



UMA FERRAMENTA VISUAL PARA ANÁLISE DE PRODUÇÕES TEXTUAIS DOS ESTUDANTES

Gustavo Lovato – CCET/UCS – glovato@ucs.br

Ricardo Dacol Gil – CCET/UCS – rdgil@ucs.br

Ana Maria Coulon Grisa – CCET/UCS – amcgrisa@ucs.br

Carine G. Webber – CCET/UCS - cgwebber@ucs.br

Elisa Boff – CCET/UCS – eboff@ucs.br

Maria de Fátima Webber do Prado Lima – CCET/UCS – mfwplima@ucs.br

Resumo. Tradicionalmente a leitura e interpretação de textos é uma habilidade que humanos realizam de forma muito superior às máquinas. Entretanto, quando se trata de buscar padrões e associações em textos, os computadores podem realizar a parte mais exaustiva e longa desta tarefa. Técnicas de aprendizado de máquina podem auxiliar professores na avaliação de produções textuais, destacando pontos fortes e identificando dificuldades. Desta forma os professores podem oferecer um feedback mais apropriado aos estudantes por meio de reflexões pertinentes às suas aprendizagens. Este artigo descreve a implementação e os resultados obtidos com o desenvolvimento de um software de apoio a análise textual. Para validar o software foram utilizadas produções textuais de estudantes da disciplina de Química Orgânica de uma Universidade (nome omitido).

Palavras chaves: Aprendizado de máquina, visualização de dados, análise textual.

A VISUAL TOOL FOR STUDENT'S TEXTUAL ANALYSIS

Abstract. Traditionally, text reading and interpretation are skills where humans perform better than machines. However, when it comes to look for patterns and associations in texts, computers could at least produce an important part of this task. Nowadays learning machine techniques can assist teachers in analyzing texts, highlighting strengths and identifying difficulties. So, teachers can offer a more appropriated feedback to students through relevant reflections on their learning. This article describes implementation and results obtained with the development of a textual analysis software. To evaluate the software, experiments at Organic Chemistry class have been done.

Keywords: Machine learning, data visualization, textual analysis.

1. Introdução

Atualmente acredita-se que as habilidades relacionadas à produção e interpretação textual são inerentemente humanas. Contudo, graças aos avanços na área da computação cognitiva, as máquinas poderão desenvolvê-las também (IBM, 2016; Microsoft 2016). Fazendo uso das tecnologias de aprendizado de máquina, em conjunto com mecanismos para representação de conhecimento, processamento de linguagem natural e raciocínio automático, diversos sistemas começam a surgir (e.g. Watson (IBM, 2016)). Um desafio desta área consiste em prover interpretações para grandes volumes

de textos de forma automática e compreensível ao ser humano. Isto implica em analisar e sumarizar texto, além de fazer uso de representações visuais apropriadas e de forma correta.

No contexto das atividades docentes, tarefas de leitura e interpretação de produções textuais de estudantes fazem parte da rotina de grande parte dos professores. Em algumas disciplinas as avaliações são feitas por meio da resolução de problemas ou provas objetivas. Poucos registros são utilizados para avaliar a aprendizagem através de produções textuais. Além de incentivar a leitura e a escrita por parte dos estudantes, percebe-se que as produções textuais poderiam complementar a avaliação que o docente realiza sobre um estudante (ou uma turma) pois constituem-se de evidências da construção de conhecimento.

As técnicas de aprendizado de máquina podem ser aplicadas a fim de organizar, categorizar ou descobrir informações contidas em bases de dados de forma ágil e automática. Tecnologias disponíveis permitem que o aprendizado seja feito a partir de dados textuais (FELDMAN e SANGER, 2006). Isso significa que elas possibilitam a identificação de padrões e a posterior classificação de dados de forma automática e precisa. Para fins de comparações e análises, os dados gerados podem ser submetidos a técnicas de visualização que auxiliam na elaboração de *insights* (Webber, Cini e Lima, 2013).

Neste trabalho busca-se investigar a aplicação de técnicas de aprendizado de máquina em cenários de análise textual de produções elaboradas por estudantes. Assume-se que qualquer análise que pretenda auxiliar o professor deva fornecer visões sobre a evolução dos estudantes individualmente e da turma como um todo. Embora existam trabalhos relacionadas a este, destaca-se aqui o uso de ferramentas de visualização para apoio as atividades docentes. A questão de pesquisa que norteia este trabalho diz respeito a análise da adequação dos recursos de visualização de dados textuais, usualmente empregados em vastos domínios, para o perfil do docente a quem ela deve servir.

