos de livros.

Os arquivos selecionados foram então exportados para o formato texto (.txt) gerando dez arquivos individuais (corpus), sobre os quais foi realizado um pré-processamento manual, removendo os componentes textuais que indicavam os autores, instituições, cabeçalho, rodapé e referencial bibliográfico, obtendo dessa forma o material que serviria como base para este trabalho.

Na sequência, foi efetuado o upload do corpus para a ferramenta, o que possibilitou o início das analises, partindo da configuração inicial da ferramenta.

Na visualização padrão da ferramenta, já é possível obter muitas informações sobre o corpus em análise. As ferramentas são interativas e ao interagir com algumas delas, como por exemplo a nuvem de palavras, tal interação é refletida em outras ferramentas de visualização da interface.

Inicialmente são apresentadas cinco formas de visualização, mas cada uma delas possui algumas (ou muitas) configurações e formas de visualização diferentes, atendendo a praticamente qualquer necessidade de análise requerida pelo usuário com relação a contexto, sumarização, tendências e muitas outras, de forma individual para cada texto do corpus, ou globalmente. Além disso, cada um desses cinco espaços da interface web pode ser personalizada pelo usuário, de forma a apresentar a ferramenta desejada e na dimensão que mais for conveniente (desde que respeitada a área de trabalho do aplicativo e suas restrições).

Uma forma interessante de visualização é o gráfico de linhas de bolhas que permite ver a frequência e distribuição dos termos (Fig. 1), mediante o qual ficou evidente a desconexão de dois textos analisados (textos 05 e 07) em relação ao estudo desejado sobre drones como ferramentas pedagógicas. Ou seja, apesar dos respectivos resumos atenderem totalmente os requisitos desejados com relação ao conteúdo esperado, realizando uma leitura mais completa desses dois trabalhos, verificou-se que realmente não se enquadram como boas referências bibliográficas, tendo conteúdo pobre em relação ao tema específico desejado.

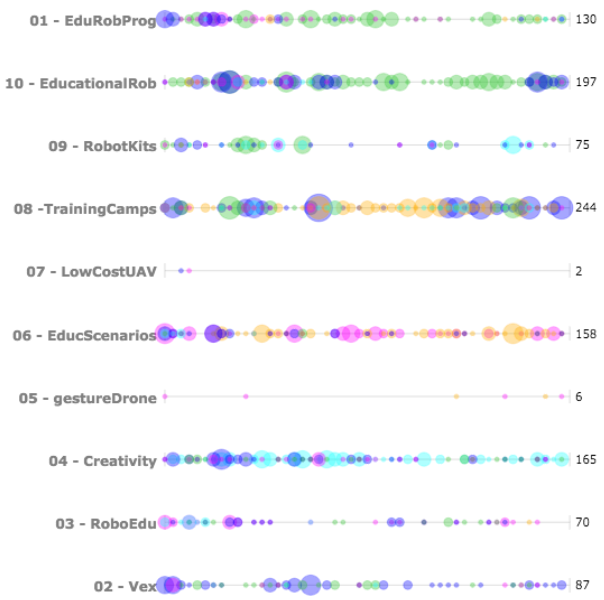

Fig. 1 – Bubblelines graphic

A nuvem de palavras (Fig. 2 - a) ressaltou uma característica dos textos que despertou a atenção de forma momentânea, mas ficou evidente após uma análise do corpus global. A palavra "drone", que deveria ser um dos temas centrais do estudo, não aparece na nuvem. Foi sendo gradualmente aumentado o número de palavras da nuvem, mas mesmo com valores acima de duzentas palavras, o termo continuava ausente. No entanto, seu sinônimo (UAV – *Unmanned Aerial Vehicle*) já aparecia com apenas 50 palavras na nuvem.

Analisando a frequência dos termos no corpus, foi detectado que a palavra "drone" estava na posição 721, com apenas sete aparições no corpus (Fig. 2 - b).

Ao verificar os documentos do corpus, foi fácil constatar que o termo drone aparece somente em um dos trabalhos (texto 05), como pode ser observado no gráfico de tendências (Fig. 3), o qual teve de ser personalizado para exibir apenas os termos desejados.

