UNIVERSIDADE DE CAXIAS DO SUL
CENTRO DE CIÊNCIAS EXATAS E DA TECNOLOGIA
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

GUSTAVO LOVATO

APLICAÇÃO DA MINERAÇÃO DE TEXTOS NA ANÁLISE DE PRODUÇÕES TEXTUAIS

CAXIAS DO SUL
2015

GUSTAVO LOVATO

APLICAÇÃO DA MINERAÇÃO DE TEXTOS NA ANÁLISE DE PRODUÇÕES TEXTUAIS

Trabalho de Conclusão de Curso para a obtenção do Grau em Bacharel em Ciência da Computação da Universidade de Caxias do Sul

Orientadora: Carine Geltrudes Webber

CAXIAS DO SUL
2015

RESUMO

Este trabalho apresenta um estudo sobre Mineração de Textos aplicada ao contexto educacional. No ensino em geral existem muitas produções textuais que precisam ser manualmente analisadas pelos professores. Existe a necessidade do desenvolvimento de softwares que auxiliem a automatizar processos de leitura e revisão de textos. Eles podem auxiliar tanto professores quanto alunos no processo de aprendizagem como extrair palavras relevantes de textos de alunos para verificar a coerência com o que foi estudado. Também pode apoiar o próprio aluno na compreensão do que está apresentando em um texto e auxiliar o aluno a entender um texto didático (original). Além disso, pode ajudar a realizar uma tarefa através da mineração de textos e da visualização dos termos mais relevantes. Porém, todas são tarefas complexas que se aplicam a diversas áreas de conhecimento. Alguns trabalhos e softwares relacionados já foram desenvolvidos, tais como o UnderMine Text Miner (CRISTOFOLI, 2012), que será utilizado neste estudo. Não basta apenas minerar os textos, os resultados da mineração precisam ser apresentados de forma que quem os interprete possa compreender e obter conclusões. Neste enfoque, foi desenvolvida a ferramenta Análise de Produções Textuais, capaz de realizar o treinamento de determinada área do conhecimento e efetuar a mineração de produções textuais de estudantes. Com base nos dados gerados, foram utilizadas visualizações desenvolvidas com a tecnologia D3 Data-Driven Documents (BOSTOCK, 2013), para fins de acompanhamento da evolução de alunos durante realização de atividades textuais, oferecendo um conjunto de ferramentas funcionais para a análise de textos educacionais.

Palavras chave: mineração de textos educacionais, visualização de resultados, análise de textos educacionais.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Com a popularização da informação e o aumento do uso do computador, surge a oportunidade de automatizar processos do cotidiano, tornando as tarefas manuais mais automatizadas. Não é diferente na área da Educação. Conforme Rodrigues, Ramos, Silva e Gomes (2013), já não se discute a importância do uso das tecnologias da informação e comunicação na Educação; o que se busca é um aperfeiçoamento dos procedimentos atuais e a descoberta de novas técnicas que auxiliem o aprendizado.

A área de mineração de textos tem muito a contribuir com a área da Informática aplicada à Educação. Entende-se por mineração de textos o processo que extrai padrões, classes, grupos a partir de um conjunto de textos usados no treinamento. O objetivo da mineração em textos é identificar palavras, termos, expressões relevantes que se destacam no domínio minerado, encontrando padrões em dados não estruturados.

Ao encontrar padrões em textos através da mineração, é possível classificar os textos de forma mais precisa. Desta forma pode-se efetuar comparações e análises através de estatísticas geradas. Estas estatísticas precisam ser visualizadas, com o objetivo de auxiliar no entendimento. Para isto é necessário um trabalho de visualização dos dados minerados, visando auxiliar a compreensão de dados que em sua forma natural são de difícil entendimento.

Baseado em sistemas já existentes, surge a oportunidade da sequência do trabalho realizado por Cristofoli (2012), Mineração de Textos Baseada em Algoritmos Imunológicos, em cujo trabalho foi pesquisado e implementado um algoritmo que realiza mineração de textos em artigos científicos, seguindo os processos de mineração de textos e embasado em metodologias imunológicas. Este algoritmo é capaz de aprender o assunto ao qual são submetidos os textos, sendo capaz de apresentar a classificação de um texto minerado conforme a área de estudo.

A sequência do trabalho já desenvolvido abre campo para o uso do algoritmo proposto para mineração de textos educacionais, sendo assim possível identificar em que situação esse tipo de abordagem instrucional proporciona melhores benefícios educacionais ao aluno (BAKER, ISOTANI, 2011). O algoritmo

existente não é suficiente para serem analisados resultados mais satisfatórios, uma vez que apenas a classificação do texto quanto a área de ensino não é o suficiente para se ter conclusões e tomar ações perante à evolução de uma turma de alunos.

Tendo em vista que apenas a aplicação do algoritmo não soluciona todos os problemas, é preciso apresentar seus resultados na forma que quem os leia possa fazer análises, observações e atestar a evolução dos alunos observados. Desta forma é possível ao educador chegar a uma conclusão sobre as dificuldades e o aprendizado de uma classe de alunos, por exemplo. Visando esta melhoria, faz-se necessário uma avaliação de ferramentas úteis para a visualização das informações mineradas, integrando-a com o algoritmo já existente e que proporcione diversas visualizações sobre textos de alunos frente ao aprendizado da ferramenta.

1.1 OBJETIVOS

O objetivo deste trabalho é estudar e testar o algoritmo de mineração de texto desenvolvido por Cristofoli (2012) no trabalho de conclusão de curso (Mineração de Textos Baseada em Algoritmos Imunológicos). O algoritmo aplicado a textos educacionais deve proporcionar recursos de visualização para facilitar a compreensão dos resultados.

Para alcançar o objetivo geral, este trabalho será dividido em quatro objetivos específicos:

a) Realizar testes no algoritmo de mineração de texto desenvolvido por Cristofoli (2012) para verificar o aprendizado dos textos educacionais.
b) Pesquisar os métodos de visualização que melhor se adaptariam para a análise dos textos minerados.
c) Efetuar os ajustes necessários para que o algoritmo possa ser utilizado em textos educacionais, integrando-o com uma interface gráfica com recursos de visualização.
d) Analisar o resultado do software desenvolvido através de textos educacionais selecionados e submetidos no sistema.

1.2 ESTRUTURA DO TRABALHO

O trabalho está estruturado em quatro capítulos, conforme descrito a seguir. O primeiro capítulo tem por objetivo introduzir o leitor ao assunto abordado, apresentando os objetivos e o detalhamento da estrutura do trabalho. O capítulo 2 procura conceituar mineração e visualização de textos, apresentando o processo de mineração e técnicas para visualização de dados. O capítulo 3 apresenta a ferramenta UnderMine Text Miner, a qual foi desenvolvida por Cristofoli (2012) e será utilizada no presente trabalho. O capítulo 4 descreve a ferramenta Análise de Produções Textuais desenvolvida, onde é apresentada a modelagem, a arquitetura e a implementação. Também apresenta os testes realizados para a validação da ferramenta. Por fim, o capítulo 5 apresenta as conclusões finais sobre a implementação realizada no trabalho.

2 MINERAÇÃO E VISUALIZAÇÃO DE TEXTOS

Um processo de mineração consiste em retirar informações importantes e úteis de dados ou textos. Quando se trata de minerar dados, existem diversas técnicas bem conhecidas que apresentam resultados satisfatórios na maioria dos casos. Contudo, a mineração em bases textuais constitui um estudo mais recente, tendo recebido mais atenção com o surgimento da web. Como textos são dados em formatos não estruturados, geralmente gerados por humanos seguindo uma gramática e semântica em um contexto, a tarefa torna-se complexa pela necessidade de buscar padrões na ausência de uma ontologia ou metadados. Para validar um processo de mineração, as ferramentas de visualização são de grande valor. Elas permitem a visualização dos resultados dos textos minerados para uma posterior análise. Apresentar estas áreas, como elas se relacionam e os desafios atuais constituem o objetivo deste capítulo.

2.1 MINERAÇÃO DE TEXTOS

Conforme Feldman e Sanger (2006), a mineração de texto pode ser definida como um processo de conhecimento intensivo em que um usuário interage com uma coleção de documentos ao longo do tempo por meio de um conjunto de ferramentas de análise. Diferentemente da mineração de dados, onde se tem como objetivo descobrir “novas” informações através da análise de grandes quantidades de dados (WITTEN, FRANK E HALL, 2011), a mineração de textos visa extrair padrões interessantes e não triviais a partir de documentos textuais não estruturados (TAN, 1999).

Segundo Aranha (2007), a mineração de textos pode ser dividida em cinco etapas (Figura 1). Para um texto ser submetido ao processo de mineração, ele deve receber um tratamento diferente daquele dado aos dados, por se tratar de um tipo de dado não estruturado.

Figura 1 - Processo de Mineração de Textos.

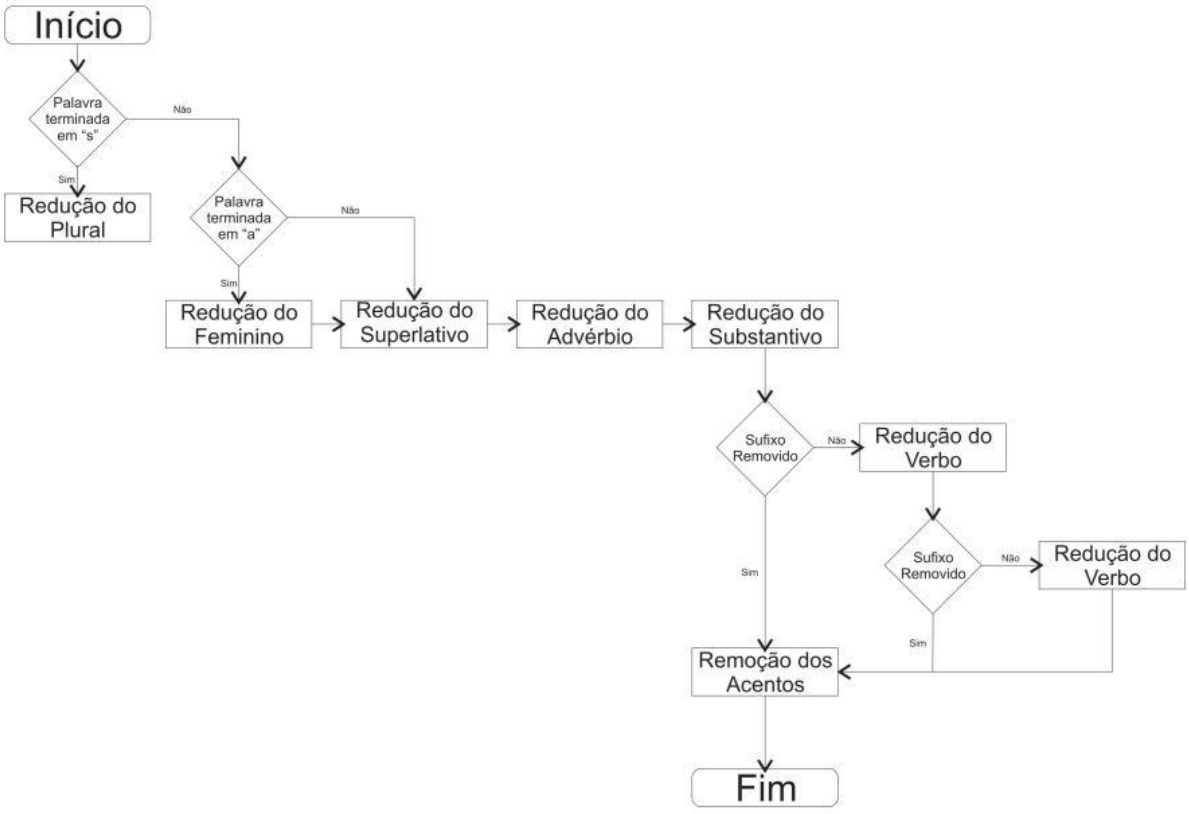
Fonte: Aranha (2007).

Na primeira etapa do processo são coletados os documentos que formarão a base de textos a serem analisados. Os textos podem ser de várias fontes, dependendo da aplicação em que serão utilizados. No caso de uma aplicação voltada à educação, os textos podem ser coletados de uma turma de alunos, visando analisar os resultados do aprendizado.

O pré-processamento, que é 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.

A etapa de indexação é composta por três fases, onde a primeira consiste em separar os termos simples dos termos compostos. Na segunda fase acontece a remoção das *stopwords*, que são as palavras que não possuem relevância para o texto e que não poderão mais ser recuperadas. A terceira fase é a da normalização morfológica, onde o objetivo é igualar as palavras sem a sua classificação e estrutura perante o texto. Nesta etapa é executado o algoritmo de *stemming*, que em inglês significa reduzir uma palavra ao seu radical. Orengo e Huyck (2001) apresentam um algoritmo de *stemming* para a língua portuguesa desenvolvido na linguagem C que é composto por oito etapas (Figura 2).

Figura 2 - Algoritmo de *Stemming* para língua portuguesa.



Fonte: Orengo e Huyck (2001).

As etapas para o processo de *stemming* são as seguintes:

1. Remoção do plural – Basicamente consiste em remover o “s” do final da palavra. Porém, há uma lista de exceções, como a palavra “lápis” por exemplo.
2. Remoção do feminino – As palavras femininas são transformadas na correspondente masculina. Exemplo: “brasileira” será transformado em “brasileiro”.
3. Remoção de advérbio – Como o único sufixo que identifica um advérbio é o sufixo “mente”, o mesmo será retirado das palavras. Para este passo também há uma lista de exceções.
4. Remoção de aumentativo e diminutivo – Remove os sufixos dos substantivos e adjetivos que identificam aumentativo e diminutivo. Exemplo: “bonitinho” e “paredão”.
5. Remoção de sufixos em nomes – Existe uma lista com 61 sufixos para substantivos e adjetivos. O programa testa a palavra contra esta

lista, caso exista o sufixo é removido e as etapas seis e sete não são executadas.

6. Remoção de sufixos em verbos – Os verbos podem variar de acordo com o tempo, a pessoa, o número e o modo. A estrutura das formas verbais pode ser representada por: radical + vogal temática + tempo + pessoa. A finalidade deste passo é reduzir as formas verbais ao seu radical correspondente.
7. Remoção de vogais – Nesta etapa é removida a última vogal (“a”, “e” ou “o”) das palavras que não foram examinadas pelas etapas cinco e seis.
8. Remoção de acentos – Na última etapa, apenas a acentuação é retirada.

Após classificados os termos no texto é efetuada a etapa de extração do conhecimento, conforme o objetivo que se deseja. Por exemplo, se a aplicação é voltada à educação, define-se a relevância dos termos e o objetivo que se deseja chegar nesta área. Nem todas as palavras possuem a mesma relevância no texto, por isso é importante o cálculo da relevância dos termos, que se baseia também na frequência em que os termos aparecem no texto. As três frequências mais comuns são: frequência absoluta, frequência relativa e frequência inversa de documentos.

A frequência absoluta representa a medida da quantidade de vezes que um termo aparece em um documento. A frequência relativa leva em conta o tamanho do documento, ou seja, a quantidade de palavras que ele possui, e os pesos são normalizados de acordo com esta informação. A frequência inversa de documentos utiliza o cálculo das duas frequências anteriores para ser calculada, assim é possível aumentar a importância dos termos recorrentes em poucos documentos e diminui a importância dos termos que aparecem muitas vezes, uma vez que os termos que aparecem poucas vezes, geralmente, tem a propriedade de descrever mais o assunto abordado.

A última etapa é a da análise dos resultados. É nesta etapa que serão apresentadas as informações para que se obtenham conclusões da mineração realizada nos textos. Esta etapa não é feita computacionalmente, mas sim, por uma figura humana de interesse aos dados e de uso da aplicação.

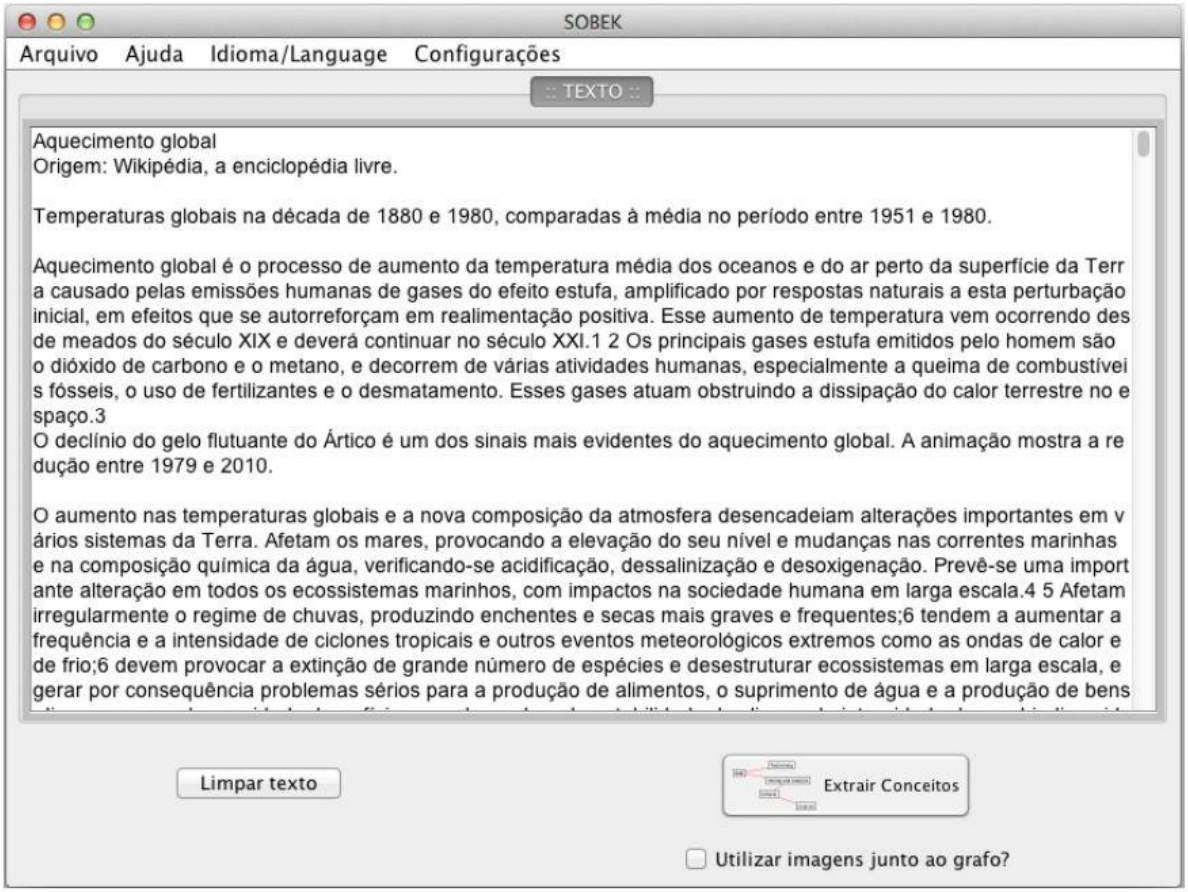
2.1.1 Aplicações Relacionadas

A mineração de textos tem sido utilizada em aplicações voltadas à educação com o objetivo de auxiliar os alunos na produção textual. Segundo Cabral (2009), os alunos têm muita dificuldade para se expressar textualmente, o que não acontece com a oralidade, uma vez que a linguagem oral não exige maiores formalidades, podendo ser utilizada a linguagem coloquial e gestos que facilitem o entendimento.

Como exemplos de ferramentas de mineração de textos no auxílio à área de educação tem-se:

- SOBEK (SOBEK, 2015)
- TagCrowd (TAGCROWD, 2015)
- Word Counter (WORDCOUNTER, 2015)
- TextAlyser (TEXTALYSER, 2015)

Figura 3 - Entrada de texto do SOBEK.



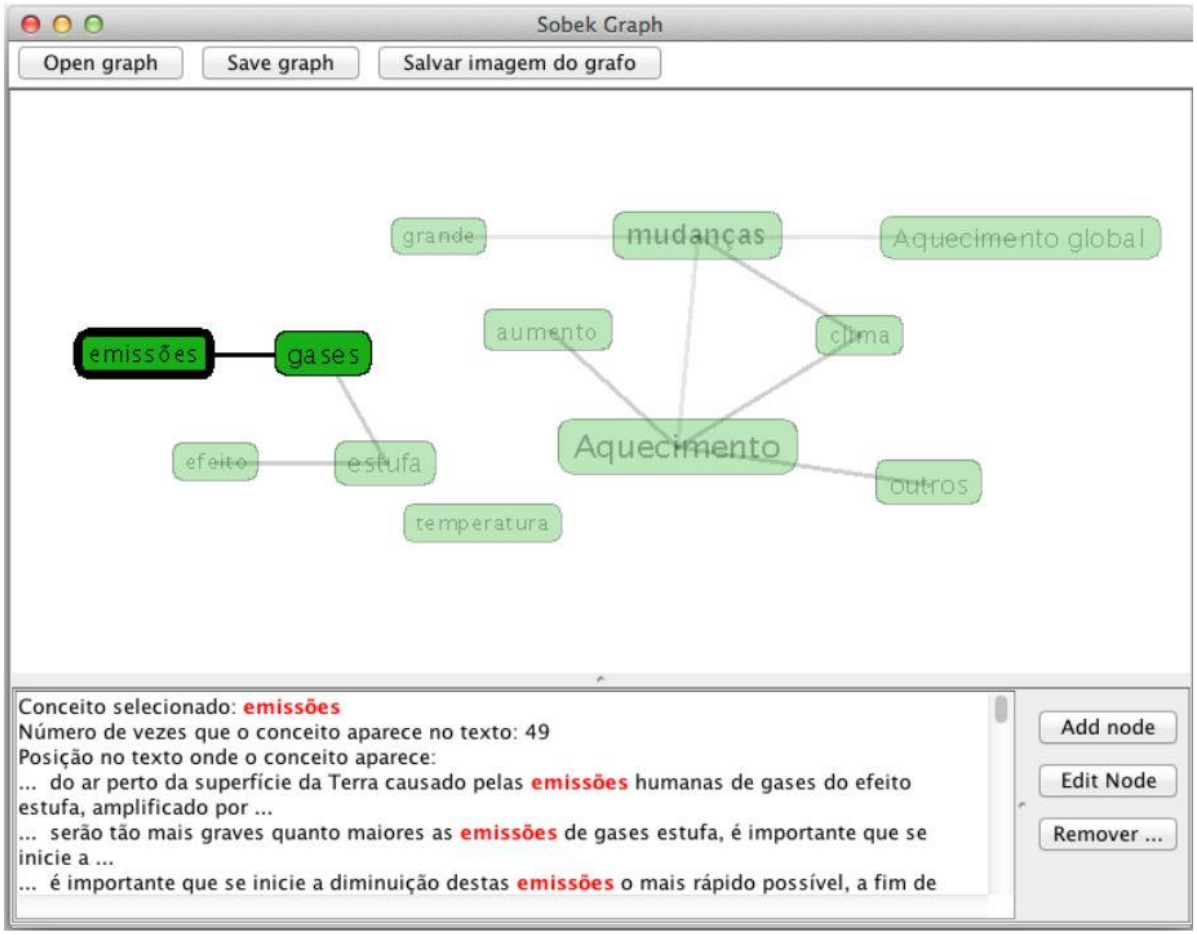
Fonte: http://sobek.ufrgs.br/uploads/2/3/3/9/23394804/sobek_quick_reference_guide_pt.pdf.