Neste cenário, o presente artigo descreve o desenvolvimento de um software de apoio a análise e visualização textual a partir de técnicas de aprendizado automático. Ele está organizado em 5 seções. A seção 2 introduz os principais conceitos associados ao aprendizado de máquina em textos. A seção 3 apresenta a ferramenta Análise de Produções Textuais desenvolvida. A seção 4 ilustra os resultados da utilização da ferramenta em dois experimentos realizados em sala de aula. A seção 5 apresenta uma análise preliminar do docente que acompanhou o experimento e as conclusões preliminares deste estudo.

2. Aprendizagem de Máquina e Visualização de Textos

Dentre as técnicas de aprendizado de máquina, denomina-se mineração de texto (MT) a que busca extrair padrões não triviais a partir de dados textuais não estruturados (TAN, 1999). Segundo Aranha (2007), a MT pode ser dividida em cinco etapas. A primeira etapa é a coleta de dados, onde são coletados os documentos que formarão a base de textos a serem analisados. O pré-processamento, a segunda etapa do processo, tem como objetivo transformar estes textos em informações mais estruturadas, efetuando a escolha dos termos mais importantes do texto dentro do objetivo que se deseja atingir. Nesta etapa efetua-se a eliminação de *stopwords*, que são os termos que não possuem relevância para o texto ou que sozinhas nada significam (MANNING et al., 2008). Na etapa de indexação, as variações morfológicas de uma palavra podem ser eliminadas. Elas são eliminadas identificando o radical de uma palavra. Retiram-se os prefixos e os sufixos e os radicais resultantes são incluídos ao índice (processo de *stemming*). Na etapa de extração do conhecimento, a relevância dos termos e o objetivo a ser

alcançado são definidos de acordo com a área que está sendo minerada. A última etapa envolve a aplicação de técnicas de análise dos resultados. Esta etapa não é realizada computacionalmente, mas por um ser humano com conhecimento.

A análise semântica latente (LSA) é um método utilizado para extrair as relações entre documentos e os termos neles contidos, por meio da produção de um conjunto de conceitos latentes relacionados com os documentos e os termos. Este método assume que as palavras semanticamente relacionadas tendem a coocorrer nos textos. Este tipo de análise emprega técnicas que avaliam a sequência dos termos no contexto dos textos, no sentido de identificar qual a sua função. Sua utilização justifica-se principalmente pela melhoria da qualidade dos resultados do processo de mineração de textos, especialmente se for incrementado por Processamento Linguístico. A técnica de LSA explora a relação existente entre termos e os textos nos quais eles aparecem para construir um espaço vetorial onde similaridades de significado podem ser estabelecidas. Este novo espaço é conhecido como espaço-conceito ou espaço-semântico (DEERWESTER, 1990), sendo que a proximidade entre significados é proporcional ao ângulo entre vetores neste espaço (BERRY, DRMAC e JESSUP, 1999). A LSA consiste na construção de uma matriz A que informa a coocorrência de termos e documentos. Após, a matriz é algebricamente decomposta seguindo a decomposição em valores singulares (SVD) como forma de aproximação da matriz A por combinação linear (BERRY, DRMAC e JESSUP, 1999).

Em conjunto com técnicas de aprendizado de máquina é importante oferecer recursos de visualização de dados. A área de visualização leva em consideração a construção de representações visuais de dados abstratos de forma a facilitar o seu entendimento e a descoberta de novas informações (FREITAS et al., 2001). Os dados coletados são transformados e organizados em tabelas relacionais. Estas tabelas são transformadas em estruturas visuais que representam os dados empregando representações gráficas, símbolos e padrões. Operações são realizadas nestas estruturas visuais a fim de mostrar novas informações sobre este conjunto de dados através de manipulações geométricas, indicações de subconjuntos/regiões de interesse ou mudança do ponto de observação (CARD et al., 1999). Técnicas são utilizadas na compreensão e na busca por informações provenientes do que está sendo visualizado. As técnicas podem ser classificadas com base em três critérios: os dados a serem visualizados, a técnica de visualização empregada e a técnica de interação utilizada para manipular as representações visuais (KEIM, 2002).