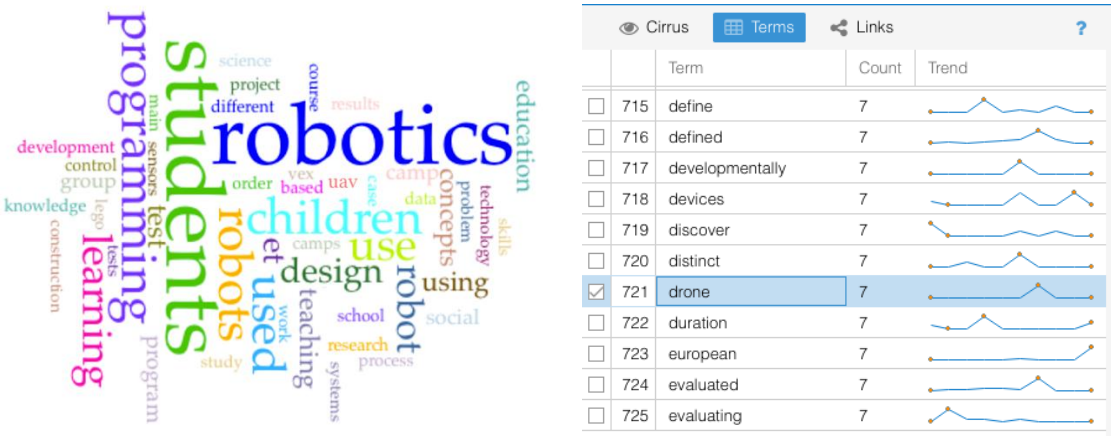

Fig. 2 – (a) Cirrus graphic – (b) Frequência de cada termo no corpus

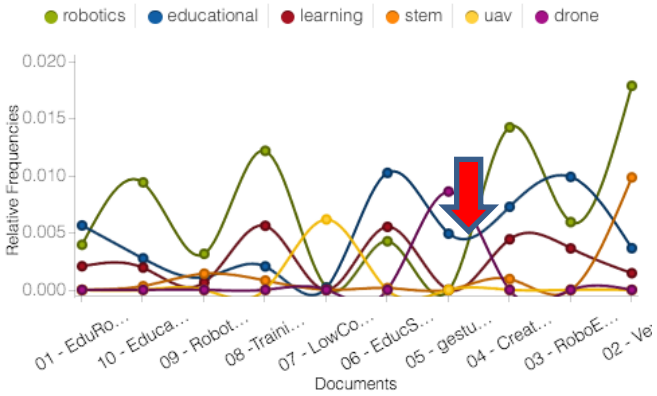

Fig. 3 – Trends graphic

Essa ausência se deve ao fato do termo drone estar comumente associado a equipamentos militares, o que faz com que ocorra um certo preconceito sobre ele. Assim, os pesquisadores preferem utilizar o termo UAV, ainda não maculado pela mídia.

Há ainda muitos recursos que podem ser explorados de acordo com a conveniência do usuário, com representação das conexões entre os termos por meio de grafos, mandalas, análise de contexto dos termos desejados, entre outros.

5. Conclusão e trabalhos futuros

Finalmente, cabe destacar que a recuperação de informação com mineração de texto busca estabelecer os mecanismos para atender as necessidades de informação do usuário, contudo, não se espera dela uma resposta a um questionamento preciso, mas sim que consiga fazer emergir dos dados, informações aparentemente ocultas e quiçá, relevantes. Em outras palavras, para Morais & Ambrósio (2007), ao fazer uso de técnicas de mineração de textos, o usuário não solicita exatamente uma busca, mas sim uma analise de um texto. Contudo, este não recupera o conhecimento em si. Assim, é importante que o resultado da consulta seja analisado e contextualizado para posterior descoberta de conhecimento.

Como resultado dos testes aqui realizados, pode-se verificar a eficiência da mineração de textos para análise de um corpus bastante reduzido. Neste caso específico, o usuário selecionou dez textos considerados importantes mediante leitura apenas do resumo e, após o processamento pela Voyant Tools, foi verificado que somente 80% desse material era efetivamente relevante. Certamente seu poderio seria muito mais evidente ao trabalhar com corpus de grande porte, possuindo centenas ou milhares de documentos, o que reduziria drasticamente o “trabalho braçal" de leitura e análise textual por parte do usuário.

Possivelmente, num futuro próximo, com os devidos e esperados avanços no PLN, as ferramentas de mineração de textos como a aqui avaliada aumentem seu poderio de análise de forma exponencial, conseguindo um efeito mais contundente no apoio aos processos de tomada de decisão e na análise de grandes quantidades de documentos com conteúdo não estruturado.

A Voyant Tools demonstrou sua versatilidade, agrupando diversas funcionalidades que não são comumente encontradas em conjunto em outras ferramentas existentes de mineração de texto, entretanto pode melhorar possibilitando a seleção ou até mesmo a implementação de novos algoritmos de aprendizado de máquina por parte do usuário, como já ocorre em outras ferramentas do tipo.

Como trabalho futuro, estuda-se a possibilidade de criar um corpus maior, uma vez que a quantidade de artigos específicos se mostrou bastante pequena, possibilitando expandir os termos de busca para constituir um conjunto de textos relativamente grande, difícil de analisar e avaliar "braçalmente", mas que podem ser minerados pela Voyant Tools de maneira a reduzir o material a uma pequena parcela, provavelmente de alta relevância para a pesquisa.

Referências

ABREU, João V. V. d'; BASTOS, Bruno L. Robótica Pedagógica e Currículo do Ensino Fundamental: Atuação em uma Escola Municipal do Projeto UCA. **Revista Brasileira de Informática na Educação**, Volume 23, Número 3, 2015.

FERRI, Andreu B. **Desarrollo de una plataforma de tiempo real para la implementación de algoritmos de control multivariables**: Ampliación al control de orientación de vehículos aéreos. Dissertação de Mestrado em e Automação e Informática industrial. Valência: Universidad de Politécnica de Valencia, 2010.

GONZÁLEZ, Juan J.; JIMÉNEZ, Jovani A. La robótica como herramienta para la educación en ciencias e ingeniería. **Revista Iberoamericana de Informática Educativa**. nº 10, Jul - Dez 2009, p.31-36. IE Comunicaciones: Espanha, 2009.

HEARST, Marti. **Untangling text data mining**. In: *Proceedings of ACL'99: the 37th annual meeting of the Association For Computational Linguistics*, junho, 1999. Acessado em: 25/10/16. *http://www.sims.berkeley.edu/~hearst/papers/acl99/acl99-tdm.html*

MORAIS, Edison Andrade Martins; AMBROSIO, Ana Paula L. **Mineração de Textos**. Relatório Técnico. Universidade Federal de Goiás, 2007.

SULLIVAN, Dan. ***Document warehousing and text mining***. New York [etc.]: Wiley Computer Publishing, 2001, xviii, 542 p.

VIEIRA, José C. S. **Plataforma Móvel Aérea QuadRotor**. Dissertação de Mestrado. Universidade do Minho - Escola de Engenharia. Portugal: UMinho, 2011.