O SOBEK é uma ferramenta utilizada para extrair termos frequentes em documentos, encontrando os relacionamentos entre estes, servindo de apoio aos professores no acompanhamento de trabalhos de escrita colaborativa (MACEDO *et al.*, 2009). Esta ferramenta se destaca por apresentar de forma gráfica os resultados da relação dos termos importantes no texto, auxiliando o aluno a formular a sua própria compreensão sobre o texto lido.

Na primeira tela do programa, pode ser digitado ou copiado o texto em tela ou carregado um arquivo nos formatos *txt*, *doc* ou *pdf* (Figura 3).

No próximo passo os conceitos principais do texto são extraídos, gerando um grafo com os principais termos e seus relacionamentos. É possível realizar a alteração dos nodos conforme a necessidade (Figura 4).

Figura 4 - Resultado do SOBEK.



Fonte: http://sobek.ufrgs.br/uploads/2/3/3/9/23394804/sobek_quick_reference_guide_pt.pdf.

Os alunos têm dificuldade em interpretar e retirar a ideia principal de um texto para elaborar um resumo. Segundo Rapkiewicz e Moretto (2013), o grafo

gerado pela ferramenta Sobek auxilia o aluno na compreensão do texto com a visualização dos principais termos.

Uma ferramenta mais simples é a TagCrowd, que permite criar nuvens de palavras, destacando as palavras mais importantes conforme a frequência que aparecem no texto. Este aplicativo é interessante pelo formato visual em que as palavras são apresentadas, porém a ferramenta não apresenta relação entre os termos, apenas a sua frequência no texto sem efetuar a remoção das *stopwords* (Figura 5 e Figura 6).

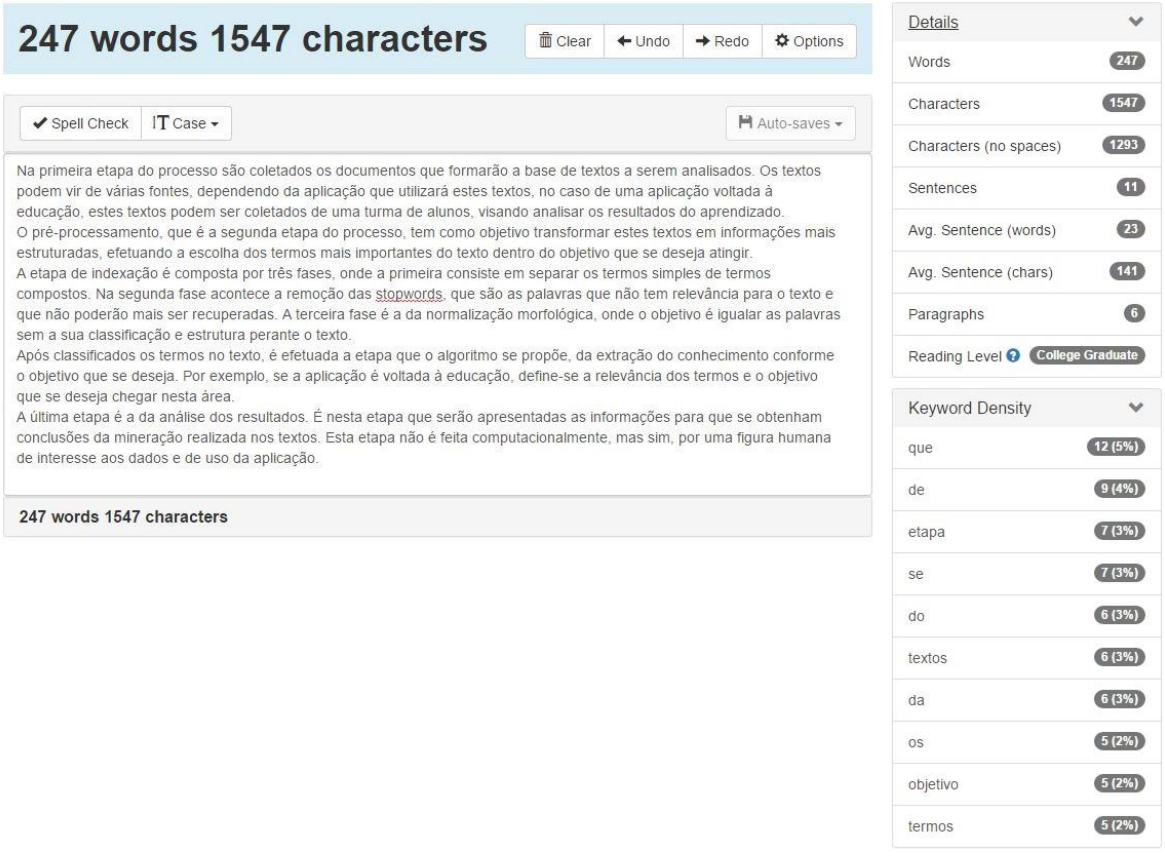
Figura 5 - Entrada de texto do TagCrowd.

Fonte: http://tagcrowd.com/.

Figura 6 – Resultado do TagCrowd.

Fonte: http://tagcrowd.com/.

Figura 7 - Entrada de texto e resultado do Word Counter.

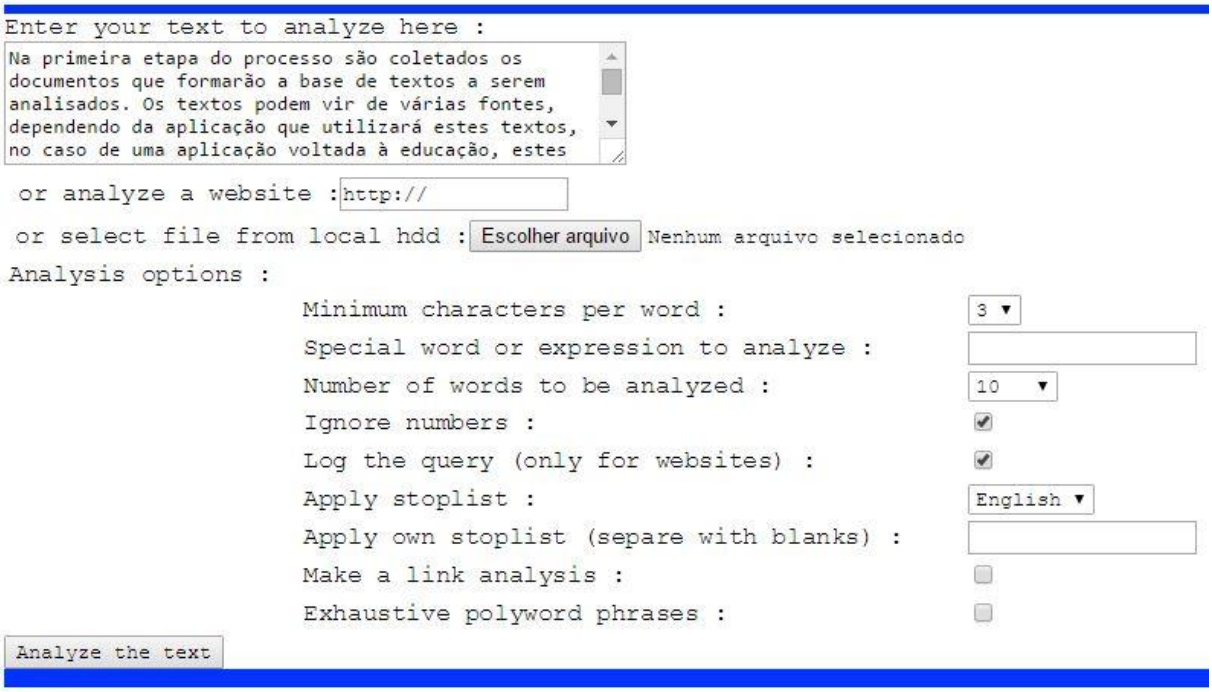


Fonte: http://www.wordcounter.net/.

O Word Counter também é uma ferramenta online com o objetivo de apresentar a relação das palavras mais usadas em um texto. Conforme Klemann, Reategui e Rapkiewicz (2011), essa ferramenta é útil pois mostra as palavras repetidas ou redundantes em uma lista. A ferramenta não apresenta os resultados graficamente, apenas no formato de uma lista conforme Figura 7, onde o texto é inserido no quadro à esquerda e o resultado é apresentado na lista à direita.

De forma similar, TextAlyser é uma ferramenta online que analisa as palavras e expressões do texto, apresentando como resultado a contagem e uma série de estatísticas dos termos e palavras utilizadas. Esta ferramenta apresenta um índice de "facilidade de leitura" (*readability*) do texto analisado, utilizando como critério o tamanho das frases e as estatísticas encontradas pelo analisador. Este programa apresenta o resultado na forma de estatística, mas não possui uma interface gráfica elaborada para isto. Na Figura 8 é mostrado a entrada de dados da aplicação e na Figura 9 é mostrado o resultado, apresentando estatísticas tais como: contagem de palavras, número de palavras diferentes e número de caracteres.

Figura 8 - Entrada de texto do TextAlyser.



Fonte: http://textalyser.net/.

Figura 9 - Parte do resultado do TextAlyser.

Textalyser Results

The complete results, incuding compexity factor, and other features

Total word count :	181
Number of different words :	113
Complexity factor (Lexical Density) :	62.4%
Readability (Gunning-Fog Index) : *(6-easy 20-hard)*	17.9
Total number of characters :	1557
Number of characters without spaces :	1244
Average Syllables per Word :	2.27
Sentence count :	12
Average sentence length (words) :	25.45
Max sentence length (words) :	41
(os textos podem vir de várias fontes dependendo da aplicação que utilizará estes textos no caso de uma aplicação voltada à educação estes textos podem ser coletados de uma turma de alunos visando analisar os resultados do aprendizado)	
Min sentence length (words) :	9
(a última etapa é a da análise dos resultados)	
Readability (Alternative) beta : *(100-easy 20-hard, optimal 60-70)*	-11

Fonte: http://textalyser.net/.

Klemann, Reategui e Rapkiewicz (2011) apresentam uma tabela comparativa destas ferramentas de mineração de textos, tornando claros os propósitos de cada uma (Tabela 1). Com relação ao cenário de disponibilidade via web (coluna on-line), os três softwares, TextAlyser, Word Counter e TagCrowd dispõem de versõs, enquanto que o Sobek possui apenas versão desktop. Os softwares TextAlyser e Word Counter possuem a funcionalidade de contagem de termos, enquando o TagCrowd e Sobek não possuem esta função (coluna Contagem de termos). Quanto a apresentação dos termos, apenas o software Word Counter disponibiliza todos os termos (coluna Apresentação de todos os termos), enquanto que os softwares TagCrowd e Sobek apresentam apenas os termos

relevantes (coluna Apresentação de termos relevantes). Os softwares TextAlyser, Word Counter e Sobek identificam a frequência dos termos (coluna Frequência dos termos), enquanto que o TagCrowd não possui esta funcionalidade. Apenas o software Sobek é capaz de analisar o relacionamento dos termos (coluna Relacionamento dos termos), enquanto que os softwares TextAlyser, Word Counter e TagCrowd não são capazes. No que se refere a apresentação dos dados, os softwares TagCrowd e Sobek possuem visualização gráfica dos termos (coluna Visualização Gráfica dos termos) e o Sobek também possui visualização gráfica dos termos relacionados (coluna Visualização Gráfica dos termos relacionados), enquanto que e os softwares TextAlyser e Word Counter não possuem nenhuma visualização gráfica.

Tabela 1 - Análise comparativa das ferramentas de mineração de textos.

	Online	Contagem de termos	Apresenta-ção de todos os termos	Apresenta-ção de termos relevantes	Frequência dos termos	Relacio-namentos dos termos	Visualiza-ção Gráfica dos termos	Visualização Gráfica dos termos e relacionamentos
TextAlyser	X	X			X			
Word Counter	X	X	X		X			
TagCrowd	X			X			X	
Sobek				X	X	X	X	X

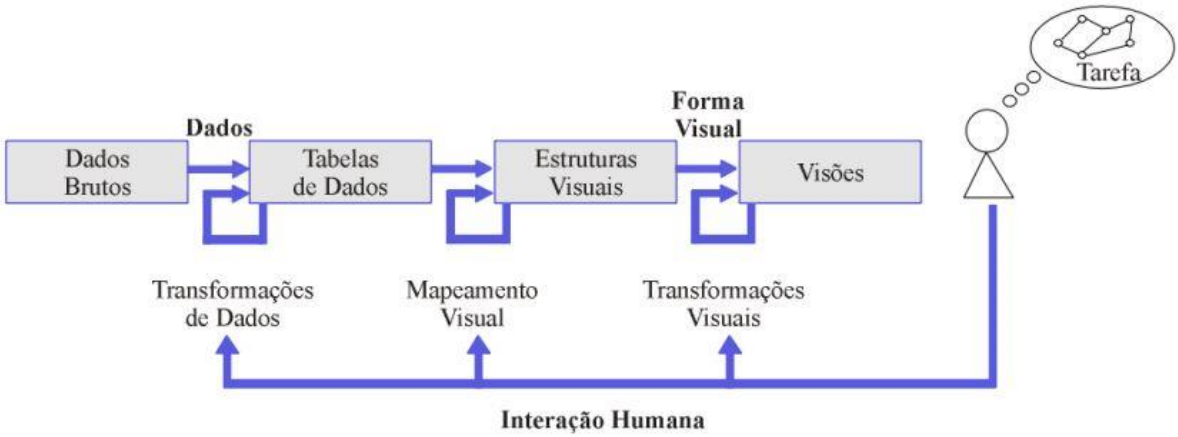
Fonte: Klemann, Reategui e Rapkiewicz (2011).

2.2 VISUALIZAÇÃO DE INFORMAÇÃO

Durante o processo de descoberta de conhecimento é de extrema importância a apresentação dos dados para uma análise por parte de uma figura humana. Afinal, o resultado tem de levar a uma conclusão precisa do que foi avaliado. Card *et al.* (1999) definem Visualização de Informação como sendo o uso de representações visuais de dados abstratos suportadas por computador e iterativas para ampliar a cognição.

Card *et al.* (1999) sugerem um modelo de referência para construir visualização de informações (Figura 10).

Figura 10 - Modelo de Referência de Informação.



Fonte: Card *et al.* (1999).

A primeira etapa, chamada de Transformação de Dados, transforma os dados brutos em uma representação mais estruturada, na forma de tabelas. Estes dados sofrem uma separação seguindo critérios, como é proposto por Freitas *et al.* (2001), onde os dados são classificados segundo seus critérios (Tabela 2).

Tabela 2 - Classificação de Informações.

Critério	**Classe**	**Exemplo**
Classe da informação	Característica Escalar Vetor Tensor Relacionamento	Gênero Temperatura Grandeza física associada a um fluído Grandeza física associada a um fluído *Link num hiperdocumento*
Tipo dos valores	Alfanumérico Numérico Símbolo	Gênero Temperatura *Link num hiperdocumento*
Natureza do domínio	Discreto Contínuo Contínuo-Discretizado	Marcas de automóveis Superfície de um terreno Anos (tempo discretizado)
Dimensão do domínio	1D 2D 3D n-D	Medida de uma grandeza no tempo Superfície de um terreno Volume de dados médicos Dados de uma população

Fonte: Freitas *et al.* (2001).

Na etapa de Mapeamento Visual, após efetuada a caracterização, os dados das tabelas são transformados em uma estrutura visual que represente visualmente os dados das tabelas. Este processo é feito computacionalmente através de uma função matemática executada por um algoritmo.

A última etapa de Transformações Visuais, transforma as imagens estáticas, vindas da fase de Mapeamento Visual, em visões que possibilitam ao usuário parametrizar a forma de visualização de acordo com a sua necessidade.

2.2.1 Técnicas de Visualização de Dados

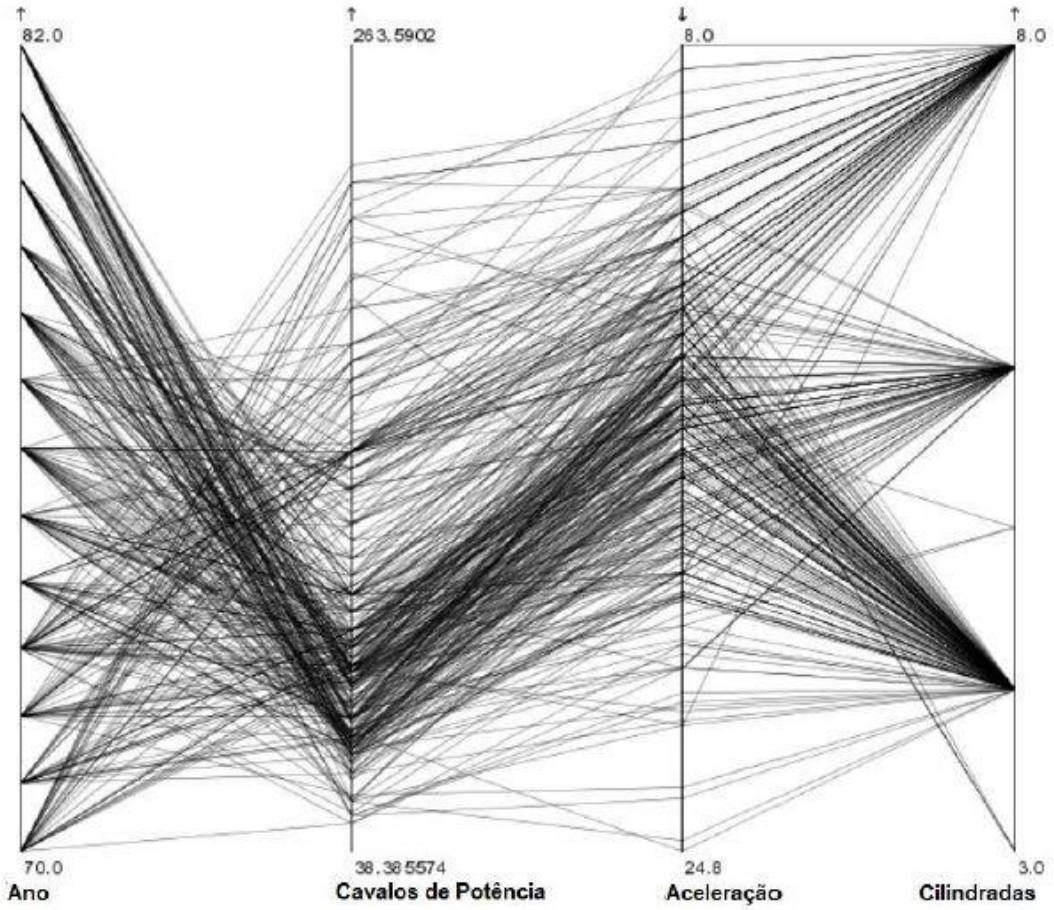
Muitas formas podem ser usadas para a apresentação da informação, dependendo na necessidade para cada conjunto de dados analisados. Por exemplo, pode-se criar um banco de dados multidimensional originado da mineração de textos. Esta seção aborda de forma resumida as técnicas de visualização de informação, pois um estudo anterior (CINI, 2013) já se aprofundou neste assunto.

Hoffman (1999) e Keim e Kriegel (1996) definem as seguintes técnicas para a esta visualização:

a) Técnicas de projeção geométrica
b) Técnicas iconográficas
c) Técnicas orientadas a pixel
d) Técnicas hierárquicas
e) Técnicas baseadas em grafos
f) Híbridas

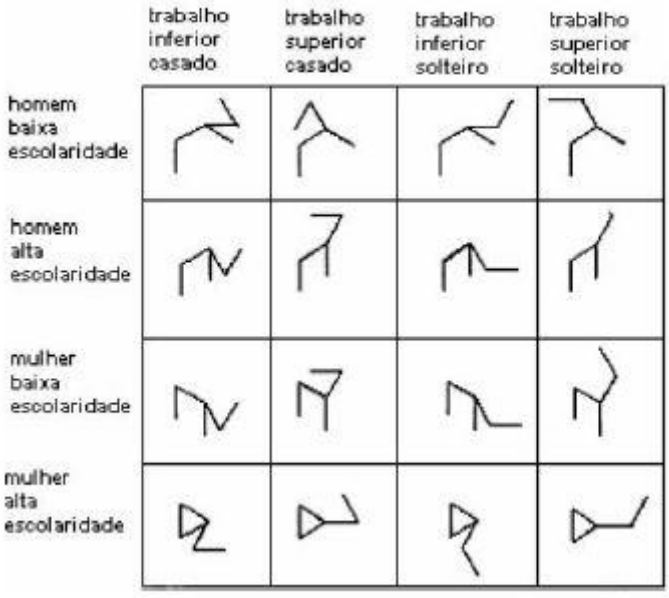
As técnicas de projeção geométrica têm por objetivo encontrar projeções “interessantes” em um conjunto de dados de um banco multidimensional. Como exemplo pode ser citada a técnica de Coordenadas Paralelas, que visa dimensionar linearmente, através dos eixos x ou y, um item a partir do seu valor mínimo ao valor máximo. Cada item é representado por uma linha poligonal. A Figura 11 representa dados de quatro dimensões. O número de itens é restrito, pois, quanto mais itens, mais difícil fica a visualização, visto que as linhas podem se sobrepor.

Figura 11 - Coordenadas Paralelas.



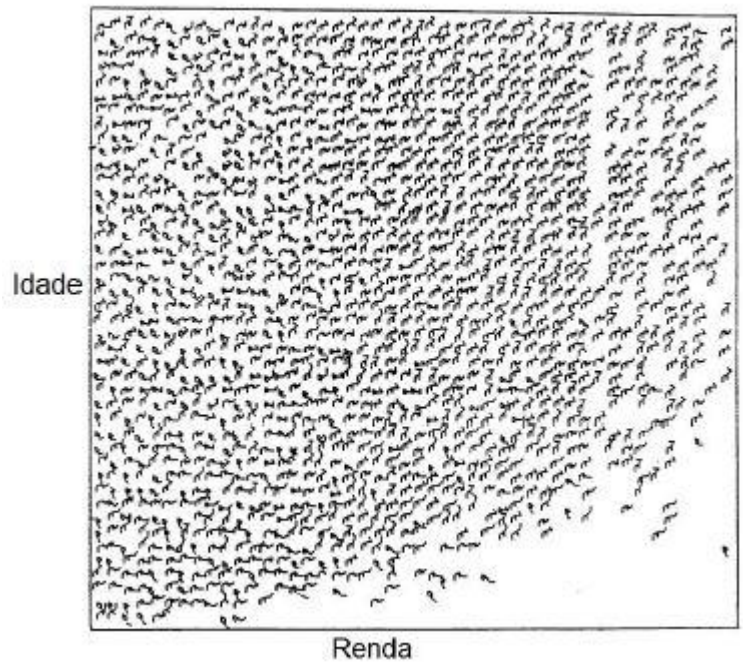
Fonte: Hauser, Ledermann e Doleisch (2002).

Figura 12 - Atributos mapeados em ícones.



Fonte: Pickett e Grinstein (1988).

Figura 13 - Amostra populacional.