De maneira muito apropriada Judelman (2004) definiu três categorias de visualização: complexidade, contexto e dinâmica. A modalidade de complexidade compreende a visualização de árvores e hierarquias, mostrando topologias ou caminhos. A categoria de contextos exhibe relacionamentos semânticos, mostrando os objetos espacialmente organizados de acordo com a similaridade ou importância. A visualização de aspectos dinâmicos envolve diagramas de processos exibindo contexto espaço temporal e usualmente um eixo representando o tempo. Para a visualização e a comparação entre textos (objeto deste estudo), a categoria que melhor se adapta é a de contexto, com base nas propriedades semânticas a serem analisadas.

3. Ferramenta de Análise de Produções Textuais

A ferramenta de Análise de Produções Textuais (APT) foi desenvolvida em linguagem C#, integrando visualizações implementadas em *Java Script* com a ferramenta *D3 Data-Driven Documents* (BOSTOCK, 2013). A comunicação entre a aplicação e as visualizações *web* é feita por meio de arquivos no formato *JSON*. Estes arquivos são salvos no diretório de publicação do servidor *web* e são lidos pelas aplicações desenvolvidas para a camada *web*.

O software APT (Figura 1) opera em duas etapas. A primeira etapa é a de treinamento, onde o sistema realiza o aprendizado a partir de textos e materiais didáticos do professor. A segunda etapa consiste na análise dos textos redigidos pelos estudantes. Os resultados do processo são transformados e ficam disponíveis por meio de diversas formas de visualizações gráficas. A interface possui duas abas. Na aba de treinamento cabe ao professor informar o seu corpus de textos. Na segunda aba, o professor acessa e manipula recursos de visualização. A etapa de treinamento é fundamental para o sucesso das análises textuais. O professor deve selecionar textos e materiais didáticos para serem utilizados (corpus do professor). O idioma pode ser português ou inglês, mas apenas um por vez.

O objetivo da segunda etapa é analisar as produções textuais dos estudantes, gerando visualizações para a análise por parte do professor. Inicialmente devem ser informados o diretório onde as produções textuais dos estudantes estão localizadas e um valor K para determinar a precisão dos resultados do algoritmo LSA. A escolha do valor K é muito importante para obter resultados viáveis. Um valor K muito pequeno, como 2, torna o desempenho do cálculo melhor mas resulta num cálculo impreciso devido à falta de informação. Um valor K alto demais, introduz muito ruído no cálculo, prejudicando também a precisão do resultado final.

Ao clicar no botão “Iniciar Diagnóstico”, o programa importará todos os arquivos da origem, apresentando-os como “Atividades Importadas”. O sistema notifica quais artigos importou, informando a atividade, a turma e o estudante. Na tela da caixa de “Seleção”, ficam disponíveis as visualizações: uma atividade de um estudante (1-1), uma atividade de uma turma (1-n), várias atividades dos estudantes (n-n) e várias atividades de um estudante (n-1).

Atividade	Turma	Estudante	NotaLSA
CRIATIVIDADE E INOVAÇÃO	1	ESTUDANTE 01	0.649363907079919
CRIATIVIDADE E INOVAÇÃO	1	ESTUDANTE 02	0.728142071903635
CRIATIVIDADE E INOVAÇÃO	1	ESTUDANTE 03	0.262729504975261
CRIATIVIDADE E INOVAÇÃO	1	ESTUDANTE 04	0.229260997161591
CRIATIVIDADE E INOVAÇÃO	1	ESTUDANTE 05	0.631892057046383
CRIATIVIDADE E INOVAÇÃO	1	ESTUDANTE 06	0.893835331187715
CRIATIVIDADE E INOVAÇÃO	1	ESTUDANTE 07	0.314204081714883
CRIATIVIDADE E INOVAÇÃO	1	ESTUDANTE 08	0.288933509657257
CRIATIVIDADE E INOVAÇÃO	1	ESTUDANTE 09	0.257503167841342
CRIATIVIDADE E INOVAÇÃO	1	ESTUDANTE 10	0.146985821274906

Figura 1 - Tela de análise da ferramenta Análise de Produções Textuais.