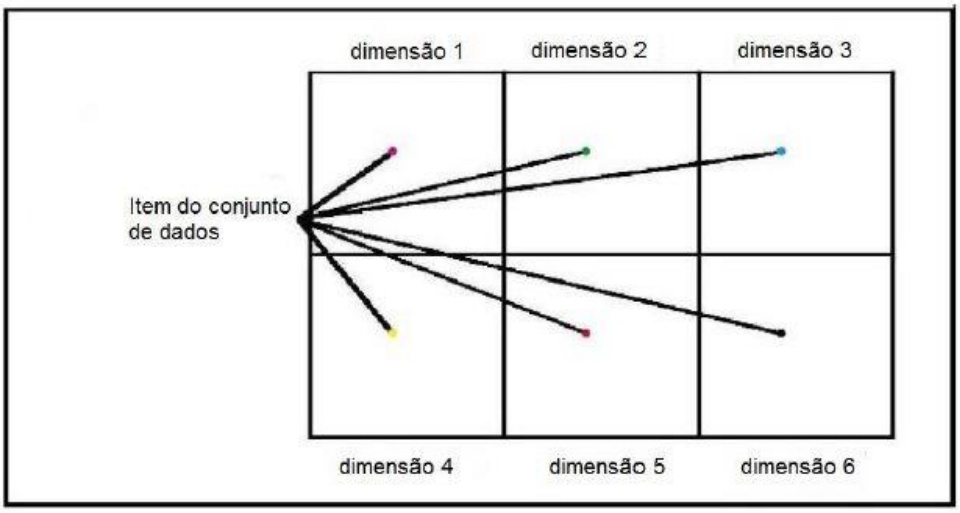


Fonte: Pickett e Grinstein (1988).

As técnicas iconográficas consistem em mapear os atributos dos dados multidimensionais para um ícone. Como técnica mais adequada para uma grande quantidade de dados, Pickett e Grinstein (1988) apresentaram a técnica *"stick figure"*, que consiste em ícones formados por linhas, onde as dimensões são mapeadas para x e y. A Figura 12 contém os atributos mapeados em ícones da técnica de *"stick figure"* e a Figura 13 contém a amostra populacional para gerar a visualização.

As técnicas orientadas a *pixel* mapeiam cada atributo de um item do conjunto de dados a um pixel colorido e apresentam suas dimensões em janelas separadas. Cada item pode ter n atributos, o que pode resultar em n dimensões. Para representar esta visualização são utilizadas técnicas como: a técnica de padrão recursivo, onde os *pixels* são organizados de uma forma recursiva, mais utilizada onde os atributos dos dados representam uma ordem cronológica, e a técnica de segmentos de círculos, onde utiliza-se um círculo dividido em segmentos que representam as dimensões e cada item é representado por um único *pixel* colorido (Figura 14).

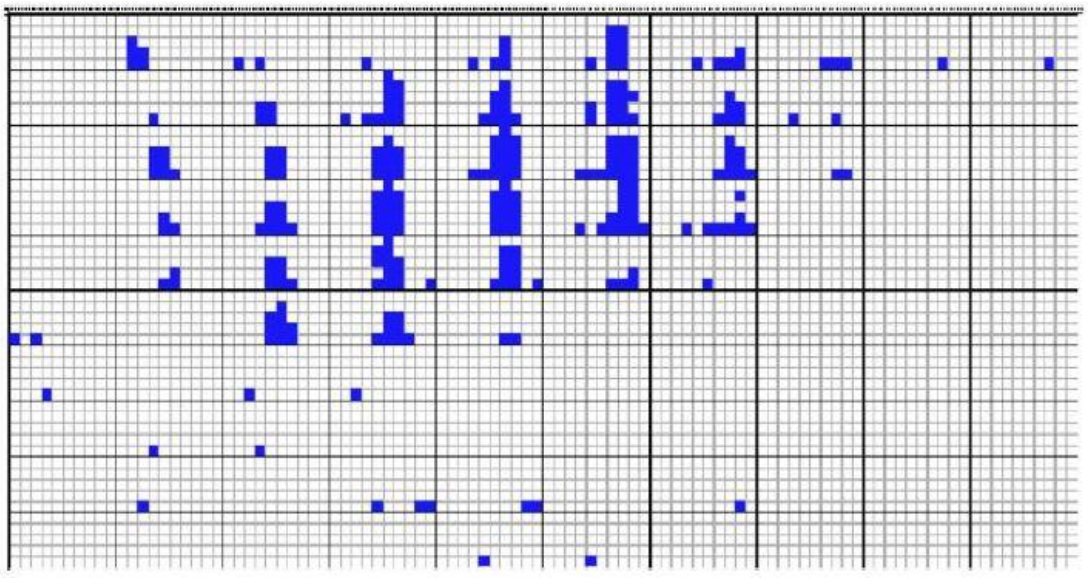
Figura 14 - Técnica orientada a *pixel* com seis dimensões.



Fonte: Keim (1996).

As técnicas hierárquicas subdividem um espaço em k-dimensões, onde cada subespaço é apresentado de forma hierárquica (KEIM e KRIEGEL, 1996). Existem diversas técnicas hierárquicas, porém, Wong e Bergeron (1997) citam como mais vantajosa a técnica de *Dimensional Stacking* pois não precisam de funções ou regras adicionais para plotar os dados em tela. Esta técnica é baseada em aplicar um sistema de coordenadas dentro de um sistema de coordenadas, onde duas dimensões formam os eixos externos e outras duas dimensões serão colocadas dentro destas dimensões, e assim sucessivamente (KEIM, 2002).

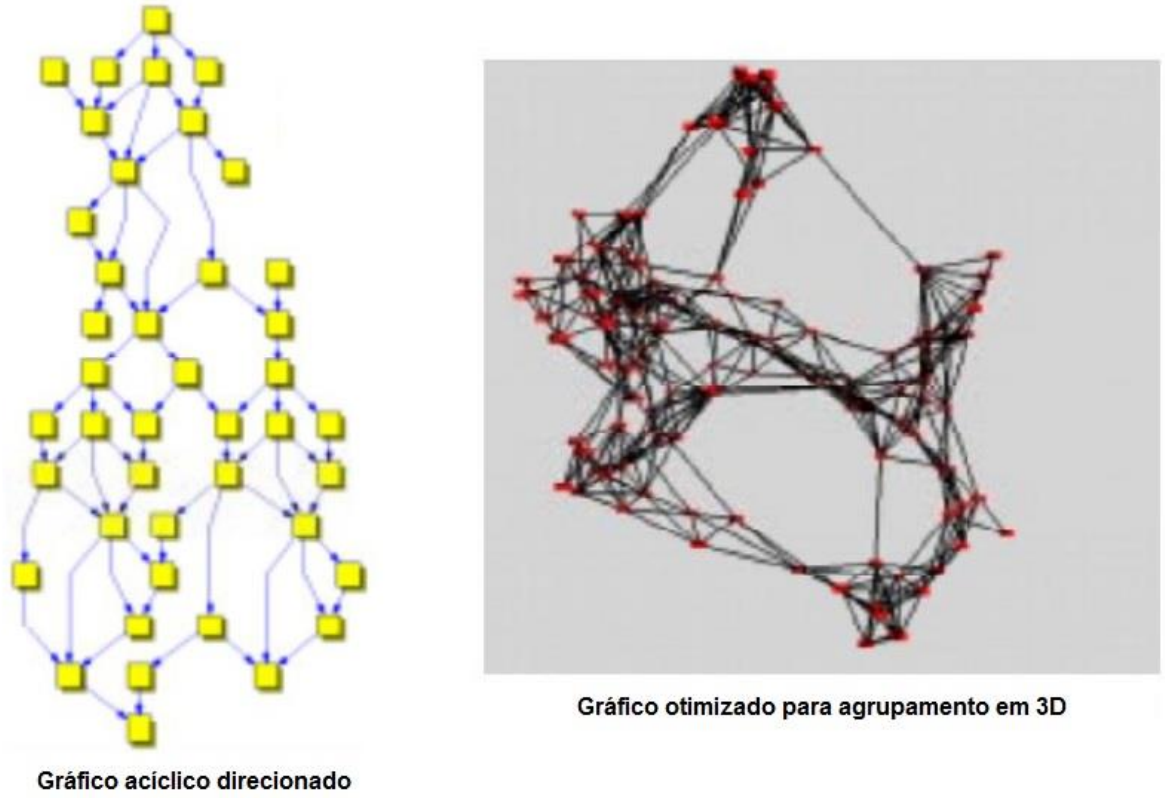
Figura 15 - Técnica de *Dimensional Stacking* para técnicas hierárquicas.



Fonte: Keim (2002).

A Figura 15 apresenta quatro dimensões para a técnica de *Dimensional Stacking*, as duas dimensões mais externas representam a latitude e a longitude de uma exploração de óleo, e as duas internas representam o grau de minério e a profundidade.

Figura 16 - Técnica baseada em grafos.



Fonte: Keim (1997).

As técnicas baseadas em grafos são utilizadas para informações hierárquicas mais complexas, altamente interconectadas e que possam conter propriedades pertencentes à teoria dos grafos. Os grafos podem ser apresentados na forma 2D ou 3D.

Caso o grafo seja 2D, as propriedades podem ser:

- Planaridade: não há arestas que se cruzem;
- Ortogonalidade: somente linhas ortogonais;
- Propriedade de grade: As coordenadas dos vértices são números inteiros.

Conforme Keim (1997), as visualizações podem apresentar propriedades estéticas, chamadas de “Metas de Otimização”, sendo elas:

- Número mínimo de cruzamento entre as arestas;
- Exibição ótima das simetrias;
- Exibição ótima dos *clusters*, - agrupamentos;
- Número mínimo de curvas em gráficos poligonais;
- Distribuição uniforme dos vértices;
- Comprimento uniforme das arestas.

A Figura 16 apresenta um exemplo de grafo em 2D e um exemplo de grafo em 3D.

2.3 TÉCNICAS PARA VISUALIZAÇÃO TEXTUAL

Judelman (2004) dividiu a visualização em três categorias, chamadas "*Meta-Estruturas*". Cada meta-estrutura tem como objetivo representar tipos diferentes de dados. As meta-estruturas são as seguintes:

a) Complexidade
b) Contexto
c) Dinâmica

A meta-estrutura chamada complexidade visa representar dados que contenham uma relação hierárquica, o que é o caso das árvores de dados e de redes.

A categoria chamada de contexto tem por objetivo a representação de dados que tenham relacionamento entre si, visando demonstrar diferenças de relações.

Já a meta-estrutura dinâmica visa a demonstração de dados que apresentam um fluxo ou uma mudança no tempo ou espaço.

As categorias de visualização são demonstradas na Tabela 3, apresentando os exemplos de aplicações de cada uma e a tendência para tamanho da estrutura: pequeno (1 – 30 objetos), médio (30 – 100 objetos) e grande (100 – milhões de objetos).

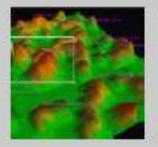
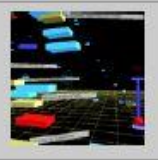
Tabela 3 - Categorias de visualização.

Meta-estrutura	**Descrição**	**Exemplos**	**Elementos Semiológicos Dominantes**	**Tendência para o Tamanho de Estrutura**
complexidade	caminhos, estruturas	árvore, hierarquia, rede	ilustração de nós e links	médio a grande
contexto	relacionamentos, conteúdos	objetos de agrupamentos	ilustração de similaridade e diferença	pequeno a médio
dinâmica	movimento, mudanças espaço-temporais	diagramas de processo e campos de vetor	eixo temporal, ilustração de fluxo	pequeno a grande

Fonte: Judelman (2004).

Para a visualização de texto e comparativo entre textos analisados, o que é tratado no presente trabalho, a categoria que melhor se adapta é a contexto. Segundo Judelman (2004), esta é a categoria mais apropriada para a visualização de objetos com base em propriedades semânticas, e cita duas técnicas para visualizar contexto: exploração e gráfico (Figura 17).

Figura 17 - Técnicas para visualizar contexto.

Meta-structure	Actions	Technique	Variations	Examples	Size	Visual structure
context	explore	clustering	overlapping zones	cluster map	small	
			magnetized pixels, search terms on surface	Kryphthästhesie, Sinnzeug	small	
	chart	spatial grouping	3D mountains	Vxinsight	medium	
			virtual environment	VR-VIBE, Starwalker	medium	

Fonte: Judelman (2004).

A técnica de exploração permite explorar o espaço semântico interativamente, alterando a distribuição dos agrupamentos de dados, dando ao usuário a oportunidade de alterar o gráfico conforme sua preferência.

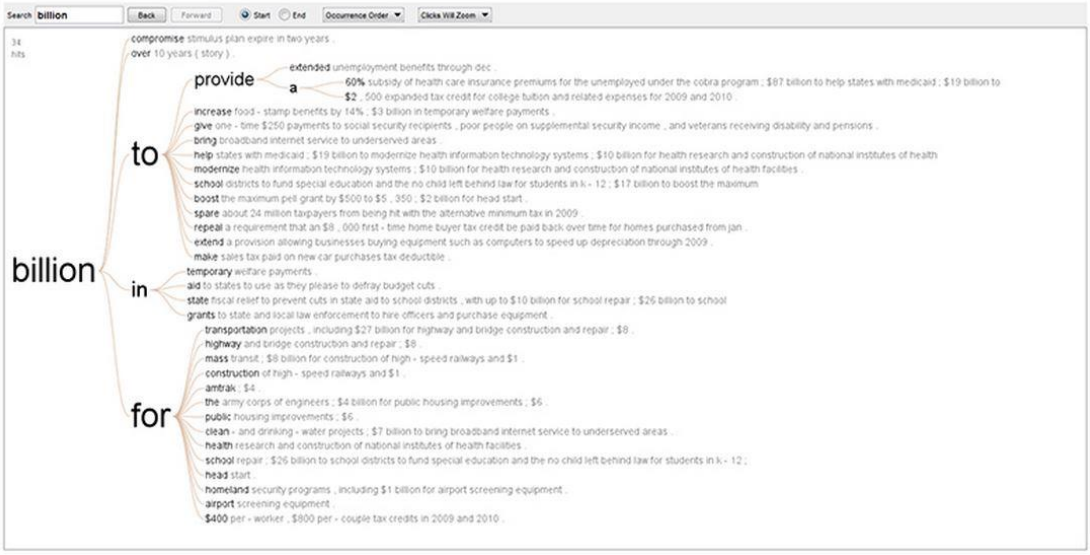
A técnica de gráfico é semelhante a técnica de exploração, porém não permitem ao usuário o mesmo grau de controle da representação gráfica. A distribuição dos grupos semânticos são geradas previamente a partir dos dados estáticos.

2.3.1 Aplicações para Visualização Textual

Para a visualização textual, existem alguns aplicativos que recebem um conjunto de dados e são responsáveis por apenas apresentar as informações já selecionadas, a exemplo tem-se: Many Eyes, Prefuse e InfoVis.

Many Eyes é uma ferramenta desenvolvida pela IBM, está disponível on-line em um site voltado ao compartilhamento das visualizações, sendo possível gerar visualizações de todos os tipos de dados, a fim de analisá-los e compartilhá-los. A Figura 18 mostra um exemplo de Word Tree gerado pelo Meny Eyes. Este tipo de visualização é apenas um dos que o aplicativo disponibiliza, também possui visualizações como: *Treemap* e *Word cloud*.

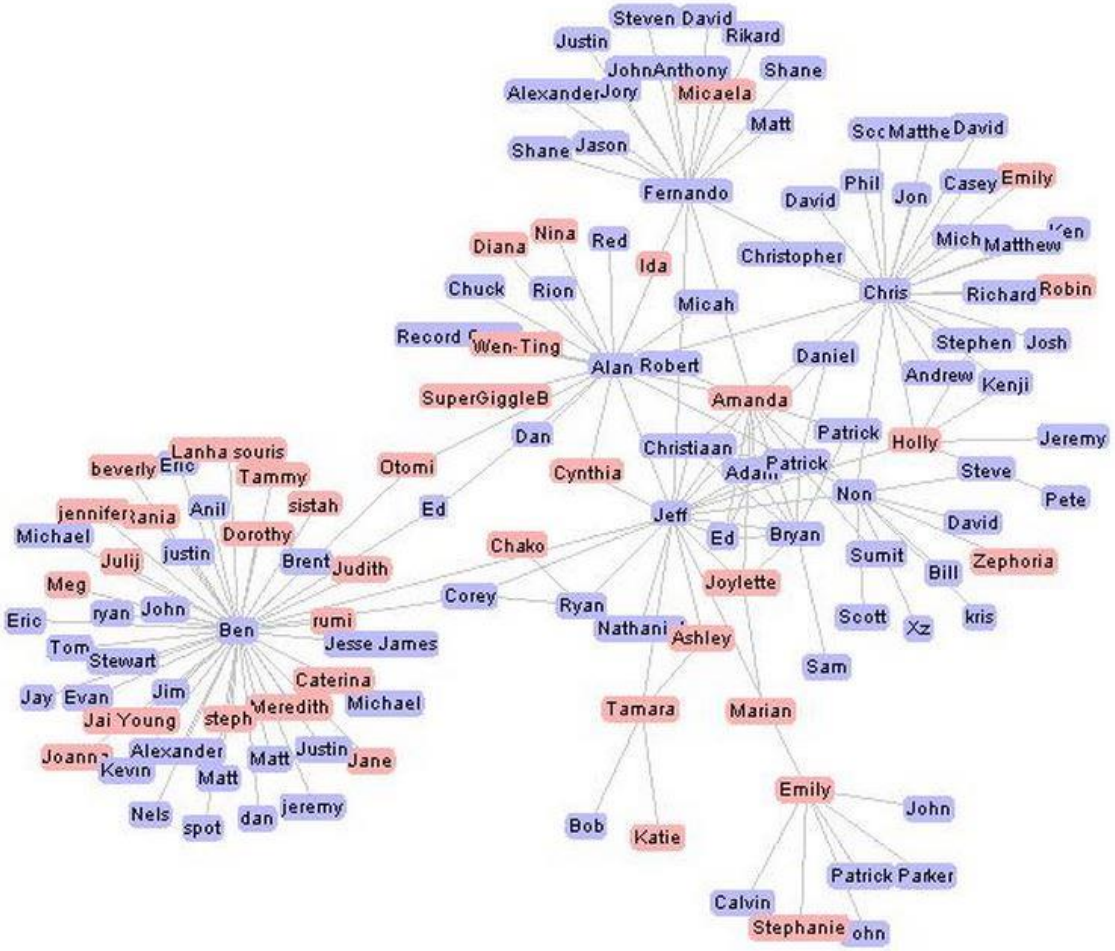
Figura 18 - *Word Tree* gerada pelo Many Eyes.



Fonte: http://www-01.ibm.com/software/analytics/many-eyes/.

Prefuse é um conjunto de ferramentas de software para a criação de ricas visualizações de dados iterativas. Trata-se de um kit de ferramentas desenvolvido na linguagem de programação JAVA, e disponível como código aberto, podendo ser utilizado para fins comerciais e não comerciais. A Figura 19 demonstra um gráfico de informações de uma rede social gerado pela ferramenta.

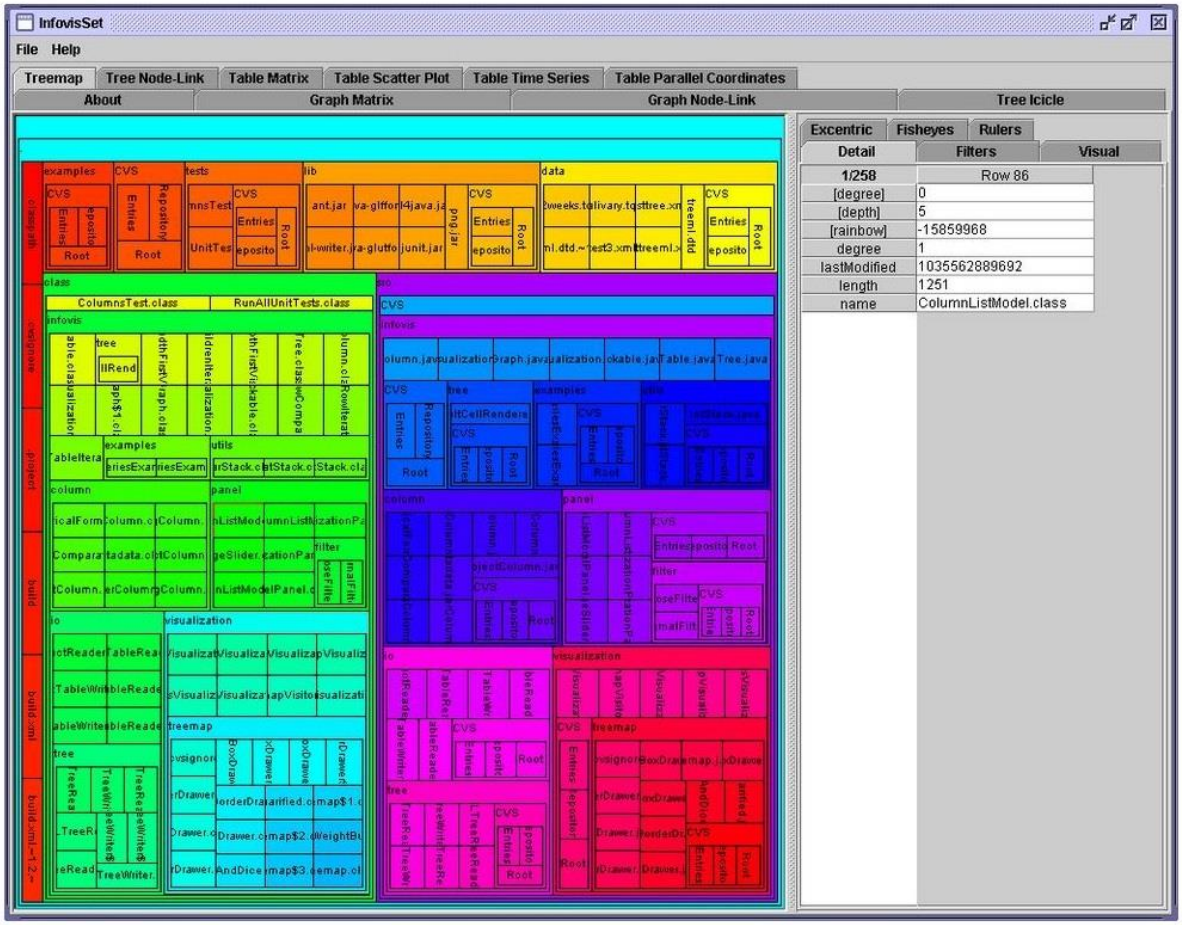
Figura 19 - Visualização de uma rede social do toolkit Prefuse.



Fonte: http://prefuse.org/.

Em seu trabalho, Cini (2013) apresenta uma ferramenta para a geração de visualizações em *Treemap*, chamada InfoVis Toolkit. Esta ferramenta é um conjunto de ferramentas gráficas interativas escrito em Java/Swing. Esta ferramenta é mantida por Jean-Daniel Fekete e colaboradores (Fekete, 2013). A Figura 20 é um exemplo de *Treemap* encontrado no site do desenvolvedor.

Figura 20 - Treemap gerado pela ferramenta InfoVis.



Fonte: (Fekete, 2013).

2.4 CONSIDERAÇÕES FINAIS

Este capítulo teve como objetivo conceituar os assuntos relacionados à Mineração de Textos e à Visualização de Informação. No que se refere à Mineração de Textos, foi apresentado o processo, que visa extrair padrões em dados textuais analisados. Também foram apresentadas aplicações já existentes, com o objetivo de demonstrar o que já foi estudado. Quanto à Visualização de Informação, foi apresentado o modelo referência para a construção de visualizações, assim como técnicas existentes para estas visualizações. Ainda, seguindo no assunto sobre visualizações, apresentou-se técnicas para visualização textual e algumas aplicações já existentes.

Pode-se notar que as ferramentas de visualização textuais chegam mais próximas ao objetivo proposto neste trabalho, tendo em vista a comparação de textos educacionais. No que diz respeito aos métodos de visualização, onde a técnica contexto se mostra mais aplicável ao universo textual, é de extrema importância na avaliação das ferramentas utilizadas para visualização.

Com o estudo teórico realizado neste capítulo, é possível avaliar a ferramenta UnderMine Text Miner desenvolvida por Cristofoli (2012) para a mineração de textos (Capítulo 3), e avaliar as ferramenta para apresentação de dados textuais para os resultados da mineração.

3 FERRAMENTA UNDERMINE TEXT MINER

UnderMine Text Miner é uma ferramenta desenvolvida por Cristofoli (2012) no trabalho de conclusão de curso (Mineração de Textos Baseada em Algoritmos Imunológicos). Esta ferramenta foi desenvolvida baseada em sistemas imunológicos artificiais. Segundo Dasgupta (1999), essa é uma área da Inteligência Artificial formada por metodologias inteligentes, baseadas no funcionamento do sistema imunológico dos seres vertebrados para a solução de problemas reais.

A ferramenta UnderMine Text Miner é composta por duas etapas. Na primeira etapa é realizado o treinamento do algoritmo que servirá como base de dados para a avaliação dos textos. A segunda etapa, denominada diagnóstico, consiste na execução do algoritmo baseado em sistemas imunológicos para a mineração dos textos. As seções seguintes detalham essas duas etapas.

3.1 ETAPA DE TREINAMENTO

A etapa de treinamento tem como objetivo realizar a leitura dos artigos de diversos assuntos e selecionar os termos que possuem maior relevância para cada tema, calculando sua frequência. Os artigos são inseridos separados pelo tema, treinando separadamente cada tema. Para isso os artigos são separados em diretórios por área de estudo, deixando o usuário responsável por selecionar um diretório por vez para o treinamento, conforme a área de estudo que está sendo treinada.

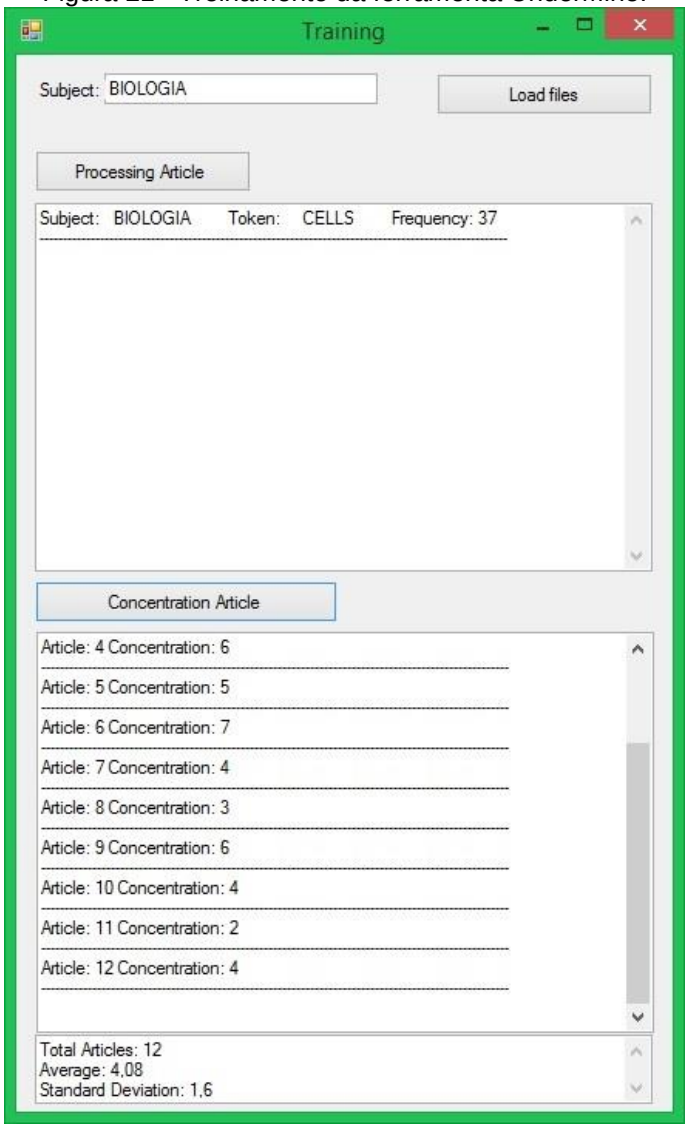
Após realizado o treinamento de um dos grupos de estudo, será criado um arquivo com os termos chamados de “*Tokens*”, com a frequência em que aparece nos artigos (Tabela 4). Como o treinamento é feito separadamente para cada área de estudo, será gerado um arquivo para cada.

Tabela 4 - Termos gerados pelo sistema de treinamento do UnderMine.

Termos de Biologia	**Termos de Música**
37|CELLS|BIOLOGIA	78|MUSIC|MUSICA
26|CELL|BIOLOGIA	17|PERFORMANCE|MUSICA
23|PROTEIN|BIOLOGIA	12|EMOTIONS|MUSICA
21|APOPTOSIS|BIOLOGIA	9|GENRES|MUSICA
17|INDUCED|BIOLOGIA	8|EFFECT|MUSICA
16|STRESS|BIOLOGIA	6|LISTENING|MUSICA
14|BCL|BIOLOGIA	5|ABILITY|MUSICA

Fonte: Cristofoli (2012).

Figura 21 - Treinamento da ferramenta UnderMine.



Fonte: Imagem da ferramenta UnderMine desenvolvida por Cristofoli (2012).

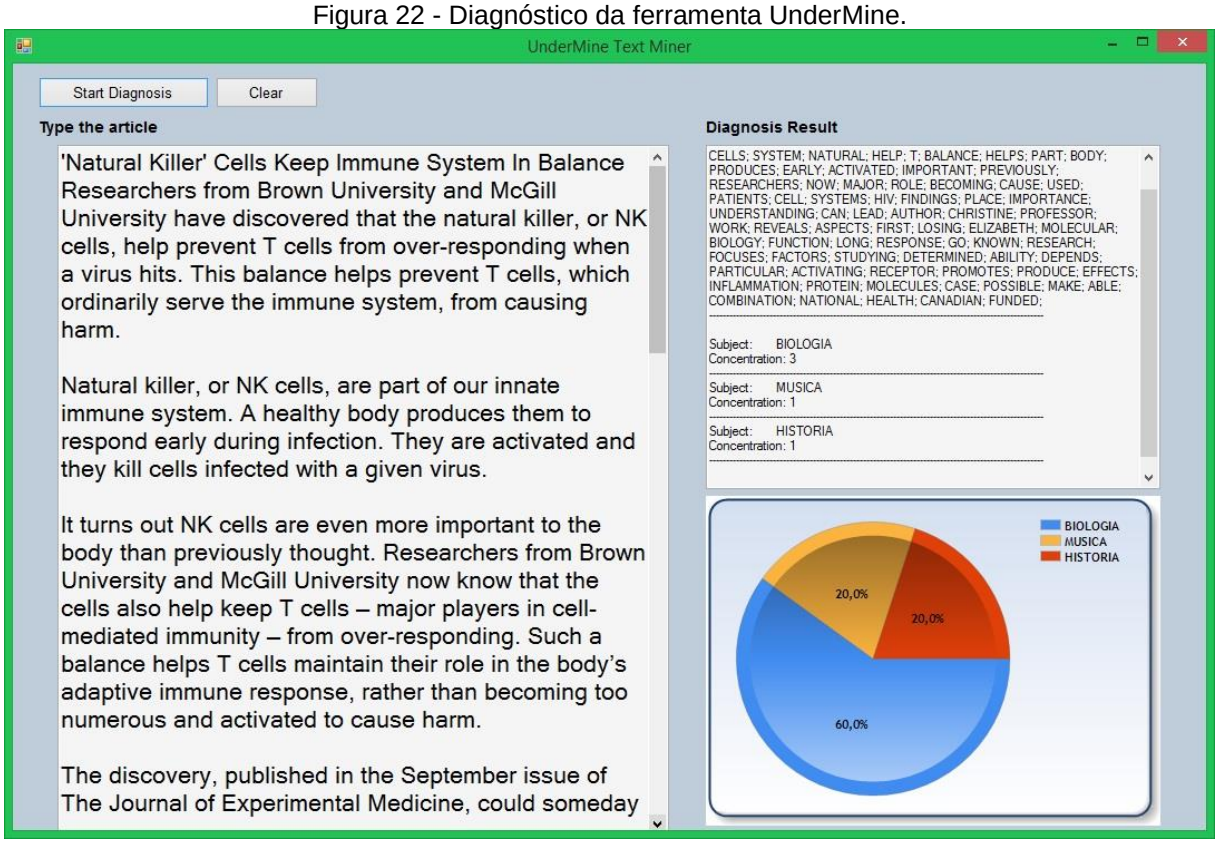
A Figura 21 apresenta a interface do sistema de treinamento da ferramenta. O botão “*Load files*” é utilizado para informar o diretório que estão localizados os artigos. Existe um diretório para cada área de estudo, sendo que apenas um por vez é informado e importado. No campo “*Subject*” é informado o assunto ao qual os artigos que serão importados se referem. Ao utilizar o botão “*Processing Article*” o sistema efetua a leitura dos artigos e calcula a frequência dos termos. No botão “*Concentration Article*” é calculada a concentração dos termos de maior frequência, que consiste em calcular o limiar do “*Token*”, o que indicará a proximidade que um artigo possui de cada área de estudo aprendida no treinamento. A fórmula do cálculo das concentrações é a soma das frequências dos termos sobre o número total de termos em comum entre o treinamento e o mesmo artigo. A concentração para cada artigo é apresentada ao usuário, assim como o total de artigos e, a média e o desvio padrão dos mesmos.

3.2 ETAPA DE DIAGNÓSTICO

A etapa de diagnóstico tem como objetivo executar o algoritmo desenvolvido conforme os sistemas imunológicos. Uma vez que a etapa de treinamento criou os anticorpos ao separar os “*tokens*” de cada domínio, é submetido o antígeno, que é o artigo bruto, para o diagnóstico. O resultado é a classificação do artigo, conforme as áreas de estudo que o sistema foi treinado na primeira etapa.

A Figura 22 apresenta a interface do sistema de diagnóstico onde é inserido o artigo para avaliação. Ao selecionar o botão “*Start Diagnosis*” a leitura do artigo é iniciada e o sistema separa cada “*Token*”, verificando se já foi aprendido e armazenando esta informação em uma lista. Ao lado, no quadro “*Dagnosis Result*” é apresentada esta lista de “*Tokens*” e a concentração que o artigo possui para cada domínio. Para uma visualização mais intuitiva é apresentado um gráfico de pizza com a porcentagem da proximidade de cada domínio.

Figura 22 - Diagnóstico da ferramenta UnderMine.



Fonte: Imagem da ferramenta UnderMine desenvolvida por Cristofoli (2012).

3.3 CONSIDERAÇÕES FINAIS

Conforme apresentado, a ferramenta UnderMine Text Miner tem por objetivo implementar o algoritmo imunológico para a mineração de textos, onde o objetivo é a classificação de artigos conforme a área de estudo e consequentemente testar a sua eficiência para tal. Este algoritmo foi desenvolvido utilizando a linguagem de programação C#. O presente trabalho abordará a mineração de textos voltada a textos educacionais, utilizando o algoritmo proposto por Cristofoli (2012) para verificar o aprendizado de alunos perante a determinado texto proposto pelo professor.

4 FERRAMENTA ANÁLISE DE PRODUÇÕES TEXTUAIS

O presente capítulo tem por objetivo apresentar a ferramenta Análise de Produções Textuais. É uma ferramenta capaz de efetuar o treinamento sobre determinada área do conhecimento, efetuar a análise de textos redigidos por alunos e apresentar os resultados.

Tendo em vista a utilização da mineração de textos para a análise dos textos, foi utilizado o algoritmo contido no software UnderMine Text Miner, uma ferramenta desenvolvida por Cristofoli (2012) no trabalho de conclusão de curso (Mineração de Textos Baseada em Algoritmos Imunológicos). Após os dados minerados, um conjunto de ferramentas foi utilizado para a apresentação dos dados para uma análise por parte do professor.

4.1 MODELAGEM

No trabalho de conclusão de curso desenvolvido por Cristofoli (2012) o objetivo era o treinamento do algoritmo sobre determinada área de estudo e a classificação dos artigos analisados. Esta análise apresentava o resultado em um gráfico de “*pizza”* conforme a porcentagem da pontuação do texto analisado com o texto original. Este sistema possuía duas etapas que eram executadas separadamente: Treinamento, encarregada da aprendizagem dos *tokens* dos artigos utilizados e a etapa de Classificação, responsável pela classificação do artigo em cada um dos domínios testados. A arquitetura de modelagem utilizava três camadas: camada de interface, camada de dados e camada de verificadores do sistema imune.

Seguindo a lógica do software proposto por Cristofoli (2012), o sistema Análise de Produções Textuais foi desenvolvido contendo duas etapas: Treinamento, responsável pela aprendizagem dos *tokens* dos artigos usados para treinar determinada atividade e Análise, encarregada da análise dos textos dos alunos e apresentação dos resultados visuais. A diferença é que estas duas etapas são apresentas na mesma aplicação, separadas por abas. Quando há a

necessidade de um novo treinamento, o mesmo pode ser realizado, caso contrário, pode ser realizada a análise dos textos a qualquer momento.

O sistema análise de produções textuais implementa uma camada a mais do que o UnderMine Text Miner desenvolvido por Cristofoli (2012), a camada de visualização, responsável por estruturar os dados e apresenta-los em tela. Além de outras alterações nas camadas já existentes que serão descritas no capítulo 4.3. A Tabela 5 apresenta as classes do sistema.

Tabela 5 - Classes do sistema Análise de Produções Textuais. Adaptado do sistema UnderMine Text Miner (CRISTOFOLI, 2012).

Nome da Classe	**Objetivos**
Treinamento	
Mensagem	- Representa os atributos do objeto do tipo Mensagem.
Token	- Representa os atributos do objeto do tipo Token.
Arquivo	- Classe responsável pela leitura dos textos para treinamento do sistema.
Artigo	- Classe centralizadora das lógicas do treinamento.
StopWord	- Classe responsável por remover as StopWords dos textos.
Análise	
AgenteDecompositor	- Decompor o artigo em *tokens*. - Gerar Listas com os termos decompostos
VerificadorConcentracaoTokens	Calcular a concentração dos *tokens* no artigo.
AgenteVerificadorDiagnostico	- Utilizando os dados das verificações anteriores, realiza o diagnóstico do artigo.
AnaliseProdText	- Classe de interface do programa principal.
AlteraConfig	- Classe de interface para configuração do sistema.
Diagnosticar	- Classe que carrega os dados necessários para o sistema e inicializa os agentes.

Análise	
QuadroNegro	- Classe onde todos os agentes escrevem informações. Essa classe é acessível para todo o sistema.
Artigo	- Classe que representa os atributos de uma produção textual de um aluno.
ListaArtigo	- Classe que armazena todas as produções textuais dos alunos a serem analisadas após terem sido lidas pelo sistema.
CarregarDados	- Classe responsável pela leitura das produções textuais dos alunos.
Termos	- Representa os atributos do objeto do tipo Termo.
Token	- Representa os atributos de um objeto do tipo Token.
ResultadoDiagnóstico	- Representa os atributos de um objeto do tipo ResultadoDiagnóstico. Nesta classe ficam armazenados os resultados da análise em tempo de execução.
Visualizações	
BarChart	- Classe responsável por estruturar os dados necessários para a visualização do tipo Bar Chart e efetuar a chamada da respectiva visualização.
BubbleChart	- Classe responsável por estruturar os dados necessários para a visualização do tipo Bubble Chart e efetuar a chamada da respectiva visualização.
CollapsibleTree	- Classe responsável por estruturar os dados necessários para a visualização do tipo Collapsible Tree ou Rotating Cluster e efetuar a chamada da respectiva visualização.

Visualizações	
TreeMap	- Classe responsável por estruturar os dados necessários para a visualização do tipo TreeMap e efetuar a chamada da respectiva visualização.
WordCloud	- Classe responsável por estruturar os dados necessários para a visualização do tipo Word Cloud e efetuar a chamada da respectiva visualização.

No anexo A será apresentado o diagrama de classes para o sistema, assim podendo ser visto de forma mais detalhada os atributos de cada classe.

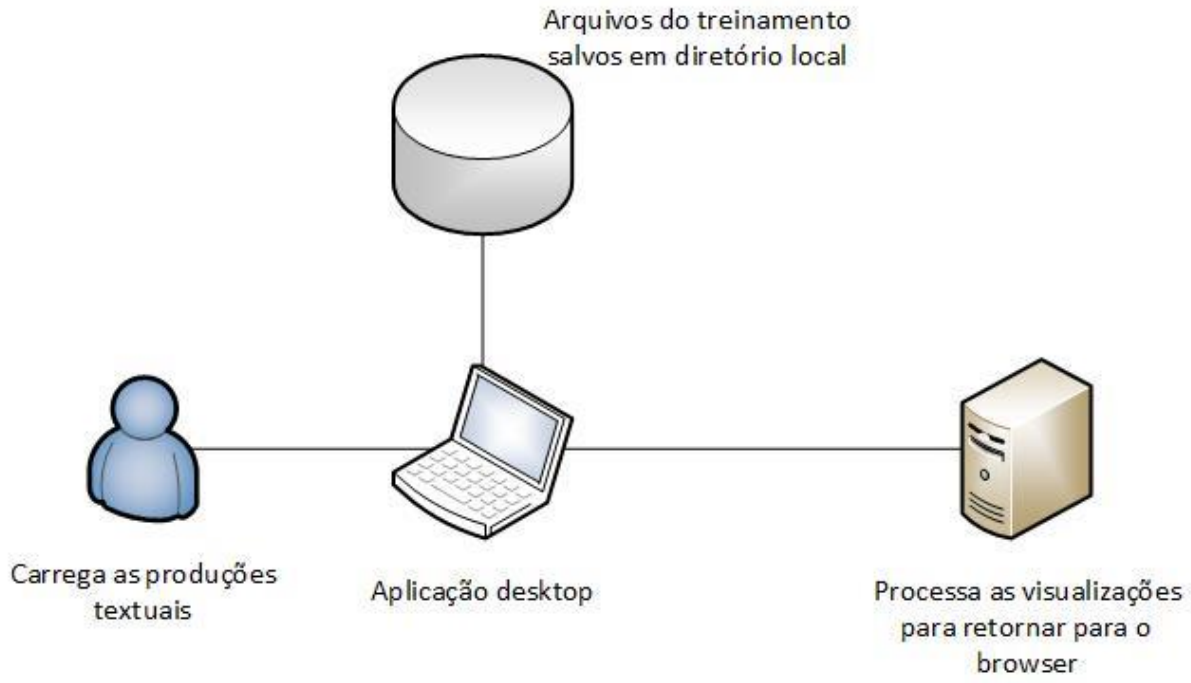
4.2 ARQUITETURA

Visando a utilização do algoritmo de mineração de textos da ferramenta UnderMine Text Miner (CRISTOFOLI, 2012), a ferramenta Análise de Produções Textuais foi desenvolvida para desktop, integrando visualizações desenvolvidas em Java Script. Existindo esta integração entre desktop e web, foi necessária a utilização de um servidor web.

Ao ser inicializado, o sistema funciona totalmente como desktop, tanto a funcionalidade de treinamento da aplicação quanto a análise das produções textuais. Após a análise dos textos na parte desktop, desenvolvida em C#, as visualizações são disponibilizadas no Browser padrão da máquina, visto que estas visualizações são desenvolvidas em JavaScript, usando a tecnologia D3 Data-Driven Documents (BOSTOCK, 2013).

A comunicação entre a camada desktop e as visualizações web é feita através de arquivos no formato *JSON*, estes arquivos são salvos no diretório de publicação do servidor web, assim podem ser lidos pelas aplicações desenvolvidas para a camada web. A Figura 23 representa a arquitetura do sistema.

Figura 23 - Arquitetura sistema Análise de Produções Textuais.



4.3 IMPLEMENTAÇÃO

Este capítulo apresenta a implementação da ferramenta Análise de Produções Textuais. A descrição da implementação contempla o que precisou ser alterado com relação ao software Undermine Text Miner (CRISTOFOLI, 2012), tanto no treinamento, quando na análise de textos. Também irá descrever o novo módulo para geração de visualizações.

O sistema Análise de Produções Textuais foi desenvolvido com o intuito de deixar acessível ao usuário todas as funcionalidades em apenas uma tela. Esta tela é separada em abas, que dividem a etapa de treinamento da etapa de análise das produções textuais. Sendo assim, o usuário pode efetuar algum treinamento antes de analisar as produções textuais ou poderá utilizar a etapa de análise independentemente do treinamento. Porém, para que a análise seja efetiva, o treinamento do assunto relacionado deve ter sido efetuado.

4.3.1 Treinamento

A etapa de treinamento foi alterada, principalmente, quanto ao idioma que a ferramenta suporta. Anteriormente o sistema apenas aceitava textos em inglês. Foi alterado para, além de suportar textos em inglês, suportar textos em português. Para isto, uma alteração significativa é quando às *Stop Words,* que são removidas no processo de treinamento. Foi alterada a classe StopWord, incluindo as *Stop Words* em português (Tabela 6).

A tela também foi alterada para melhor usabilidade por parte do professor. Foi incluído um campo que informa os textos que foram importados a partir do diretório informado. Também informa o final do treinamento com uma mensagem “Treinamento Encerrado!”. Adicionado um campo que informa as atividades já treinadas. Desta forma informando o históricos dos treinamentos. Devido à alteração do idioma do treinamento, foi incluído um “*Radio Button*” para poder informar o idioma que o treinamento será efetuado. Lembrando que o texto utilização para o treinamento deve estar no idioma que foi selecionado.

A funcionalidade do programa é conforme a ordem que se pode visualizar na Figura 24, informa-se a atividade, o diretório que contém os textos que serão utilizados no treinamento e o idioma. Clicando no botão “Processar Artigos” o processo será executado.

Figura 24 - Treinamento da ferramenta Análise de Produções Textuais.