O sistema calcula o limiar do documento, que indica a sua similaridade em relação ao conteúdo analisado no corpus do professor. O algoritmo LSA foi utilizado pelo software para realizar a análise dos textos. A coluna “NotaLSA” (Figura 1) demonstra resumidamente o resultado da análise. O algoritmo foi implementado em 11 etapas:

1. Inicialização das HashTables onde são indexados termos, documentos e a relação termo/documento. É realizada a leitura do texto de treinamento, contando a frequência de cada termo.
2. Indexação do documento de treinamento: os StopWords são filtrados, os termos restantes são inseridos na HashTable, é gerada a relação Termo/Documento

- junto à frequência de cada termo e o documento de treinamento é inserido.
3. Similar a segunda etapa, porém indexa-se os textos a serem analisados.
 4. Criação da matriz Termo/Documento, onde cada linha corresponde a um termo, e cada coluna corresponde a um documento. Cada célula corresponde a frequência Termo/Documento (Peso Local) gerada anteriormente.
 5. Cálculo do peso global de cada termo, isto é, quantos documentos dentro da coleção contêm a palavra. A função utilizada foi a frequência inversa de documentos.
 6. Cálculo dos fatores de normalização dos termos de cada documento, utilizando o peso global do termo e o peso local do termo no respectivo documento de todos os termos.
 7. Criação da matriz Termo/Documento Ponderada, onde o valor de cada célula considera o peso local de cada termo, o peso global de cada termo e o fator de normalização de cada documento.
 8. Cálculo do SVD, permitindo a reconstrução da matriz Termo/Documento Ponderada original.
 9. Preparação das matrizes reduzidas, utilizadas no cálculo de similaridade dos documentos. As matrizes reduzidas são calculadas baseadas no valor K (definido na interface gráfica).
 10. Armazenamento em disco de diversas variáveis necessárias para o cálculo de similaridade de documentos, diminuindo o uso de memória durante o processamento, uma vez que o conjunto de matrizes pode crescer rapidamente.
 11. Redução das matrizes, através de operações de transposição e multiplicação de matrizes.

A figura 2 ilustra uma produção textual de um estudante, destacando o cabeçalho do arquivo. O cabeçalho deve conter as informações sobre a turma, nome do estudante e indicador de atividade. Tais informações são necessárias para que o sistema registre as turmas, estudantes e atividades para gerar visualizações corretas.

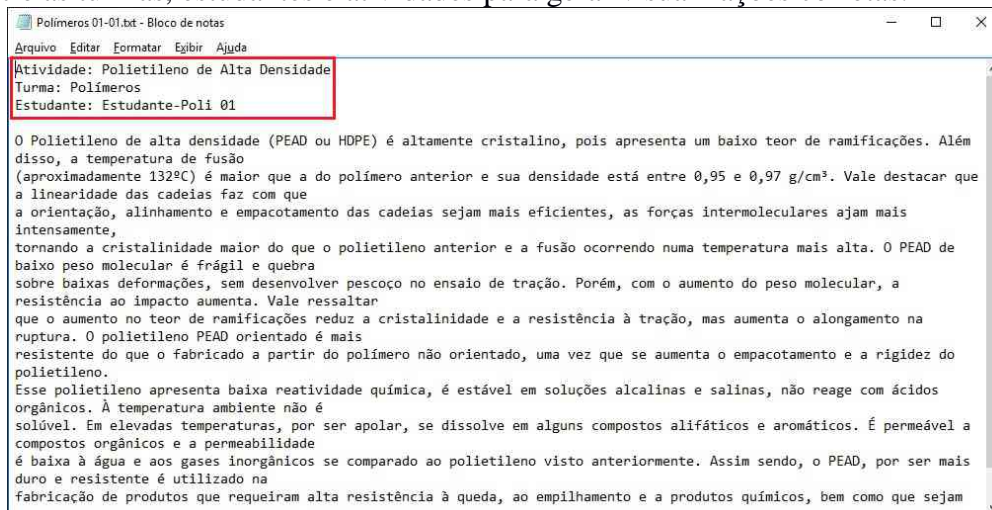


Figura 2 - Exemplo de produção textual de um estudante da turma de Química Orgânica

O software APT possibilita sete visualizações distintas, implementadas utilizando a tecnologia *D3 Data-Driven Documents* (BOSTOCK, 2013). As classes que geram as visualizações são responsáveis por fazer a chamada da visualização que abrirá no navegador padrão da máquina. As visualizações disponibilizadas são: Bar Chart (Gráfico de Barras), BubbleChart (Gráfico de Bolhas), CollapsibleTree (Árvore Flexível), RotatingCluster (Agrupamento Rotativo), Treemap (Mapa Árvore), Word Cloud (Nuvem de Palavras) e LSA ScatterPlot. O nome das visualizações foi mantido em

inglês na interface do software por serem nomes padronizados.