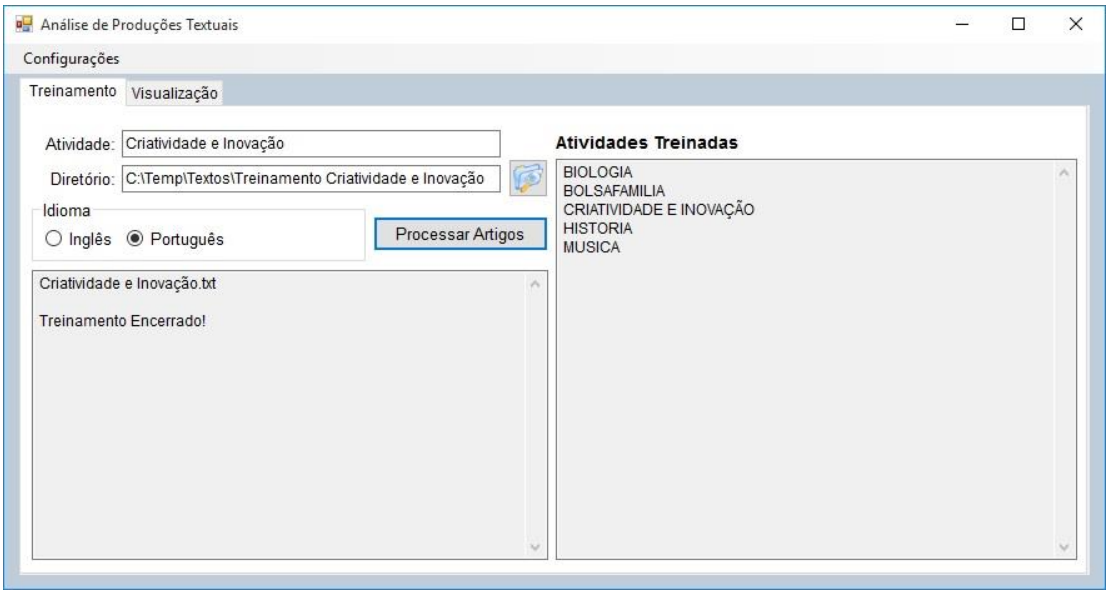


Tabela 6 - Lista de Stop Words em Português.

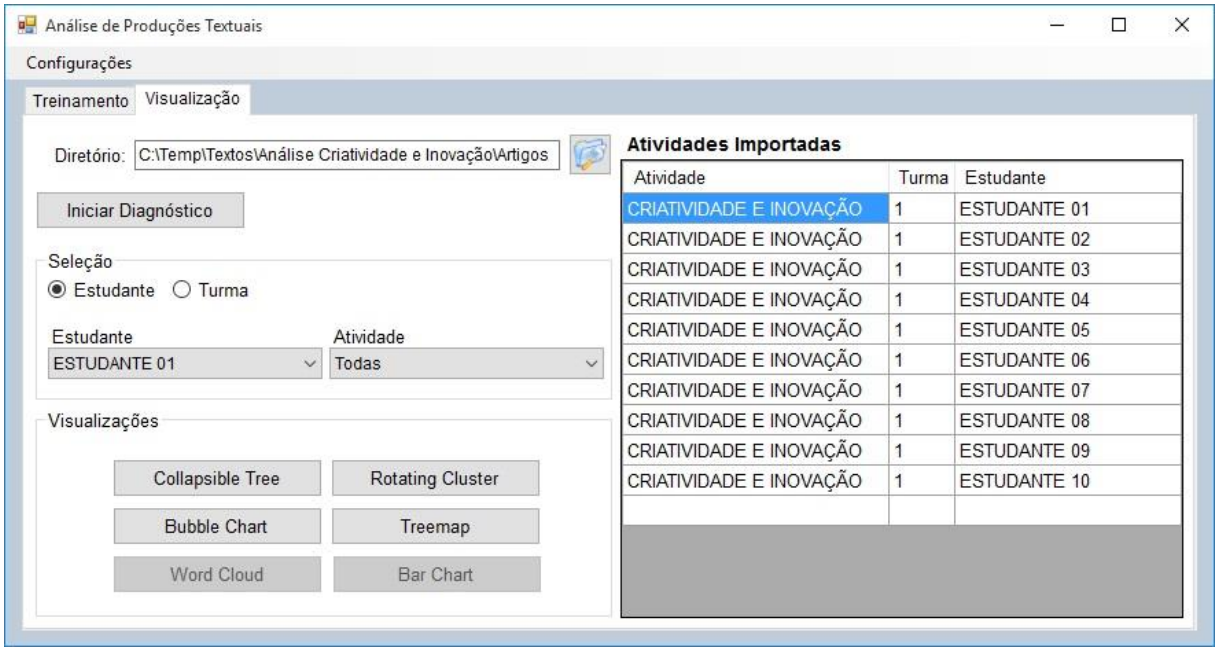
Lista de Stop Words					
A	DESSAS	ESTÃO	MUITO	PODER	SUA
AGORA	DESSE	ESTAS	MUITOS	PODERIA	SUAS
AINDA	DESSES	ESTAVA	NA	PODERIAM	TALVEZ
ALGUÉM	DESTA	ESTAVAM	NÃO	PODIA	TAMBÉM
ALGUM	DESTAS	ESTÁVAMOS	NAS	PODIAM	TAMPOUCO
ALGUMA	DESTE	ESTE	NEM	POIS	TE
ALGUMAS	DESTE	ESTES	NENHUM	POR	TEM
ALGUNS	DESTES	ESTOU	NESSA	PORÉM	TENDO
AMPLA	DEVE	EU	NESSAS	PORQUE	TENHA
AMPLAS	DEVEM	FAZENDO	NESTA	POSSO	TER
AMPLO	DEVENDO	FAZER	NESTAS	POUCA	TEU
AMPLOS	DEVER	FEITA	NINGUÉM	POUCAS	TEUS
ANTE	DEVERÁ	FEITAS	NO	POUCO	TI
ANTES	DEVERÃO	FEITO	NOS	POUCOS	TIDO
AO	DEVERIA	FEITOS	NÓS	PRIMEIRO	TINHA
AOS	DEVERIAM	FOI	NOSSA	PRIMEIROS	TINHAM
APÓS	DEVIA	FOR	NOSSAS	PRÓPRIA	TODA
AQUELA	DEVIAM	FORAM	NOSSO	PRÓPRIAS	TODAS
AQUELAS	DISSE	FOSSE	NOSSOS	PRÓPRIO	TODAVIA
AQUELE	DISSO	FOSSEM	NUM	PRÓPRIOS	TODO
AQUELES	DISTO	GRANDE	NUMA	QUAIS	TODOS
AQUILO	DITO	GRANDES	NUNCA	QUAL	TU
AS	DIZ	HÁ	O	QUANDO	TUA
ATÉ	DIZEM	ISSO	OS	QUANTO	TUAS
ATRAVÉS	DO	ISTO	OU	QUANTOS	TUDO
CADA	DOS	JÁ	OUTRA	QUE	ÚLTIMA
COISA	E	LA	OUTRAS	QUEM	ÚLTIMAS
COISAS	É	LÁ	OUTRO	SÃO	ÚLTIMO
COM	ELA	LHE	OUTROS	SE	ÚLTIMOS
COMO	ELAS	LHES	PARA	SEJA	UM
CONTRA	ELE	LO	PELA	SEJAM	UMA
CONTUDO	ELES	MAS	PELAS	SEM	UMAS
DA	EM	ME	PELO	SEMPRE	UNS
DAQUELE	ENQUANTO	MESMA	PELOS	SENDO	VENDO
DAQUELES	ENTRE	MESMAS	PEQUENA	SERÁ	VER
DAS	ERA	MESMO	PEQUENAS	SERÃO	VEZ
DE	ESSA	MESMOS	PEQUENO	SEU	VINDO
DELA	ESSAS	MEU	PEQUENOS	SEUS	VIR
DELAS	ESSE	MEUS	PER	SI	VOS
DELE	ESSES	MINHA	PERANTE	SIDO	VÓS
DELES	ESTA	MINHAS	PODE	SÓ	
DEPOIS	ESTÁ	MUITA	PUDE	SOB	
DESSA	ESTAMOS	MUITAS	PODENDO	SOBRE	

4.3.2 Análise das Produções Textuais

A etapa de análise das produções textuais é a que mais sofreu alterações para ser adequada aos objetivos propostos. Uma vez que o objetivo é a análise de diversas produções textuais de alunos e a geração de visualizações para a análise por parte do professor. A tela responsável pela análise das produções textuais foi criada para a aplicação Análise de Produções Textuais. Nesta etapa, apenas o algoritmo de mineração de textos foi utilizado da aplicação Undermine Text Miner (CRISTOFOLI, 2012).

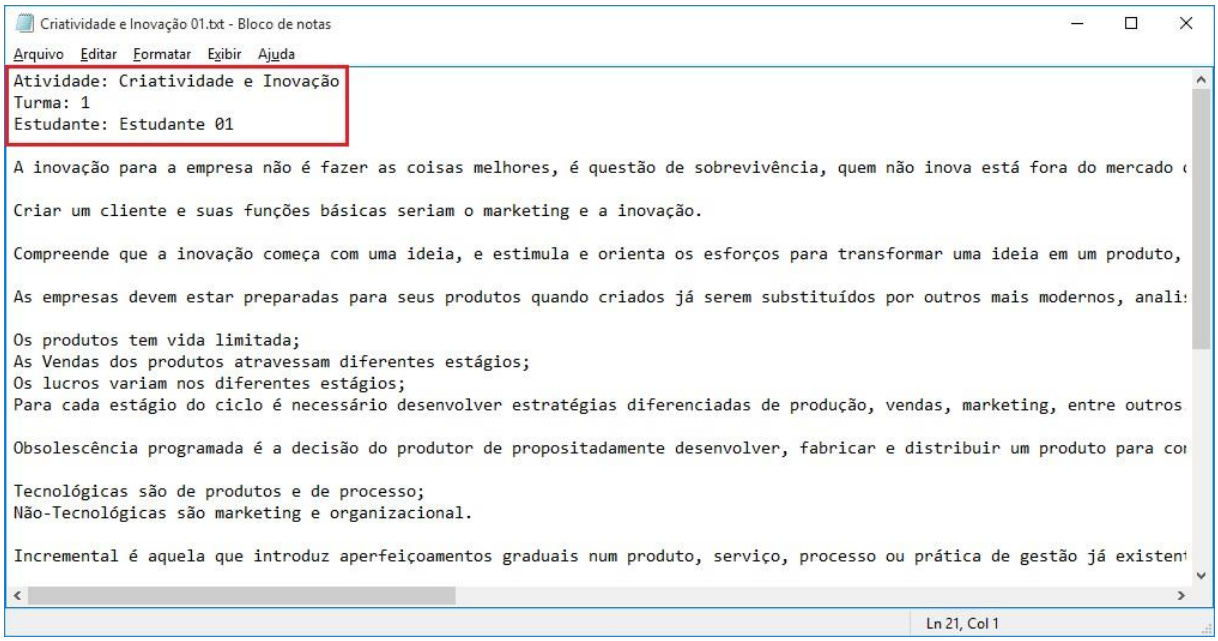
Conforme pode ser visto na Figura 25, foi incluído um campo “Diretório”, onde se pode informar o diretório das produções textuais dos alunos. Ao clicar no botão “Iniciar Diagnóstico”, o programa importará todos os arquivos e irá apresentar no campo “Atividades Importadas” todos os artigos que importou, informando a atividade, a turma e o aluno. Após a importação, também são preenchidos os campos do tipo “*combo-box*”, na “Seleção”. Ficarão disponíveis para combinar a análise que se deseja ser feita. Podendo visualizar um estudante em uma atividade ou em todas as atividades ou uma turma em uma atividade ou todas as atividades.

Figura 25 - Tela de análise da ferramenta Análise de Produções Textuais.



Para o processo de diagnóstico, antes de se gerar alguma visualização, foram necessárias algumas alterações quanto a importação e a execução da mineração de textos para diversos arquivos. Foi estipulado um cabeçalho para o arquivo, com informações da turma, nome do estudante e atividade. Esta definição foi necessária para que o programa conseguisse se estruturar internamente e classificar cada uma das produções textuais. O cabeçalho no arquivo é demonstrado na Figura 26. Para poder criar esta organização internamente, foram criadas as classes “Artigo”, que representa uma produção textual, e “ListaArtigos”, que representa a coleção de todas as produções textuais que estão sendo analisadas. Estas novas classes foram inclusas na camada de dados do sistema.

Figura 26 - Exemplo de produção textual para importação.



4.3.3 Visualização dos Resultados

Após a execução da análise, o sistema habilita a “Seleção” e as “Visualizações” na aba “Visualização”. No total, são seis visualizações disponíveis. Para criar as visualizações, foram implementadas cinco classes no projeto. A partir de cada classe será chamada a camada de visualização que utiliza a tecnologia D3 Data-Driven Documents (BOSTOCK, 2013). As classes desenvolvidas foram:

a) BarChart – Cria o arquivo TSV para a visualização Bar Chart.
b) BubbleChart – Cria o arquivo JSON para a visualização Bubble Chart.
c) CollapsibleTree – Cria o arquivo JSON para a visualização Collapsible Tree e Rotating Cluster.
d) TreeMap – Cria o arquivo JSON para a visualização Treemap.
e) WordCloud – Cria o arquivo JSON para a visualização Word Cloud.

Cada classe é responsável por criar o arquivo, responsável pela comunicação, para a visualização de mesmo nome. Apenas a classe CollapsibleTree é responsável por criar, também, a visualização RotatingCluster. Apenas a classe BarChart cria um arquivo no formato diferente de JSON, que é o TSV.

As classes que geram as visualizações são responsáveis por fazer a chamada da visualização que abrirá no browser padrão da máquina. As visualizações estão disponibilizadas no diretório de publicação do servidor web. As visualizações seguem o mesmo padrão de nomenclatura utilizando o nome da visualização, são elas:

a) bar-chart – Visualização Bar Chart.
b) bubble-chart – Visualização Bubble Chart.
c) cooalpsible-tree – Visualização Collapsible Tree.
d) rotating-cluster – Visualização Rotating Cluster.
e) treemap – Visualização Treemap.
f) word-cloud – Visualização Word Cloud.

No anexo B será apresentado o manual do produto que explica de forma detalhada a instalação e utilização do sistema Análise de Produções Textuais.

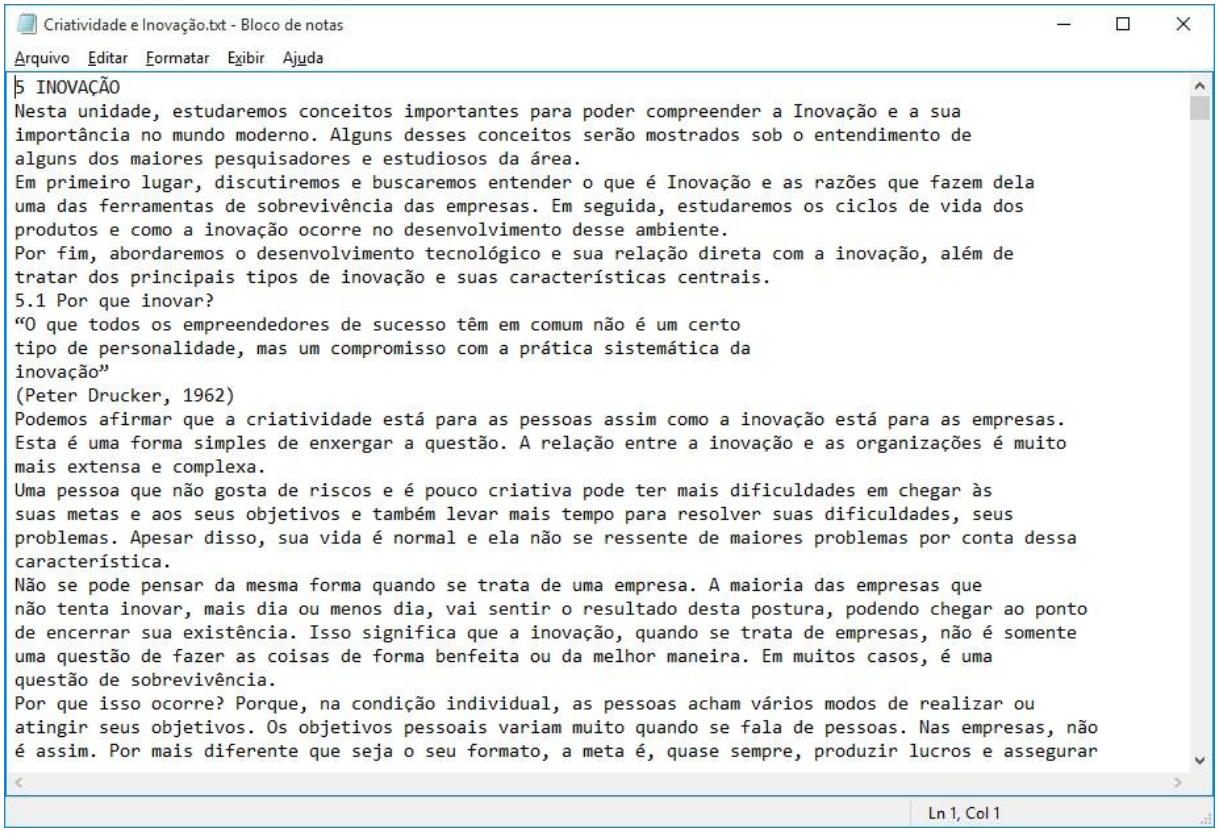
4.4 TESTES

A sessão de testes visa apresentar o funcionamento do sistema Análise de Produções Textuais e os testes realizados. O objetivo é apresentar todas as etapas

da utilização do sistema e todas as possibilidades de geração de visualizações para análise dos resultados da mineração das produções textuais.

Para este cenário de teste foram utilizados seis textos para o treinamento da ferramenta abordando os assuntos: Criatividade e Inovação, Polietileno de baixa densidade (PEBD ou LDPE), Polietileno de alta densidade (PEAD ou HDPE), Polietileno linear de baixa densidade (PELBD ou LLDPE), Polietileno de ultra alto peso molecular (PEUAPM ou UHMWPE) e Polietileno de ultra baixa densidade (PEUBD ou ULDPE). A Figura 27 demonstra um arquivo utilizado para treinar a ferramenta, neste caso apresentando o assunto Criatividade e Inovação.

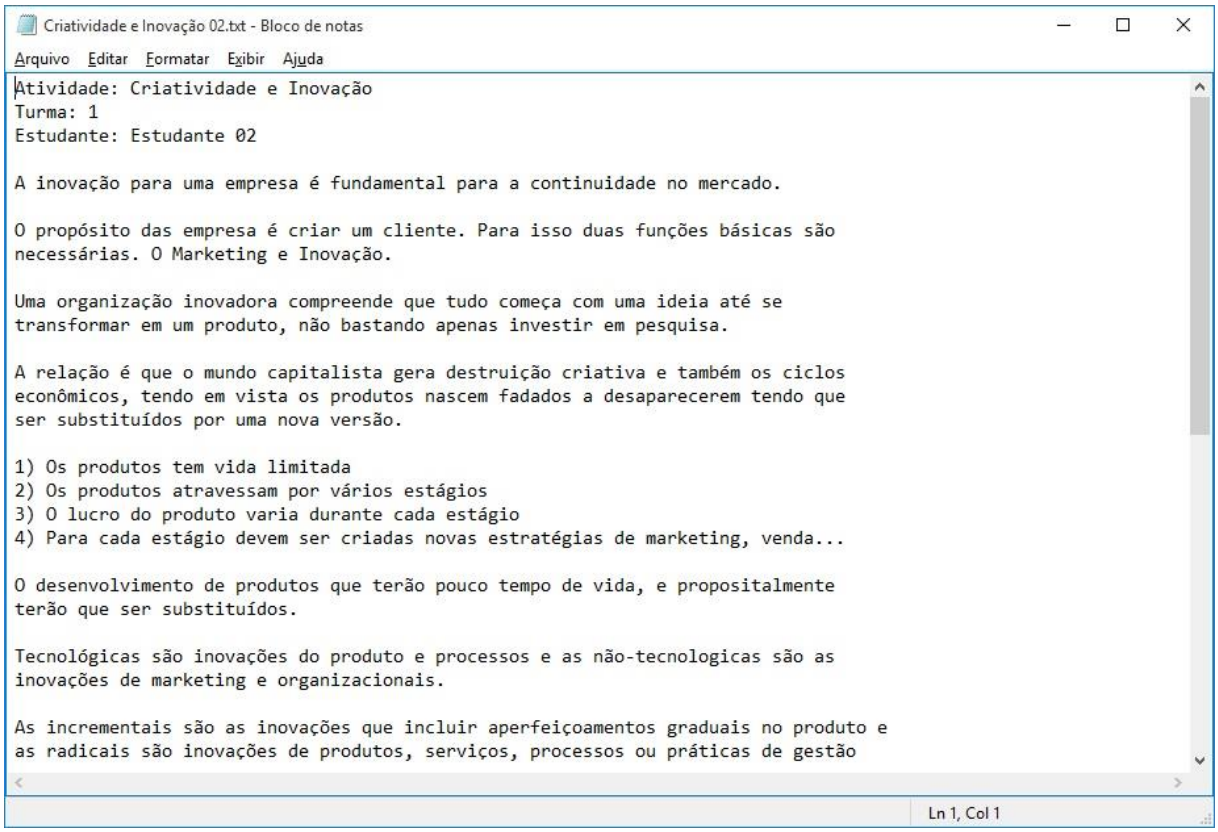
Figura 27 - Texto utilizado para treinamento de Criatividade e Inovação.



Após o treinamento da ferramenta com os textos dos assuntos desejados, são submetidas as produções textuais dos alunos. A Figura 28 apresenta um exemplo de produção textual utilizado para submeter à análise da ferramenta. Destacando o detalhe do cabeçalho do arquivo, o qual identifica de qual atividade este texto se refere, a turma que o aluno está inserido e o nome do estudante. Estas

informações são essenciais para que a ferramenta possa classificar o texto e criar as visualizações.

Figura 28 - Exemplo de produção textual para análise.