4. Resultados Obtidos

A amostra de estudantes alvo do estudo foi constituída durante uma atividade realizada na disciplina de Química Orgânica do curso de Engenharia de Materiais e Tecnólogo em Polímeros. A atividade teve o objetivo de expor as ferramentas de visualização à análise de um professor que não fosse da área da informática. Esta atividade foi iniciada com a solicitação aos estudantes da leitura do artigo Polietileno: Principais Tipos, Propriedades e Aplicações (COUTINHO, MELLO e MARIA, 2003). O artigo descreve as principais características e aplicações do polietileno de baixa densidade, de alta densidade, linear de baixa densidade, ultra-alto peso molecular e de ultrabaixa densidade, correlacionando a sua estrutura química com suas propriedades. A ferramenta APT foi aplicada nos textos de sete estudantes com relação ao estabelecimento da relação entre estrutura e propriedade dos vários tipos de polietileno que justifiquem suas aplicações. As produções textuais dos estudantes foram separadas para facilitar as análises pelo tipo de polietileno, gerando uma produção textual para cada tipo de polietileno (cinco tipos) e para cada estudante (sete estudantes). As visualizações geradas foram disponibilizadas ao professor, sendo brevemente descritas a seguir. Na tela do computador as imagens são clicáveis e se modificam conforme a interação do professor avaliador. Cabe destacar que as imagens são exibidas em tamanho grande para o professor e a resolução na tela é consideravelmente melhor do que na versão que consta neste artigo.

A primeira visualização é a Collapsible Tree (Figura 3). As ramificações iniciaram com o tema polímero e tipos de polímeros e após os estudantes da turma que realizaram a atividade. Clicando em cada estudante têm-se uma listagem das palavras mais relevantes utilizadas na sua produção textual. No exemplo selecionou-se “polietileno de alta densidade” e “estudante-poli-3” (padrão de identificador utilizado para os estudantes). Ao descrever as propriedades do polietileno de alta densidade o estudante utilizou termos importantes como: maior, molecular, ramificações, rigidez, tipo, usado. O professor avaliou que o texto do aluno apresentava palavras-chave relevantes, e que a própria visualização contribuiu para avaliar quais as propriedades que o estudante compreendeu e soube inserir em seu texto.

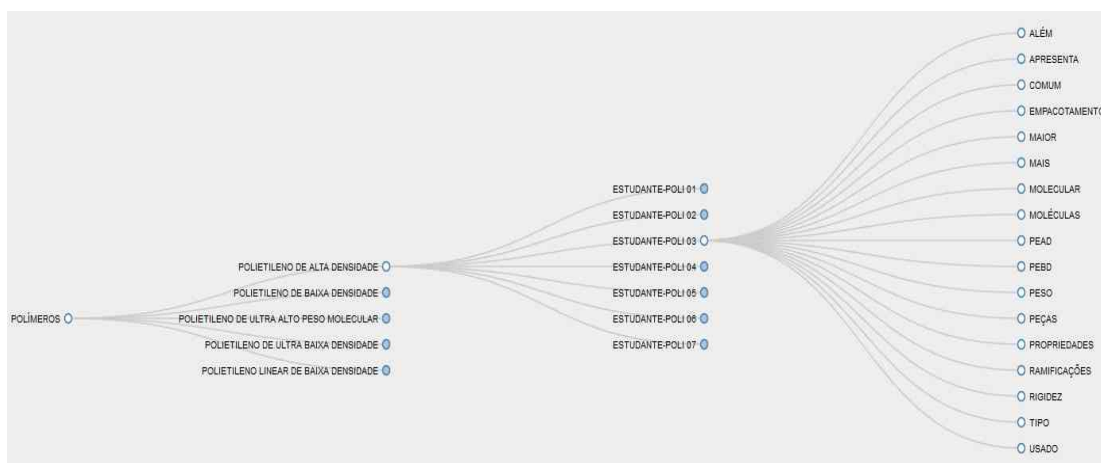


Figura 3 - Collapsible Tree (Árvore Flexível) dos estudantes em todas as atividades.

A Figura 4 apresenta o Bubble Chart com a análise de todas as atividades de um estudante. Nela os macrotemas são representados pela área da bolha e o texto do centro da bolha identifica o macrotema. Os polímeros são caracterizados por cores, onde cada cor representa um tipo de polímero. Ao posicionar o cursor sobre uma das bolhas é

apresentada a palavra completa acompanhada de sua pontuação. A visualização distribui os polímeros com maior volume de termos em macrotemas na região central e na periferia estão os com menos termos. Nesta visualização é possível perceber a representatividade dos termos mais significativos comparando o raio das bolhas. Enquanto o gráfico da Figura 3 facilita a visualização das tarefas de cada aluno individualmente, o gráfico de bolhas possibilita que se visualize todos os alunos em uma tarefa ou todas as tarefas de um aluno. Já a visualização de todas as tarefas de todos os alunos se torna mais complexa. O gráfico de bolhas foi avaliado pelo professor como um bom recurso para avaliar a progressão do estudante em termos do vocabulário empregado, bem como perceber se as propriedades esperadas constam do texto.

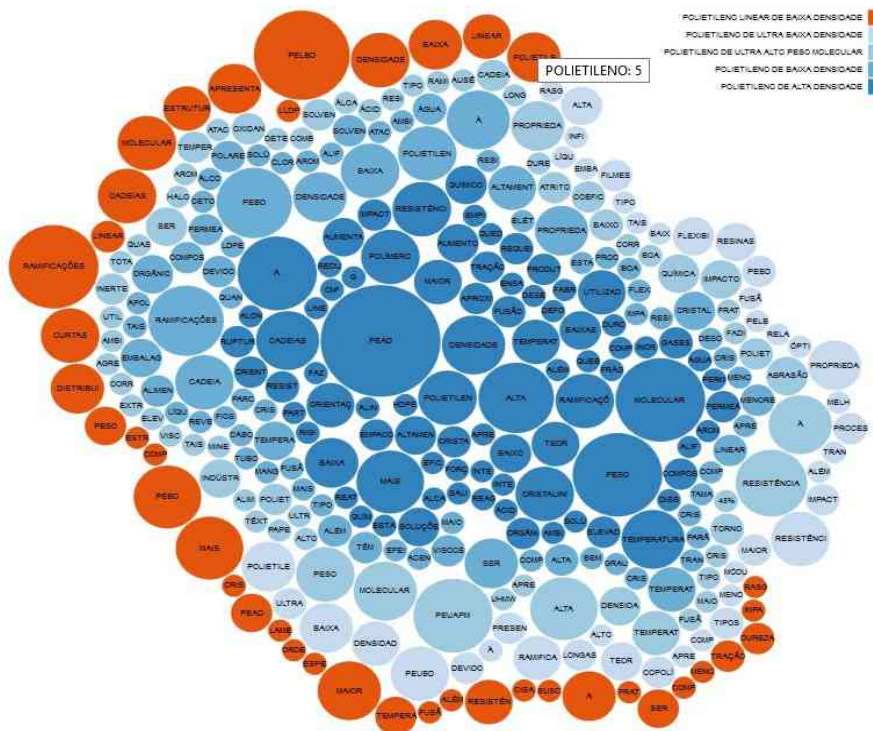


Figura 4 - Bubble Chart (Gráfico de Bolhas) de um estudante em todas as atividades.

A Figura 5 apresenta a visualização Word Cloud. A nuvem de palavras representa a visualização de dados de texto e é representada por palavras isoladas. A importância de cada palavra é mostrada com um tamanho da fonte ou cor. Neste formato pode-se perceber os termos mais proeminentes, sua importância relativa e a frequência da ocorrência das palavras no texto fornecido. Quanto maior for o número de ocorrências de palavras, maior será a nuvem de palavras. Esta visualização não mostra a pontuação, apenas apresenta visualmente, de forma intuitiva, quais as palavras mais relevantes na produção textual do estudante. Observou-se que as palavras mais importantes e de maior frequência com relação à análise do texto na atividade foram: PEAD, peso, molecular e alta.

A Figura 6 ilustra uma das visualizações utilizando Rotating Cluster. Ela é similar a visualização Collapsible Tree, apresentando contudo as palavras ficam em uma disposição diferente. O seu objetivo é gerar a listagem das palavras, não sendo possível saber quais delas são as mais relevantes na produção textual do estudante.