4.4.1 Testes Turma de Informação e Decisão

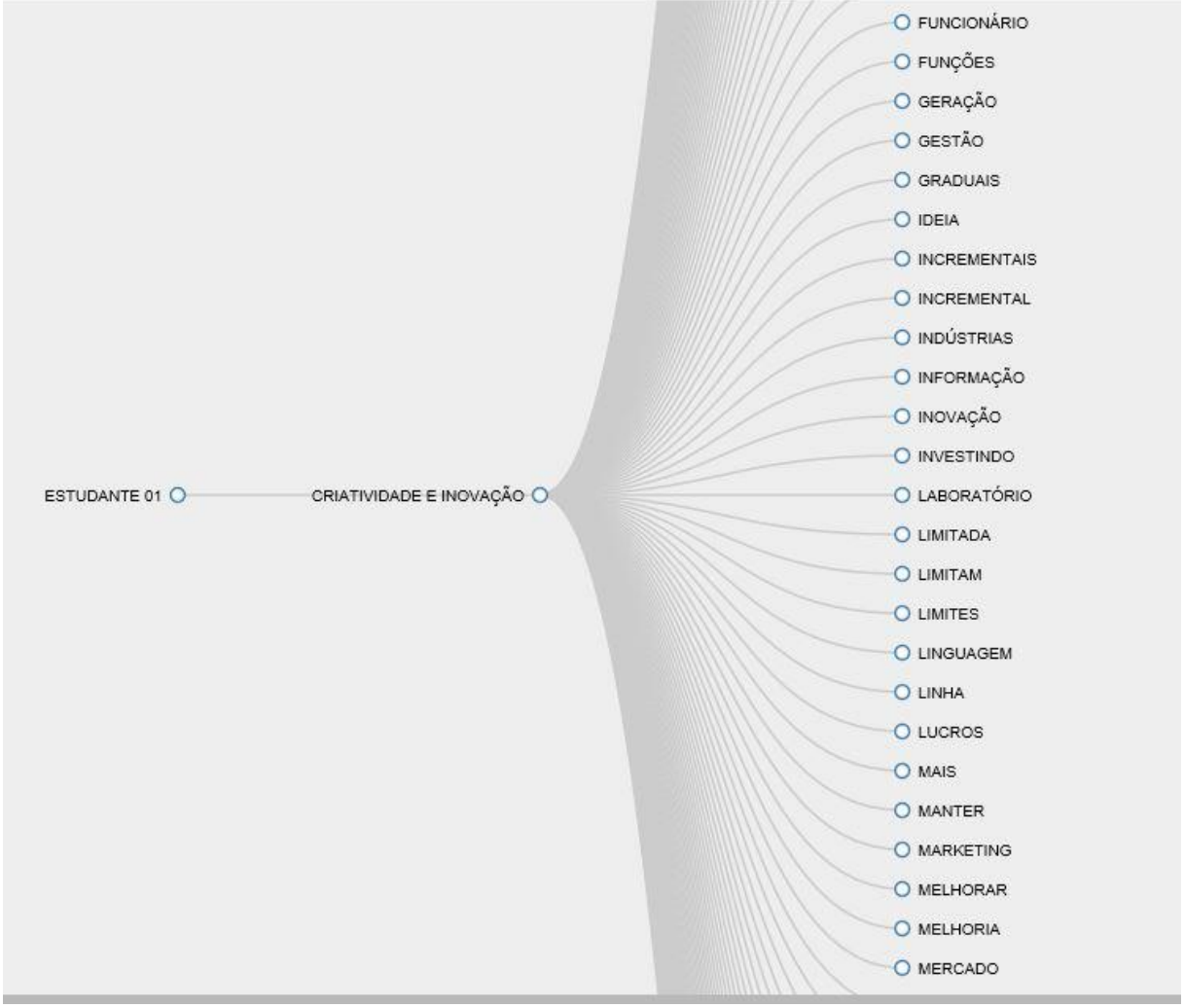
Para avaliação da ferramenta em desenvolvimento foram utilizados textos de uma turma da disciplina de Sistemas de Informação e Decisão. Esta turma teve como atividade a leitura do assunto Criatividade e Inovação, seguido pela resolução de um questionário. O sistema Análise de Produções Textuais foi treinado sobre o assunto e o questionário dos alunos foi submetido à análise. Como esta atividade foi aplicada em apenas uma turma, as visualizações geradas são de um estudante para uma atividade e da turma toda para uma atividade.

4.4.1.1 Um Estudante para Uma Atividade

Esta seção apresentará as visualizações que se pode obter combinando um estudante em uma atividade, apresentando o que pode ser analisado em cada uma delas.

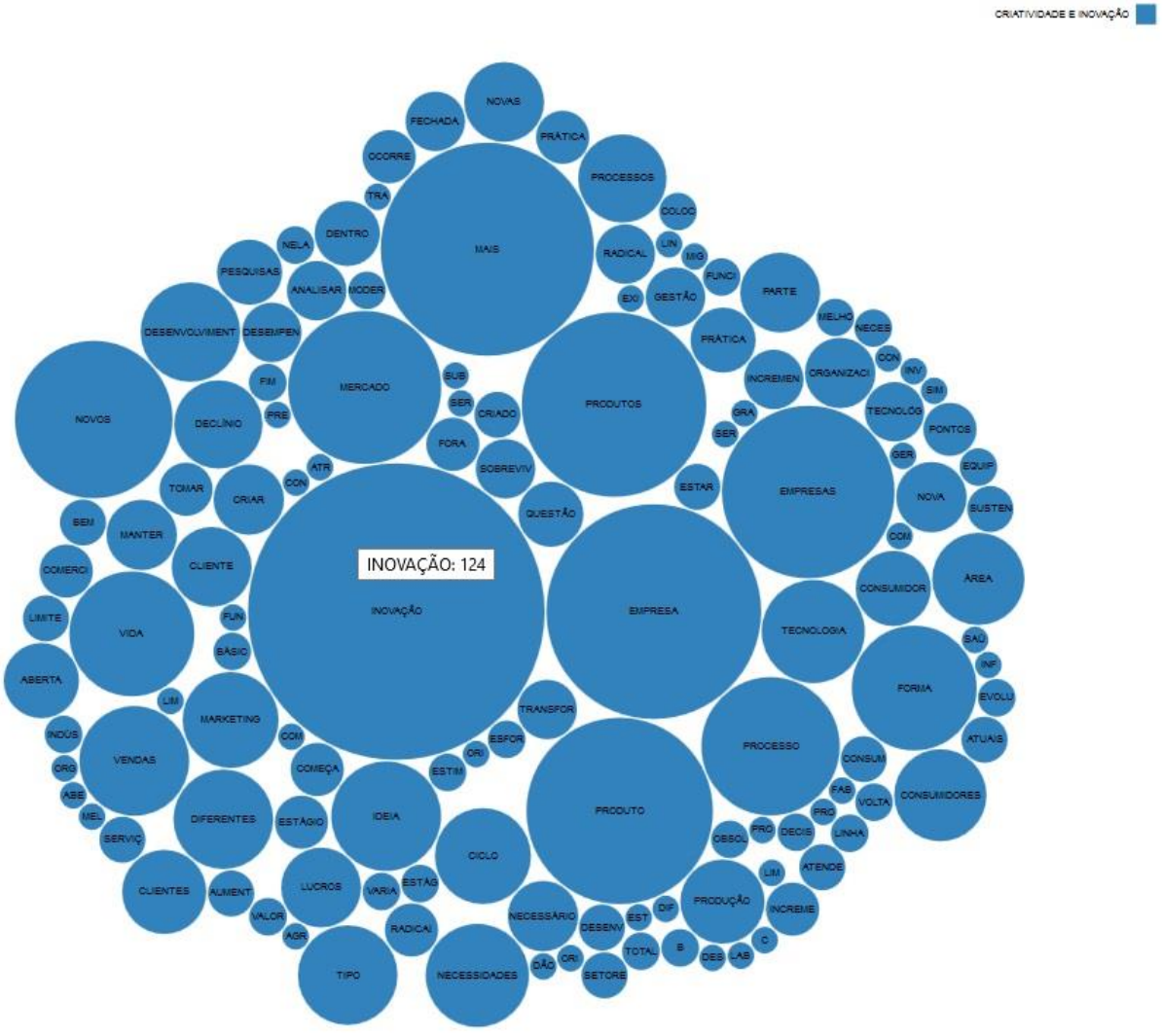
A Figura 29 apresenta a visualização Collapsible Tree (Árvore Flexível). Esta visualização tem por objetivo mostrar as palavrar relevantes utilizadas pelo estudante em sua produção textual. Esta visualização apenas apresenta as palavras, não se pode ter um comparativo entre elas e nem saber quais as mais relevantes.

Figura 29 - Collapsible Tree (Árvore Flexível) para um estudante em uma atividade.



Na Figura 30 é demonstrada a visualização Bubble Chart (Gráfico de Bolhas). O objetivo desta visualização é mostrar a relevância de cada palavra dentro da produção textual do estudante. Como está sendo visualizado apenas um estudante em uma atividade, há apenas uma cor nas bolhas. Quanto maior o tamanho da bolha, maior a sua pontuação. Colocando o ponteiro do mouse sobre a bolha desejada, pode-se visualizar a palavra por completo e também a pontuação que esta palavra obteve. No canto superior direito há uma legenda indicando a atividade que está sendo analisada.

Figura 30 - Bubble Chart (Gráfico de Bolhas) para um estudante em uma atividade.

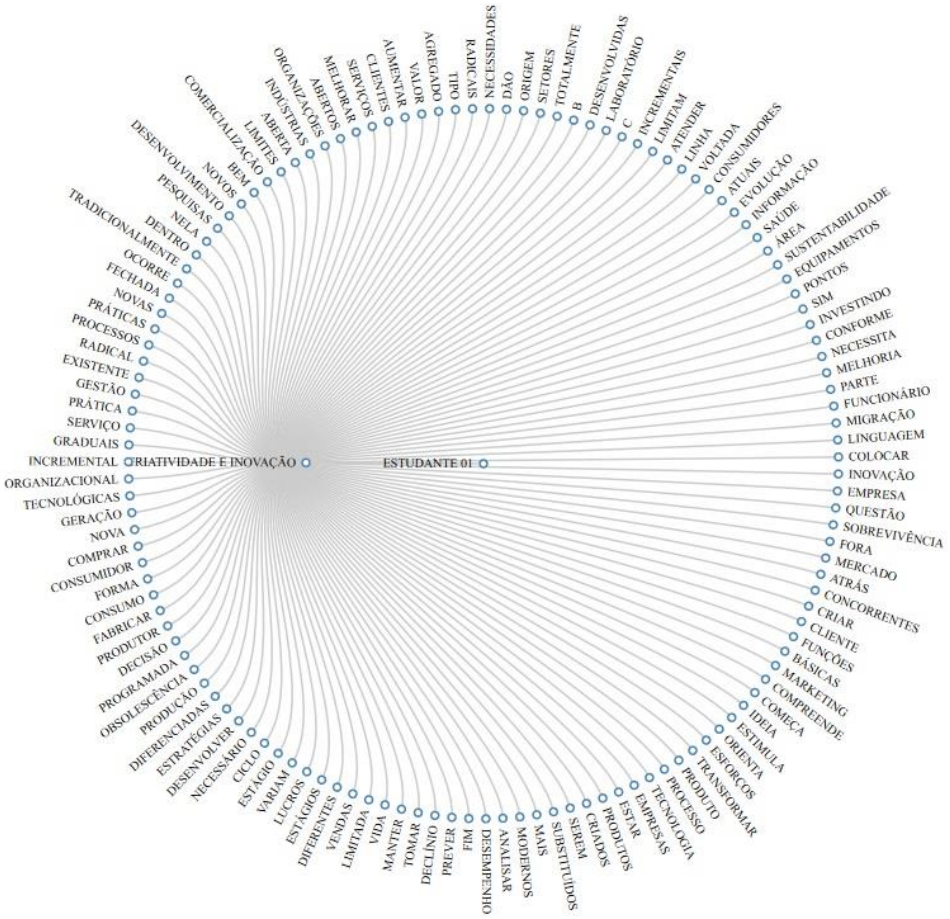


A Figura 31 mostra a visualização Word Cloud (Nuvem de Palavras). Esta visualização visa o comparativo entre as palavras. Quanto maior e mais clara a palavra, maior foi sua pontuação. Quanto menor e mais escura a palavra, menor foi

a pontuação. Esta visualização não mostra a pontuação, apenas apresenta visualmente, de forma intuitiva, quais as palavras mais relevantes na produção textual do estudante na atividade que está sendo analisada.

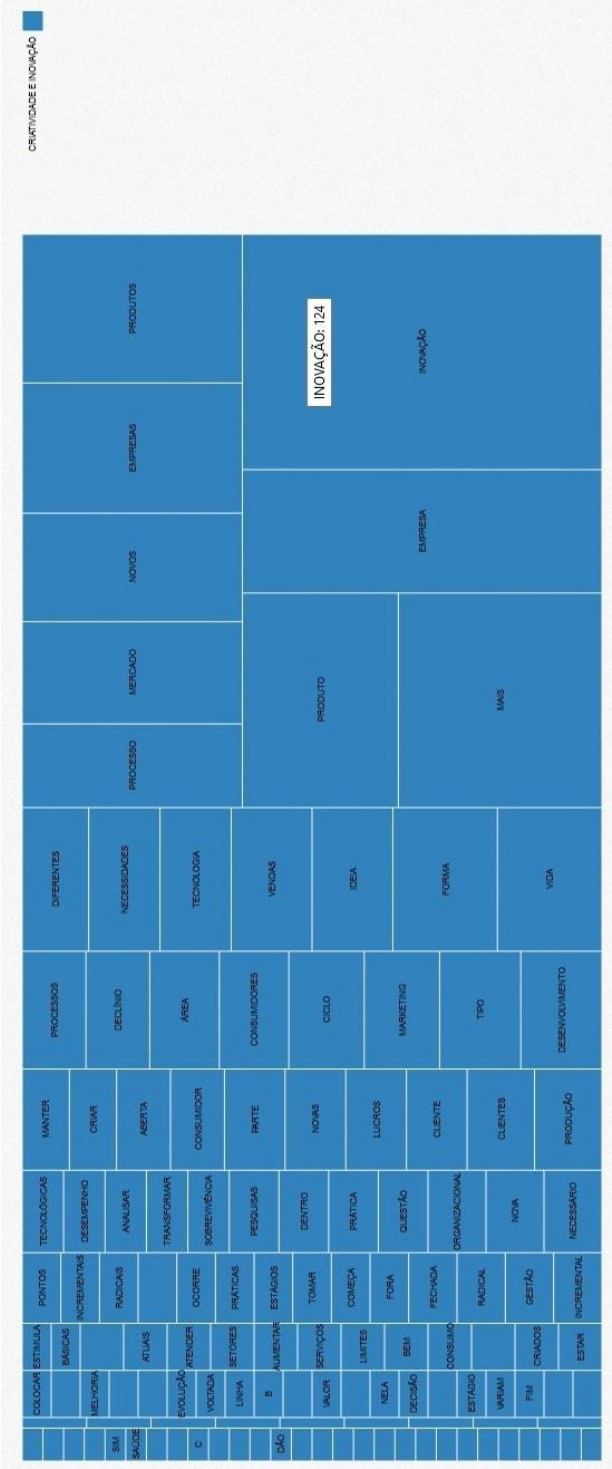
Figura 31 - Word Cloud (Nuvem de Palavras) para um estudante em uma atividade.

Figura 32 - Rotating Cluster (Agrupamento Rotativo) para um estudante em uma atividade.



A Figura 32 mostra a visualização Rotating Cluster (Agrupamento Rotativo). Sua função é muito parecida com a visualização Collapsible Tree, apenas apresentando as palavras em uma disposição diferente. O objetivo é apenas a listagem das palavras, não sendo possível saber quais delas são as mais relevantes na produção textual do estudante.

Figura 33 - Treemap (Mapa Árvore) para um estudante em uma atividade.

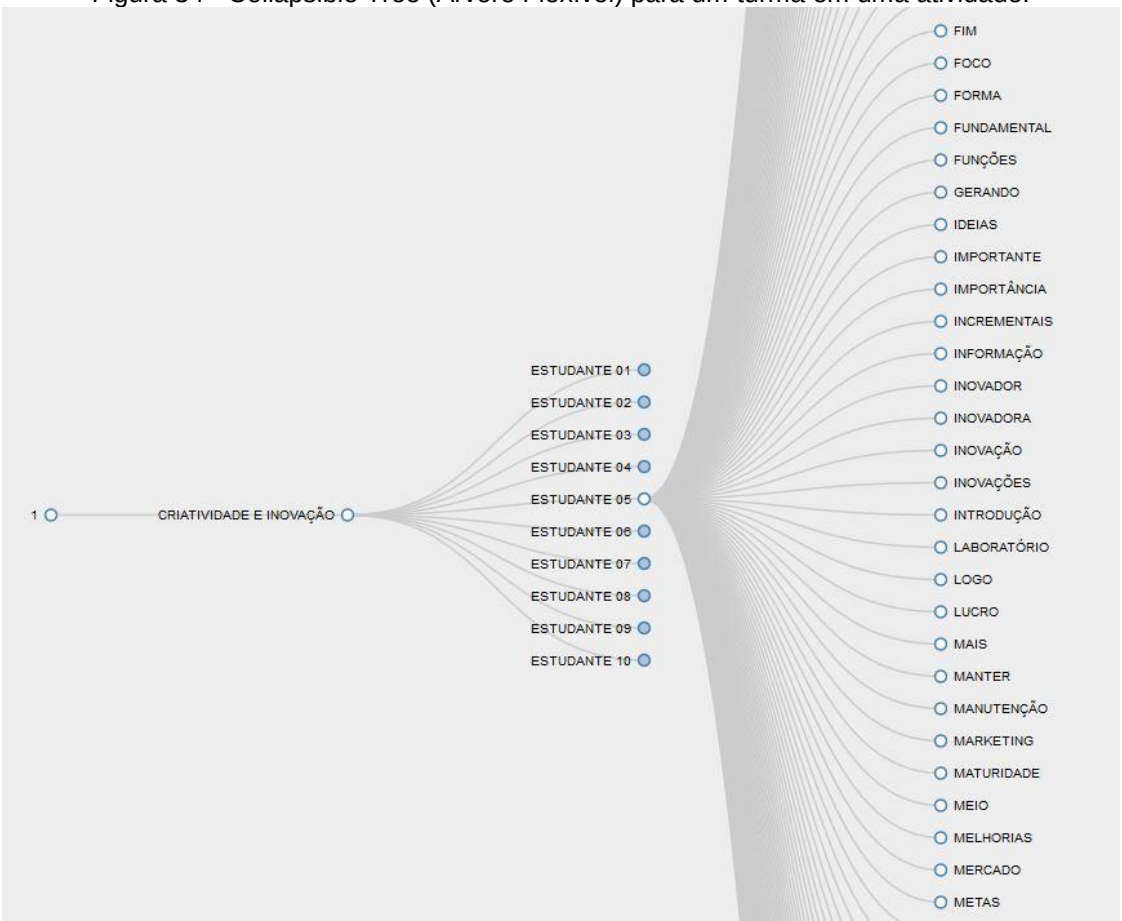


A Figura 33 apresenta a visualização Treemap (Mapa Árvore). Esta visualização tem a função de demonstrar o comparativo entre as palavras. Quanto maior a área da palavra, mais pontuação ela obteve. Quanto menor a área da palavra, menos pontuação a palavra tem dentro da produção textual. Posicionando o cursor do mouse sobre a área da palavra, pode-se visualizar a palavra completa e a sua pontuação. No canto superior direito há uma legenda indicando a atividade que está sendo analisada. Neste caso em que se visualiza apenas uma atividade de um estudante, há apenas uma cor para a visualização.

4.4.1.2 Uma Turma para Uma Atividade

Esta seção visa apresentar as visualizações que se pode obter quando analisada uma turma em uma atividade, objetivando o comparativo entre os estudantes da turma.

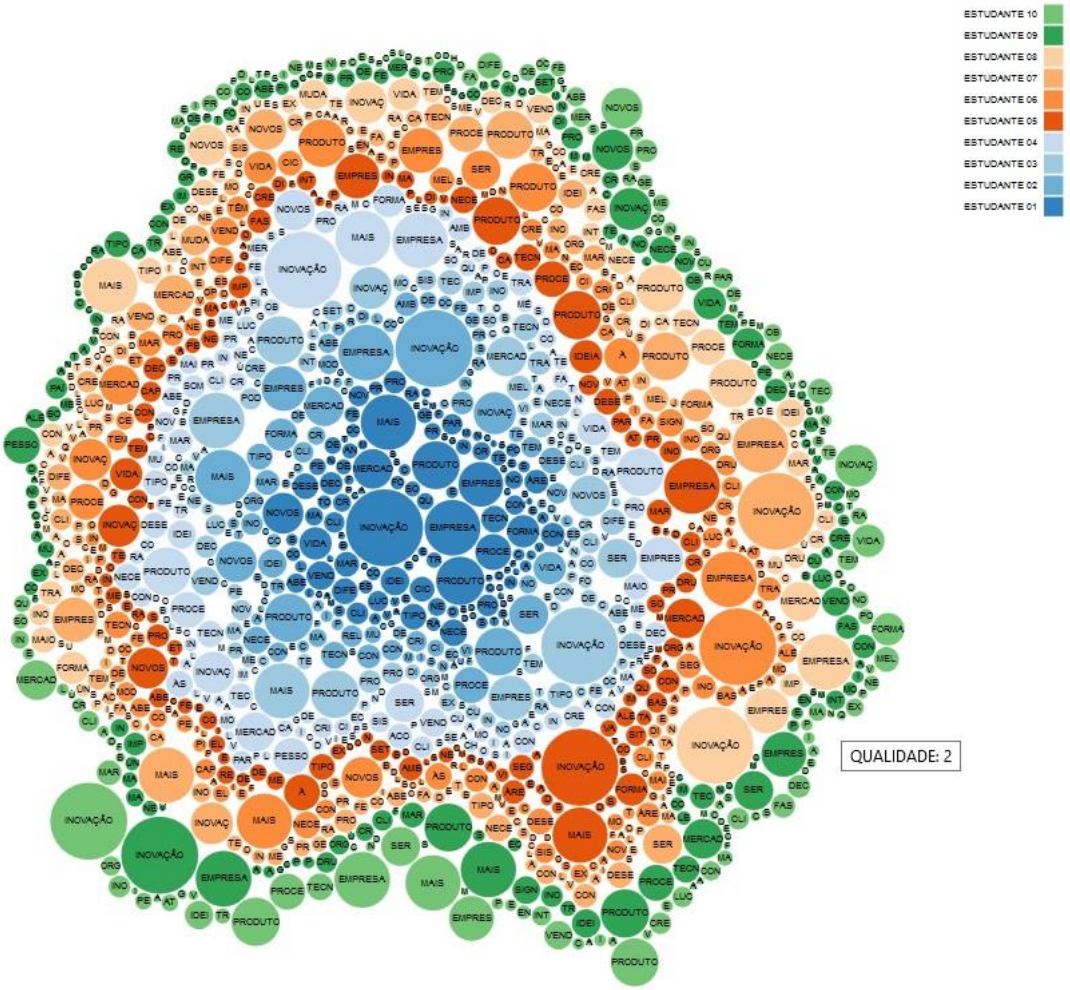
Figura 34 - Collapsible Tree (Árvore Flexível) para um turma em uma atividade.



Neste caso, a visualização Collapsible Tree (Árvore Flexível), apresenta as ramificações da turma, seguido pela atividade e, dentro da atividade, todos os estudantes da turma que realizaram a atividade em questão. Basta apenas clicar sobre um estudante para expandir as palavras relevantes que este utilizou em sua produção textual. Não sendo possível nesta visualização chegar a conclusões quanto a pontuação de cada palavra, apenas a listagem das palavras é apresentada (Figura 34).

Para a visualização Bubble Chart (Gráfico de Bolhas) na Figura 35, as cores das bolhas são de acordo com os estudantes que realizaram a atividade que está sendo analisada. A legenda no canto superior direito indica a qual estudante cada cor representa. Quanto maior a bolha, maior a pontuação da palavra na produção textual do estudante. Quanto menor a bolha, menor a pontuação. Algumas bolhas ficam com o tamanho muito reduzido, então, é possível posicionar o cursor do mouse sobre a bolha desejada para visualizar a palavra completa e sua pontuação.

Figura 35 - Bubble Chart (Gráfico de Bolhas) para uma turma em uma atividade.



O Treemap (Mapa Árvore) representado na Figura 36, apresenta cada estudante em uma cor e quadrante para a atividade que está sendo analisada. A legenda no canto superior direito auxilia na localização do quadrante referente ao estudante que se deseja analisar. Quanto maior a área da palavra, mais pontuação ela obteve na produção textual do estudante. Quanto menor a área da palavra, menos significativa ela é. Como as informações aparecem em grande número, posicionando o cursor do mouse pode ser visualizada a palavra por inteiro e a sua pontuação. Também pode-se clicar no quadrante do estudante para ter um zoom apenas da atividade deste estudante, conforme Figura 37. Clicando novamente, a visualização retorna a sua origem.