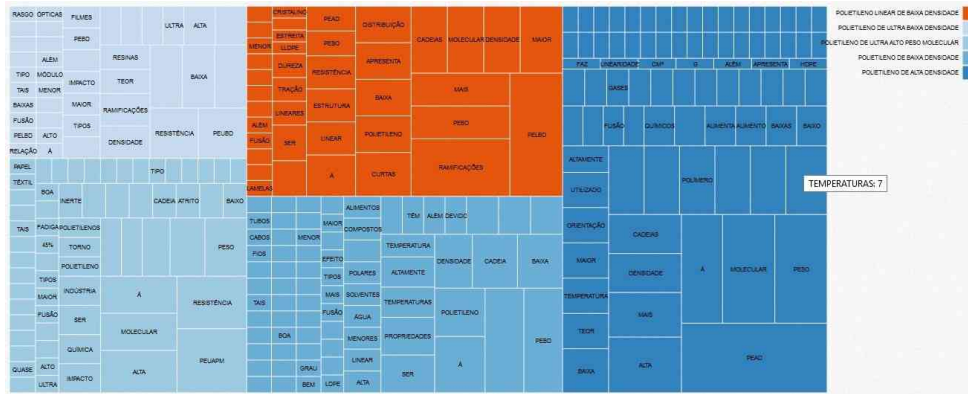


Figura 7 - Treemap (Mapa Árvore) de um estudante em todas as atividades.

A Figura 9 apresenta o LSA ScatterPlotBar, relacionando os resultados obtidos pelo algoritmo LSA com a produção textual dos alunos. Este gráfico mostra graficamente os valores produzidos pela aplicação do LSA às produções dos alunos. É possível observar a distribuição das produções textuais no espaço semântico, onde a diferença na qualidade de produção entre os estudantes de maior e menor nota são evidenciadas.

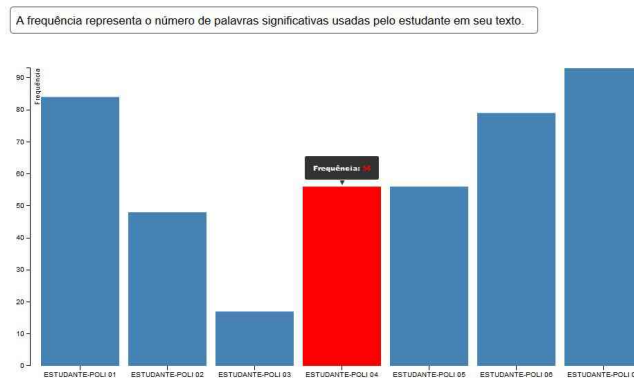


Figura 8 - Bar Chart (Gráfico de Barras) para uma turma em uma atividade.

Na análise do professor, os resultados obtidos foram satisfatórios, pois o sistema é capaz de avaliar a produção textual dos alunos e qualificá-los com referência ao texto utilizado no treinamento.

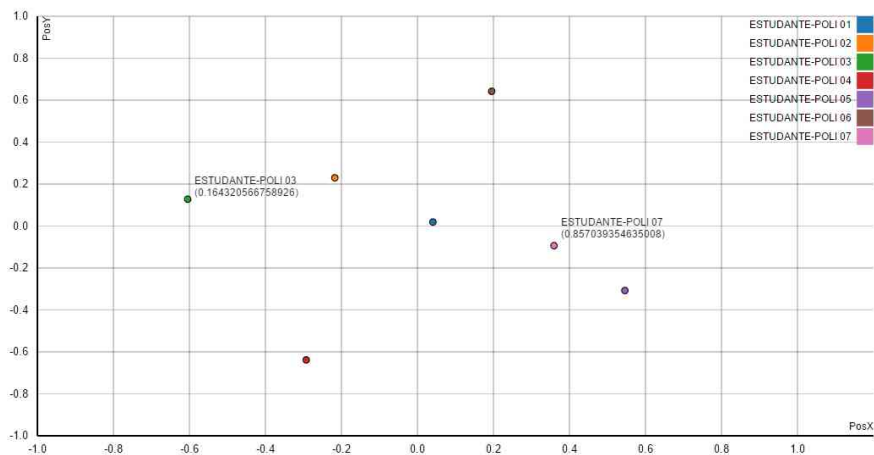


Figura 9 - LSA ScatterPlotBar para uma turma

A nota que o algoritmo atribui (coordenada x,y) não deve ser utilizada como nota absoluta, uma vez que pode haver notas negativas no resultado final. No entanto, podem

ser utilizadas para classificar as produções textuais entre si.

5. Conclusões Preliminares

O aprendizado de máquina oferece técnicas úteis na extração do conhecimento em textos, identificando palavras, expressões e termos relevantes em uma produção textual. Porém, para facilitar a análise de um texto é importante a apresentação visual destes dados. Foi então desenvolvido o software APT para a análise e visualização de produções textuais para uso educacional em qualquer área do conhecimento.

O primeiro experimento no uso do software foi realizado na disciplina de Química Orgânica. A avaliação preliminar feita pelo professor responsável pelo experimento identificou os seguintes pontos positivos no uso da ferramenta APT:

- (a) reconhecimento do vocabulário empregado pelos estudantes
- (b) ponderação do nível de coerência do texto do aluno em relação ao corpus do professor
- (c) percepção do desenvolvimento textual global da turma
- (d) comparação do nível de redação entre alunos da mesma turma
- (e) detecção das relações mais fortes entre conceitos do texto original
- (f) pontos de maior compreensão e menor compreensão do texto original

Como se trata de um projeto em andamento, novos algoritmos de aprendizado de máquina para análise das produções textuais se fazem necessários para que se produza resultados que auxiliem e contribuam ainda mais com a tarefa de leitura e avaliação textual. Para enriquecer a visualização dos dados minerados também são necessários novos produtos que proponham melhorar a interatividade do sistema. Para finalizar, conclui-se que os métodos automáticos podem operar sobre as produções textuais e produzir análises que forneçam ao professor visões sobre a evolução dos estudantes e da turma como um todo.

6. Referências Bibliográficas

- ARANHA, C. N. Uma Abordagem de Pré-Processamento Automático para Mineração de Textos em Português: Sob o Enfoque da Inteligência Computacional. Rio de Janeiro, 2007.
- BERRY, M. W.; DRMAC, Z.; JESSUP, E. R. Matrices, Vector Spaces, and Information Retrieval. SIAM Rev. Philadelphia, PA. 1999.
- BOSTOCK, M. D3 Data-Driven Documents. D3 Data-Driven Documents, 2013. Disponível em: <<http://d3js.org/>>. Acesso em: 22 junho 2015.
- CARD, S. K.; MACKINLAY, J. D.; SHNEIDERMAN, B. Information Visualization. Readings in Information Visualization: Using Vision to Think, Morgan Kaufmann Publishers. San Francisco, 1999.
- COUTINHO, F. M. B.; MELLO, I. L.; MARIA, L. C. D. S. Polietileno: Principais Tipos, Propriedades e Aplicações. Polímeros: Ciência e Tecnologia, Rio de Janeiro, v. 13, n. 1, 2003.
- DEERWESTER, S. E. A. Indexing by latent semantic analysis. Journal of the American Society for Information Science. [S.l.]. 1990.
- FELDAMAN, R., SANGER, J. Text Mining Handbook. Cambridge, Inglaterra: Cambridge University Press, 2006.
- FREITAS, C. M. D. S.; CHUBACHI, O. M.; LUZZARDI, P. R. G.; CAVA R. A. Introdução à Visualização de Informações. RITA – Revista de Informática Teórica e Aplicada, Instituto de Informática, v. 8, p. 143-158, 2001.
- IBM Watson. Disponível em: <<http://www.ibm.com/watson/>>. Acesso em: 22 setembro 2016.
- JUDELMAN, G. B. Knowledge Visualization: Problems and Principles for Mapping the Knowledge Space. Germany, 2004.
- KEIM, D. A. Information visualization and visual data mining. IEEE Trans. on Visualization and Comp. Graphics, v. 8, n. 1, p. 1-8, 2002.
- MICROSOFT. Cognitive Services. Disponível em: <<https://www.microsoft.com/cognitive-services/en-us/apis>>. Acesso em: 10 setembro 2016.
- MANNING, C. D., RAGHAVAN P., SCHUTZE, H., An Introduction to Information Retrieval. Cambridge University Press, 2008
- TAN, A.-H. Text mining: The state of the art and the challenges. In: PAKDD 1999 WORKSHOP ON KNOWLEDGE DISCOVERY FROM ADVANCED DATABASES. Beijing, 1999.
- WEBBER, C.G., CINI, G., LIMA, M.F.W.P. Facilitando a Análise se Dados Educacionais através se Ferramentas se Visualização. Revista Novas Tecnologias na Educação, RNOTE, v. 11, n. 3, 2013.