Figura 36 - Treemap (Mapa Árvore) para uma turma em uma atividade.

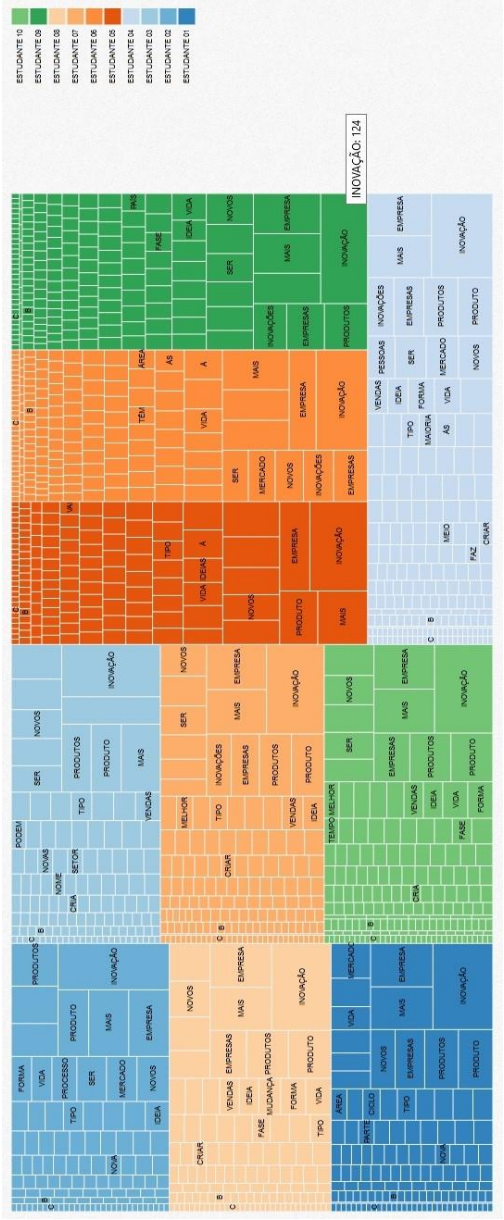
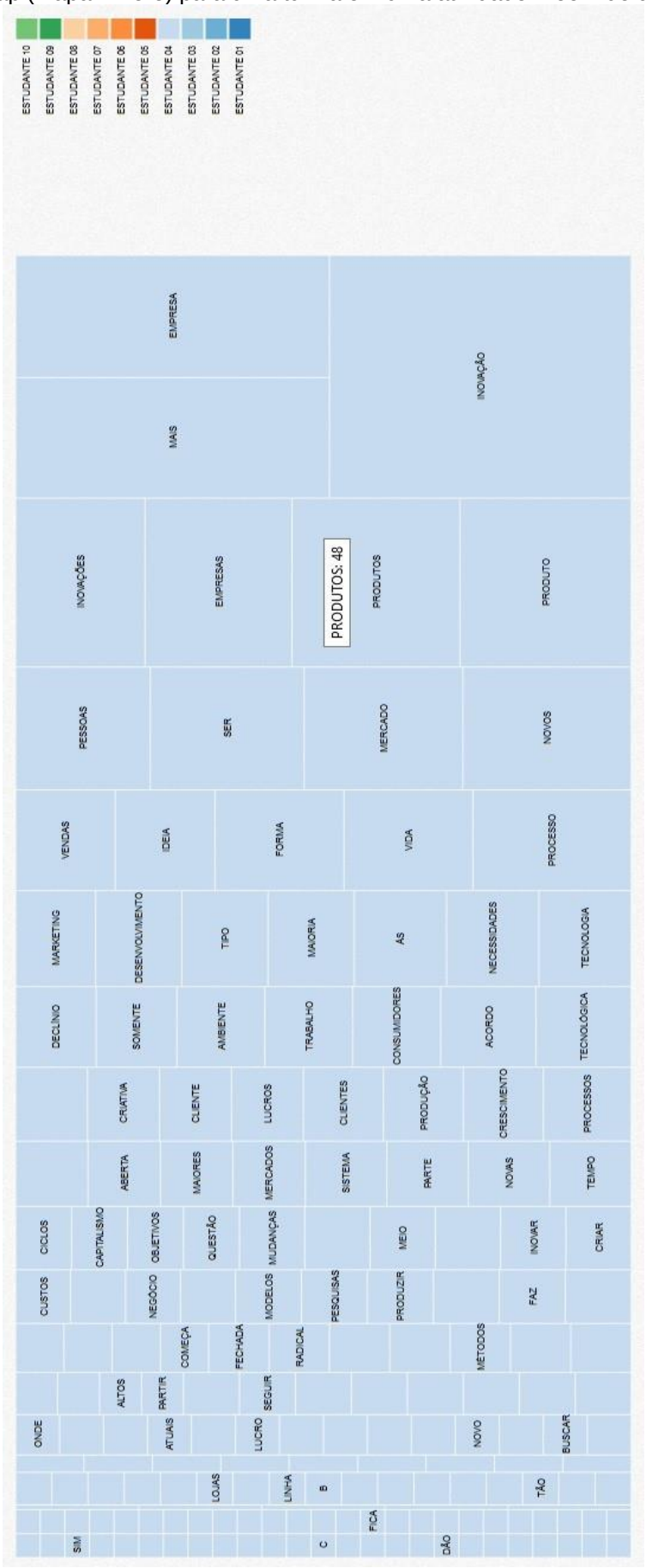
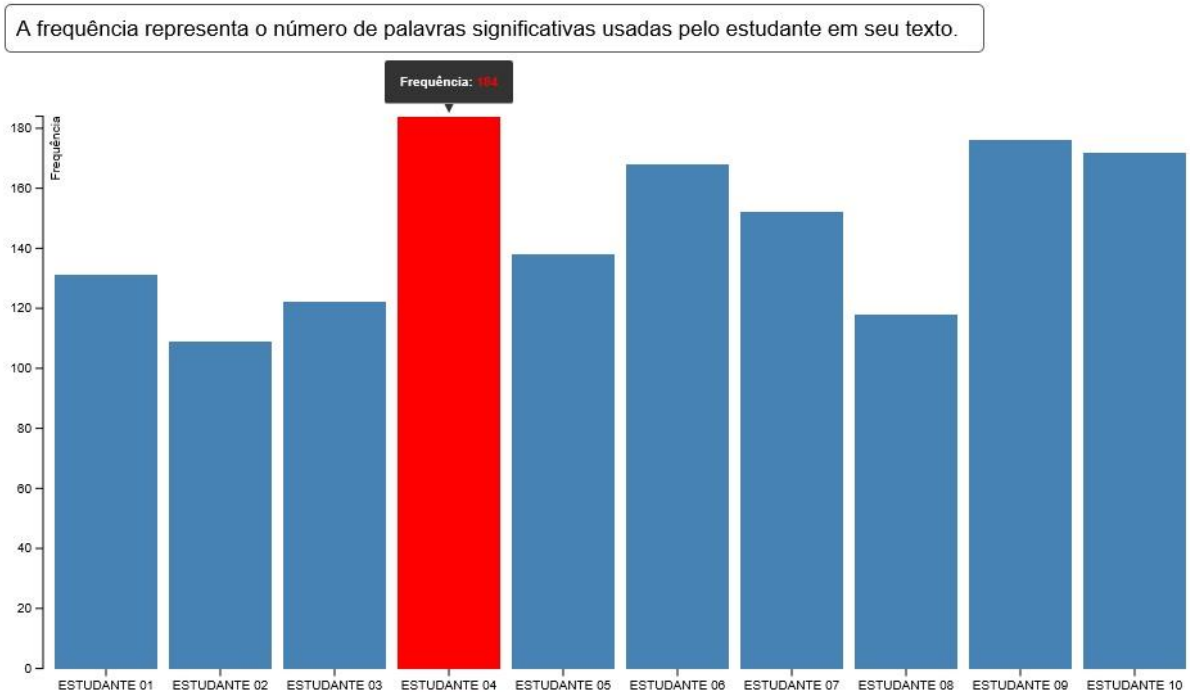


Figura 37 - Treemap (Mapa Árvore) para uma turma em uma atividade. Zoom de um aluno da turma.



A Figura 38 apresenta o Bar Chart (Gráfico de Barras), onde pode-se comparar a quantidade de palavras relevantes utilizadas pelos estudantes da turma analisada. Cada barra é referente a um estudante da turma. Posicionando o cursor do mouse sobre a barra, tem-se a informação da frequência, que é justamente a quantidade de palavras relevantes para a atividade analisada. O objetivo principal desta visualização é comparar o quanto cada estudante aproximou-se do texto original.

Figura 38 - Bar Chart (Gráfico de Barras) para uma turma em uma atividade.



4.4.2 Testes Turma de Polímeros

A amostra alvo do estudo foi constituída por uma atividade realizada pelos alunos da professora Ana Grisa do curso de Polímeros da Universidade de Caxias do Sul. A atividade teve o objetivo de expor as ferramentas de visualização à análise de um profissional que não fosse da área da informática, desta forma deixando o teste mais próximo da realidade. Esta atividade foi composta pela leitura do artigo Polietileno: Principais Tipos, Propriedades e Aplicações (COUTINHO, MELLO e MARIA, 2003). Este artigo explica sobre cinco tipos de polietileno: Polietileno de baixa densidade (PEBD ou LDPE), Polietileno de alta densidade (PEAD ou HDPE),

Polietileno linear de baixa densidade (PELBD ou LLDPE), Polietileno de ultra alto peso molecular (PEUAPM ou UHMWPE) e Polietileno de ultra baixa densidade (PEUBD ou ULDPE). Foram selecionados os textos de sete estudantes da turma, cada estudante escreveu sobre os cinco tipos de polietileno. As produções textuais dos estudantes foram separadas por tipo de polietileno, gerando uma produção textual para cada tipo de polietileno por estudante.

O sistema Análise de Produções Textuais foi treinado com o artigo que os estudantes leram, divididos em cinco atividades, uma para cada tipo de polietileno. Em seguida o sistema executou a análise das produções textuais dos alunos, onde cada aluno possuía uma produção textual por tipo de polietileno.

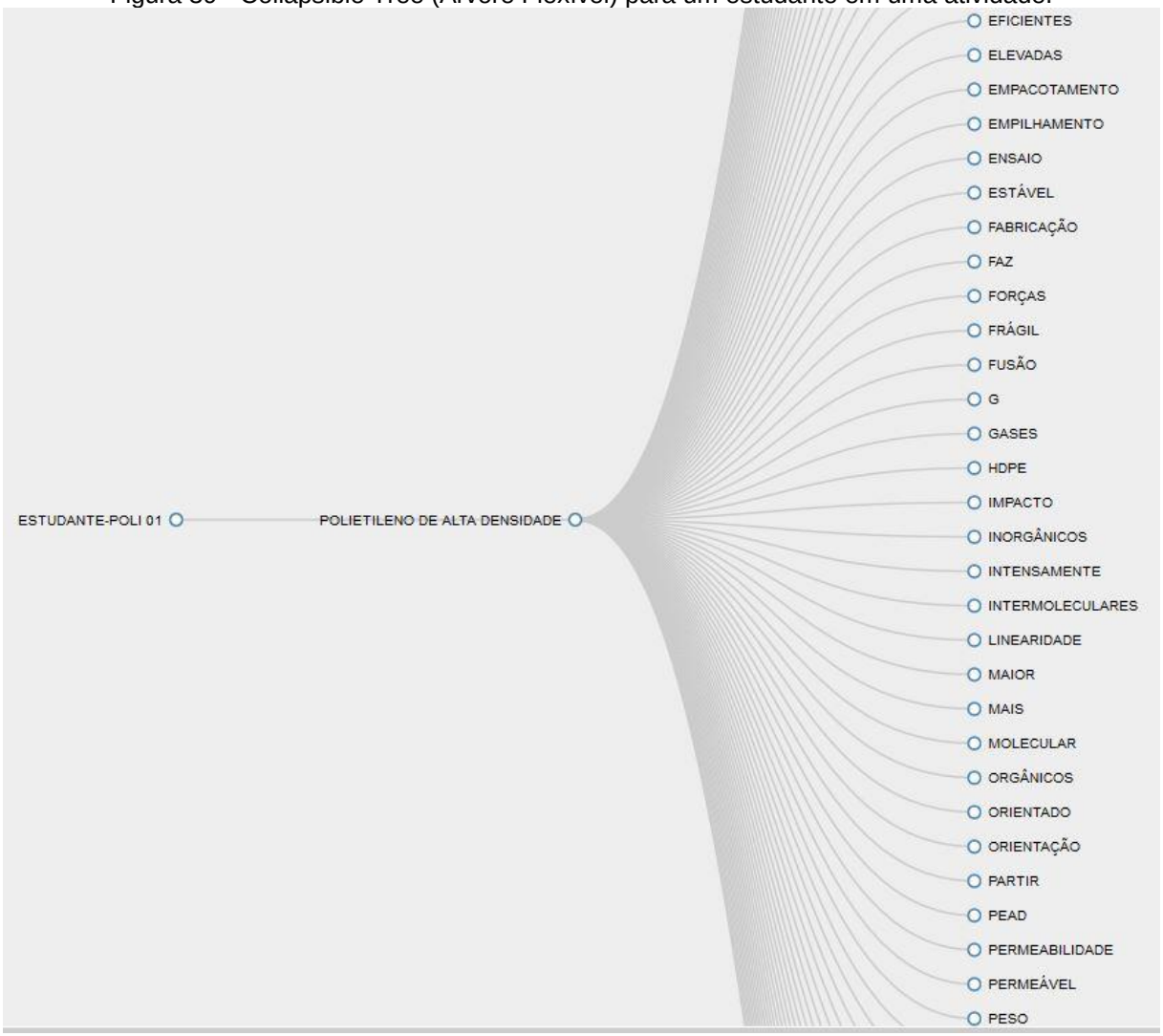
O objetivo desta seção é apresentar os resultados dos testes desenvolvidos, como cada visualização pode ser interpretada conforme a combinação feita na análise. Esta combinação pode ser feita analisando um aluno em uma atividade, um aluno em todas as atividades, uma turma em uma atividade ou uma turma em todas as atividades.

4.4.2.1 Um Estudante para Uma Atividade

Esta seção tem por objetivo apresentar as visualizações geradas quando se compara um estudante em uma atividade. Para apresentar estes testes, foram selecionadas as produções textuais do primeiro estudante da turma na atividade Polietileno de Alta Densidade.

A Figura 39 contém a visualização Collapsible Tree (Árvore Flexível). A proposta desta visualização é apresentar as palavras relevantes utilizadas pelo estudante em sua produção textual. Como só estamos visualizando um aluno em uma atividade, só possui uma ramificação possível de ser expandida. Esta visualização não apresenta a pontuação das palavras, apenas apresenta as palavras em ordem alfabética. É possível aplicar zoom para poder aumentar ou diminuir a visualização.

Figura 39 - Collapsible Tree (Árvore Flexível) para um estudante em uma atividade.



A Figura 40 é referente à visualização Bubble Chart (Gráfico de Bolhas). Esta visualização, além de apresentar as palavras utilizadas na produção textual do aluno, tem por objetivo apresentar a pontuação de cada palavra. Ao posicionar o cursor do mouse em alguma bolha, é apresentada a palavra completa e a pontuação obtida pela mesma. Quando maior a bolha, maior a pontuação obtida, e quando menor o tamanho da bolha, monos pontuação a palavra obteve. No canto superior direito da tela há uma legenda que indica de qual atividade se refere a visualização, como está sendo visualizada apenas uma atividade de um aluno, existirá apenas uma unidade na legenda.

Figura 40 - Bubble Chart (Gráfico de Bolhas) para um estudante em uma atividade.

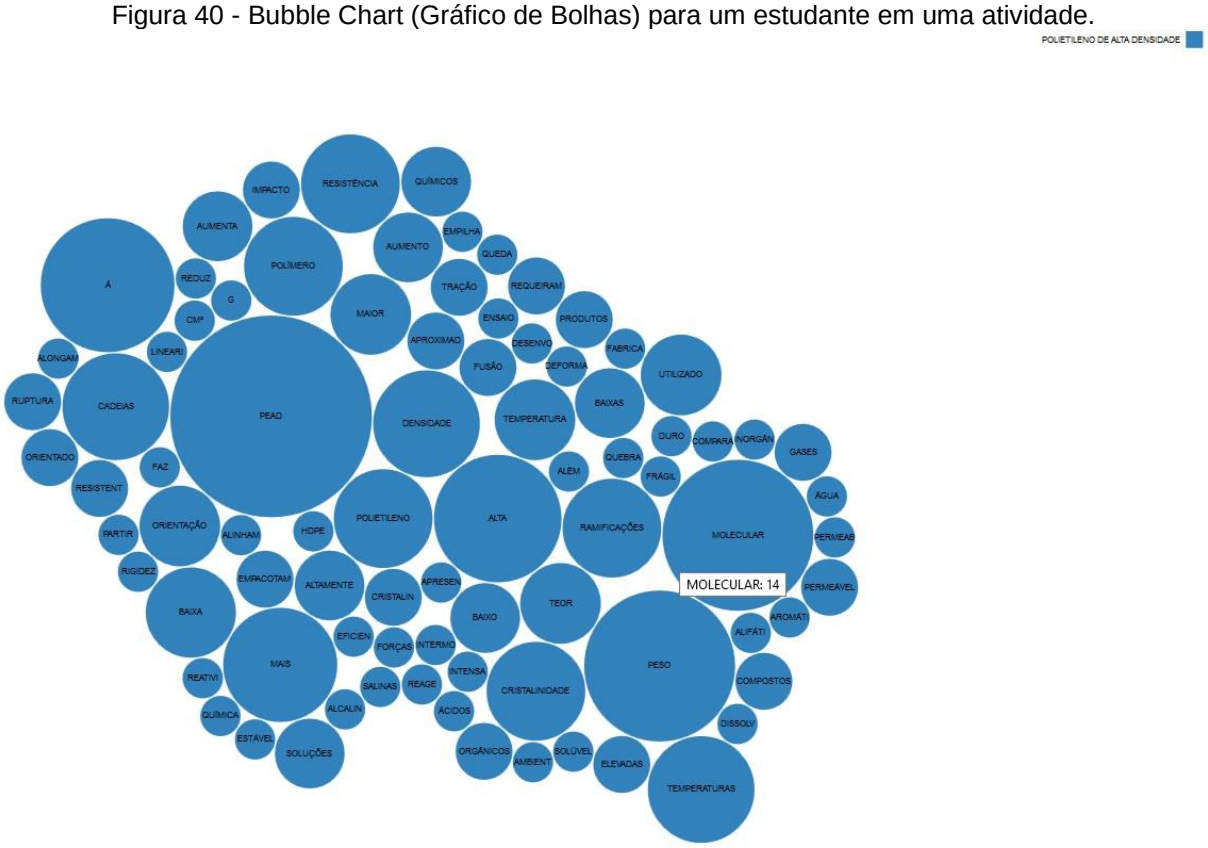
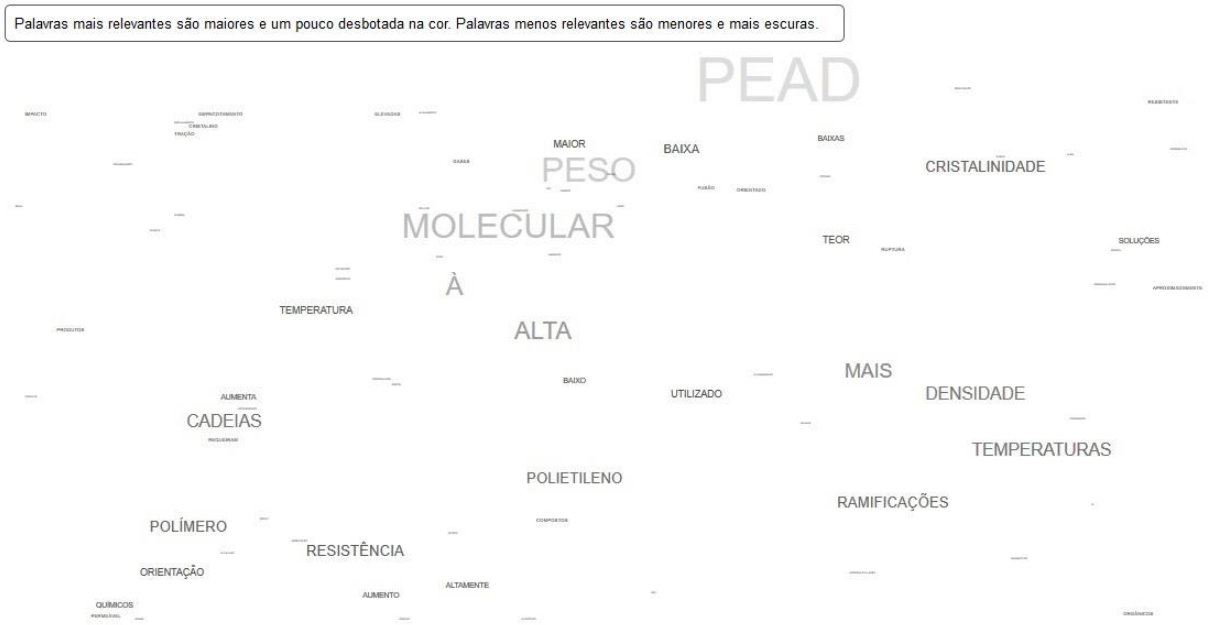


Figura 41 - Word Cloud (Nuvem de Palavras) para um estudante em uma atividade.

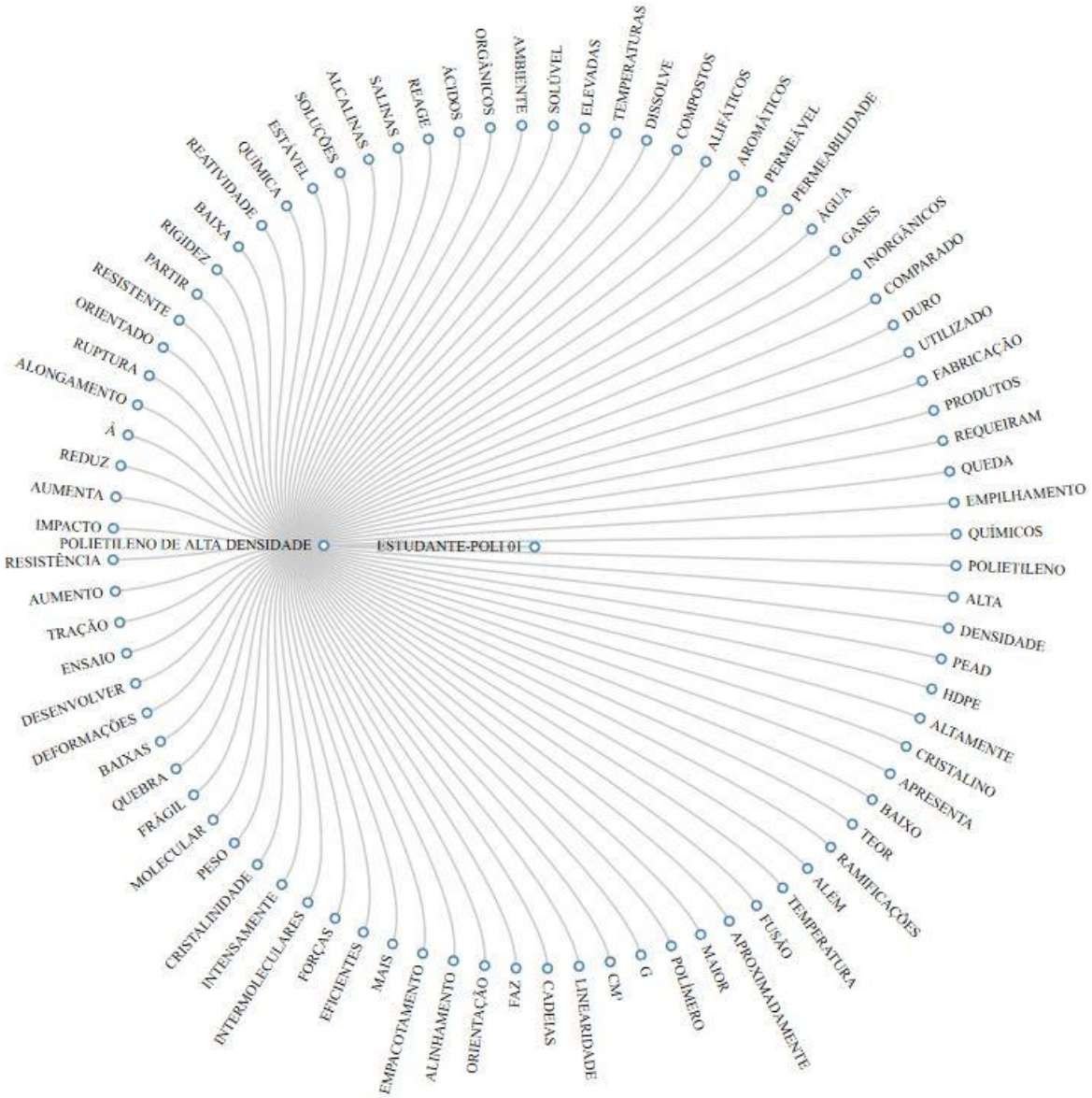


Na Figura 41 é apresentada a visualização Word Cloud (Nuvem de Palavras). Esta visualização tem por objetivo demonstrar a relevâncias das palavras utilizadas pelo estudante em sua produção textual. Quanto maior e mais desbotoada

for a palavra, maior é a pontuação e, quanto menor e mais escura for a palavra, menor é a pontuação obtida pela palavra.

A visualização Rotating Cluster (Agrupamento Rotativo), representada na Figura 42, é muito semelhante com a visualização Collapsible Tree (Árvore Flexível). Apenas apresenta as palavras relevantes utilizadas pelo estudante em seu texto, sem apresentar a pontuação de cada palavra. Esta figura é útil para a visualização como uma listagem de palavras.

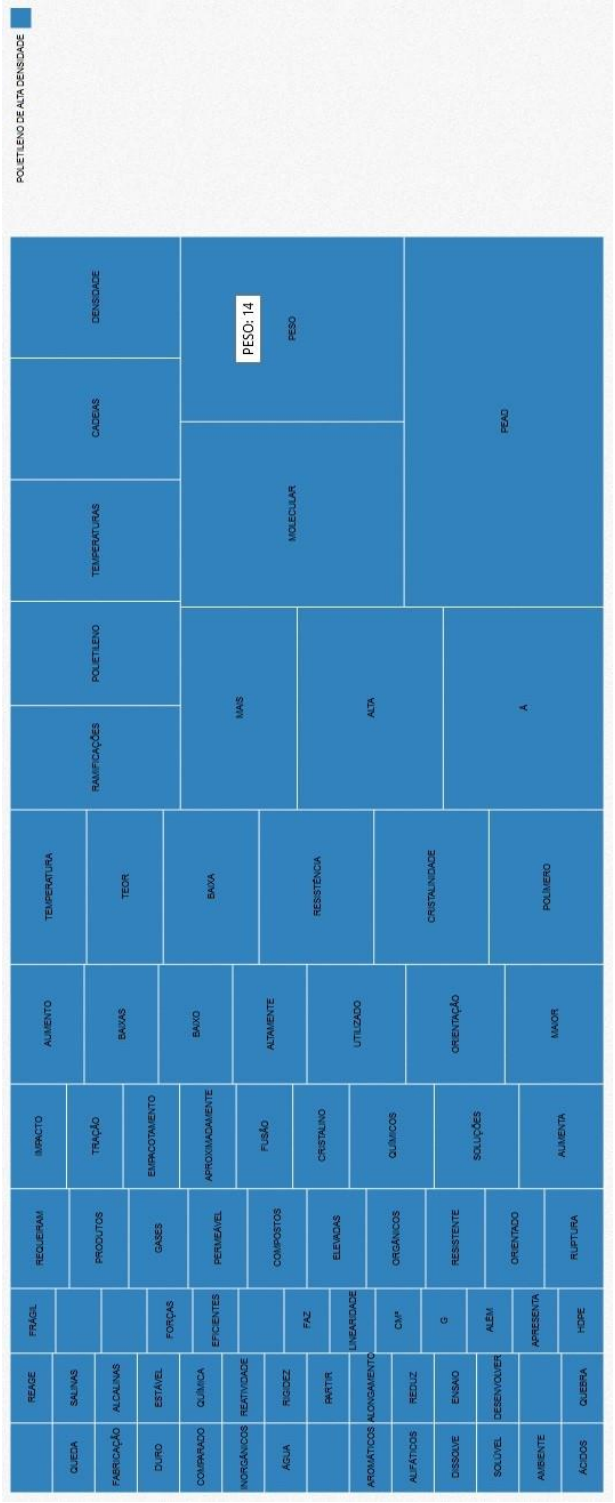
Figura 42 - Rotating Cluster (Agrupamento Rotativo) para um estudante em uma atividade.



A Figura 43 demonstra a visualização Treemap (Mapa Árvore). Esta visualização apresenta as palavras e a pontuação que cada palavra obteve. Esta informação pode ser vista posicionando o mouse sobre a área da palavra. Quando

maior for a área, maior foi a pontuação desta palavra e, quanto menor a área da palavra, menor foi a pontuação obtida pela mesma. No canto superior direito há uma legenda informando a que estudante e atividade se refere a cor da visualização. Neste caso, como está sendo analisada uma produção textual de uma estudante, há apenas uma unidade na legenda.

Figura 43 - Treemap (Mapa Árvore) para um estudante em uma atividade.

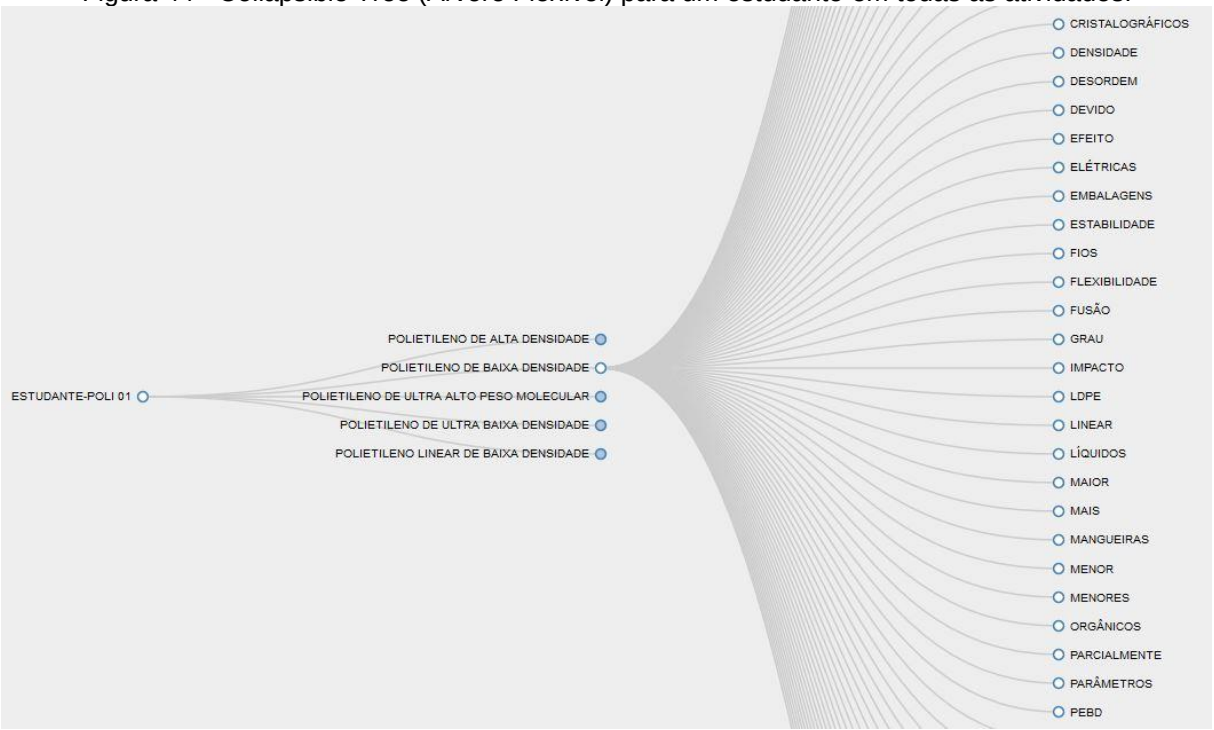


4.4.2.2 Um Estudante para Todas as Atividades

O objetivo desta seção é apresentar as visualizações disponíveis para quando se deseja visualizar o resultado da mineração de textos de uma produção textual de uma estudante em todas as suas atividades.

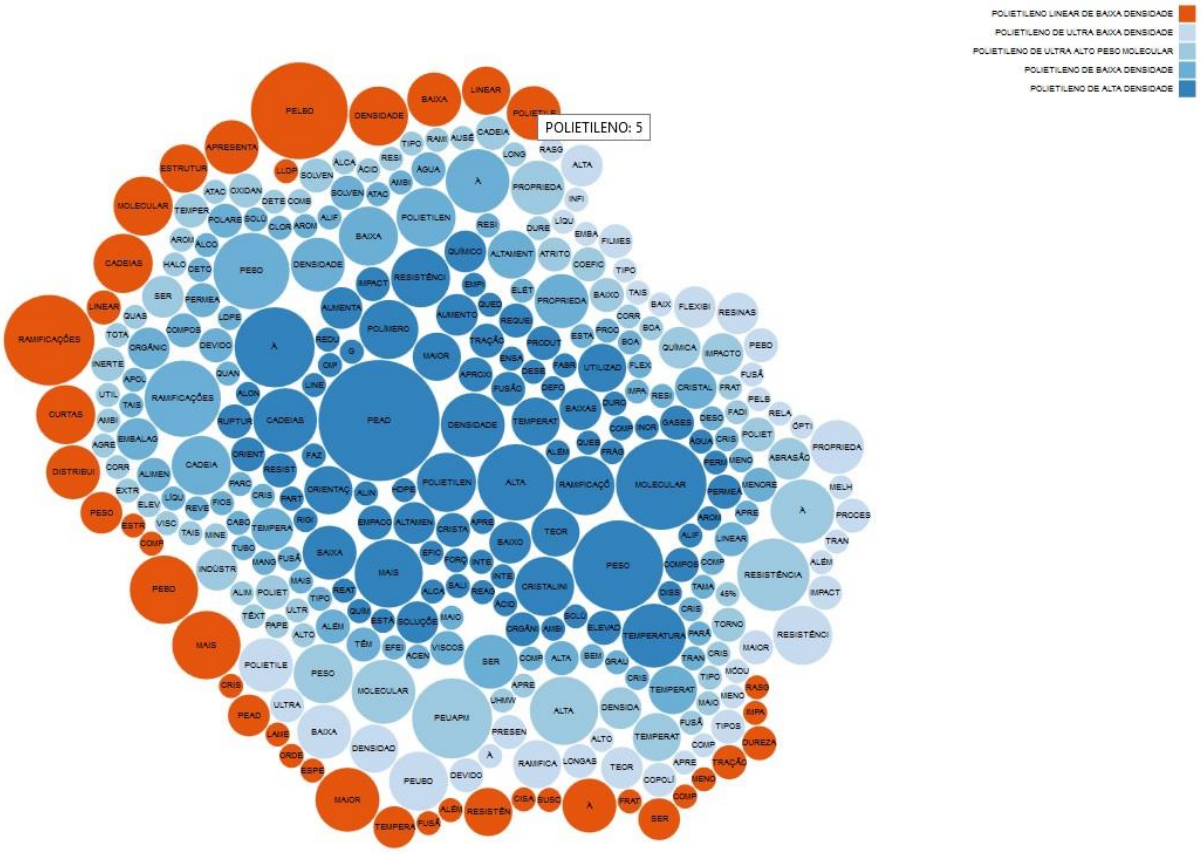
A visualização Collapsible Tree (Árvore Flexível) neste caso aparece com ramificações para as atividades do aluno. Pode-se expandir uma atividade ou todas elas. Neste caso apenas apresentando uma listagem das palavras utilizadas nas produções textuais dos alunos. Esta visualização pode ser vista na Figura 44.

Figura 44 - Collapsible Tree (Árvore Flexível) para um estudante em todas as atividades.



A Figura 45 apresenta o Bubble Chart (Gráfico de Bolhas) com a análise de duas atividades de um estudante. O exemplo mostra dois tons de cor azul, onde cada cor representa uma atividade. Esta diferenciação de cores é mostrada na legenda no canto superior direito. Desta forma pode ser comparado o desempenho do aluno em suas atividades. As bolhas maiores são as de maior pontuação na atividade e as menores são as bolhas de menor pontuação. Posicionando o cursor do mouse sobre uma das bolhas, é apresentada a palavra completa acompanhada de sua pontuação.

Figura 45 - Bubble Chart (Gráfico de Bolhas) para um estudante em todas as atividades.



No Treemap (Mapa Árvore), representado na Figura 46, é apresentado para a análise de todas as atividades de um estudante. Neste exemplo é demonstrado para duas atividades. Cada tom de azul representa uma atividade. No canto superior direito há uma legenda indicando a qual atividade pertence cada cor. Quanto maior a área da palavra, maior a pontuação da mesma, e quanto menor a área da palavra, menor a pontuação na produção textual. Posicionando o cursor sobre a área da palavra, pode ser visualizada a palavra por completo acompanhada da sua pontuação.

Quando há mais do que uma atividade sendo analisada nesta visualização, é acrescentada uma nova funcionalidade que é a de zoom nas palavras de determinada atividade, conforme pode ser visto na Figura 47. Este zoom é apresentado quando clicado sobre alguma área das palavras da atividade. Ao clicar novamente, a visualização retorna ao estado original.

Figura 46 - Treemap (Mapa Árvore) para um estudante em todas as atividades.

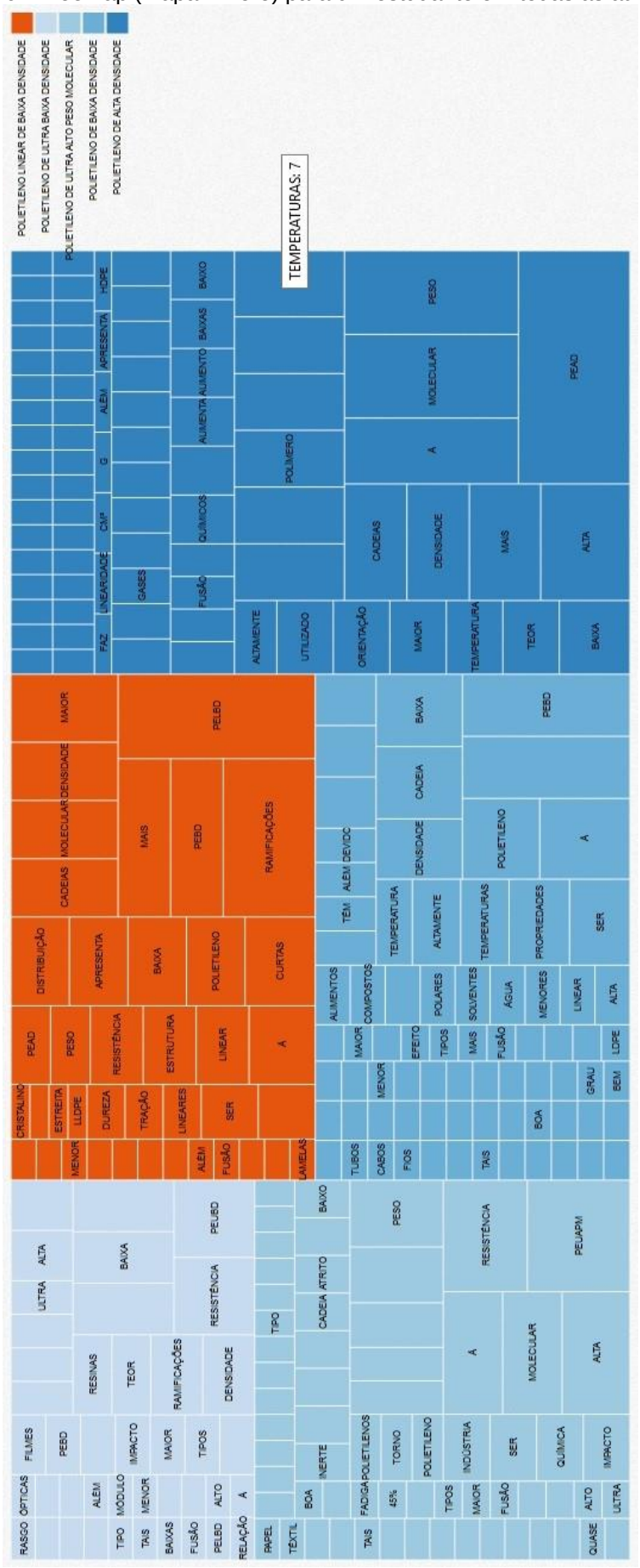
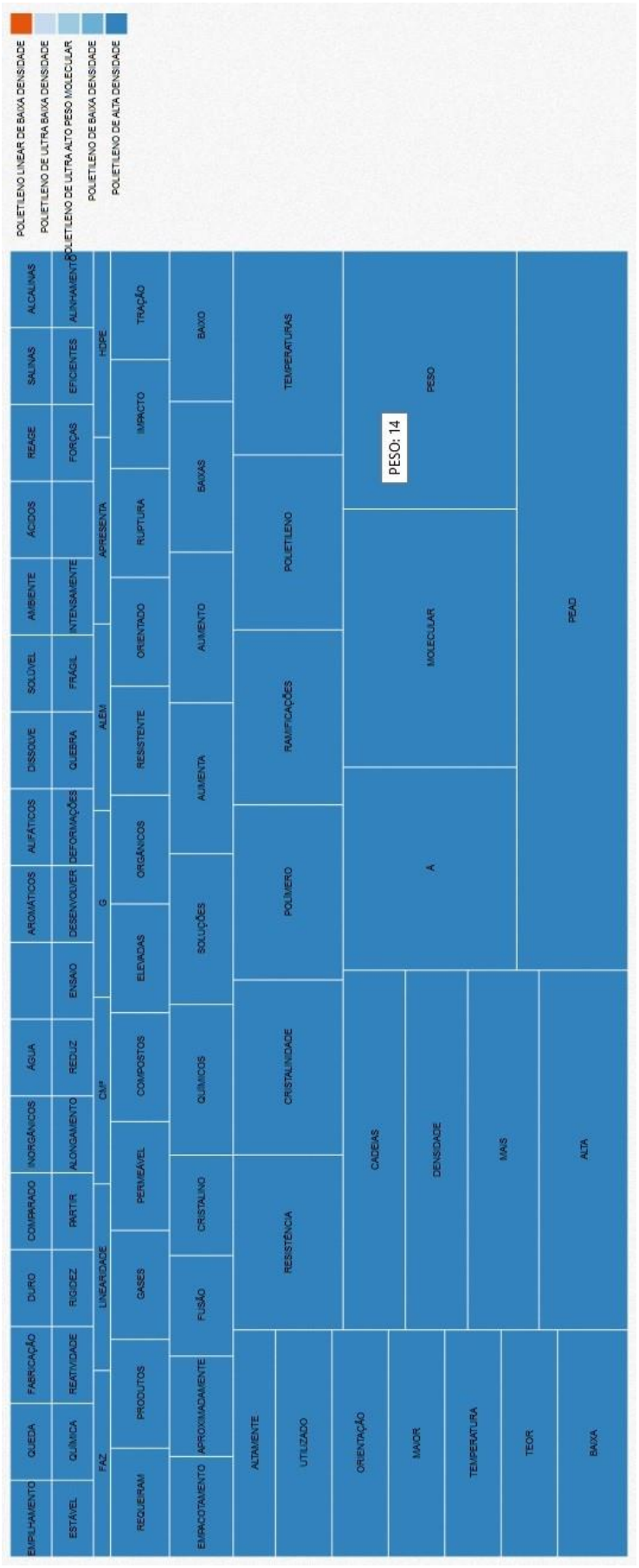


Figura 47 - Treemap (Mapa Árvore) para um estudante em todas as atividades. Zoom de uma atividade.

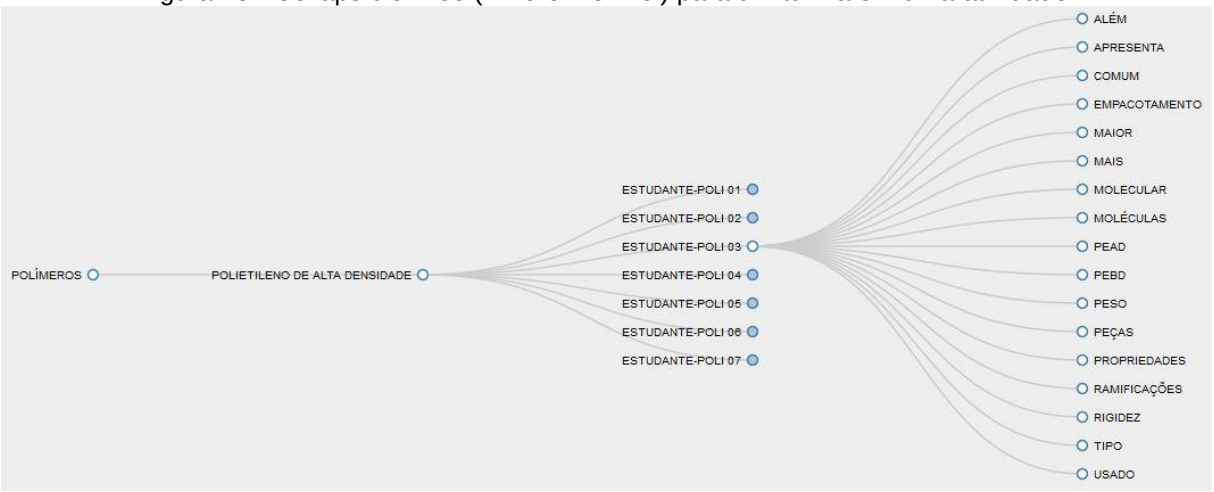


4.4.2.3 Uma Turma para Uma Atividade

Esta seção visa apresentar as visualizações possíveis quando analisada uma turma de estudantes em apenas uma atividade. Desta forma, sendo possível o comparativo entre os estudantes, pondo analisar os estudantes que mais se destacam em cada atividade.

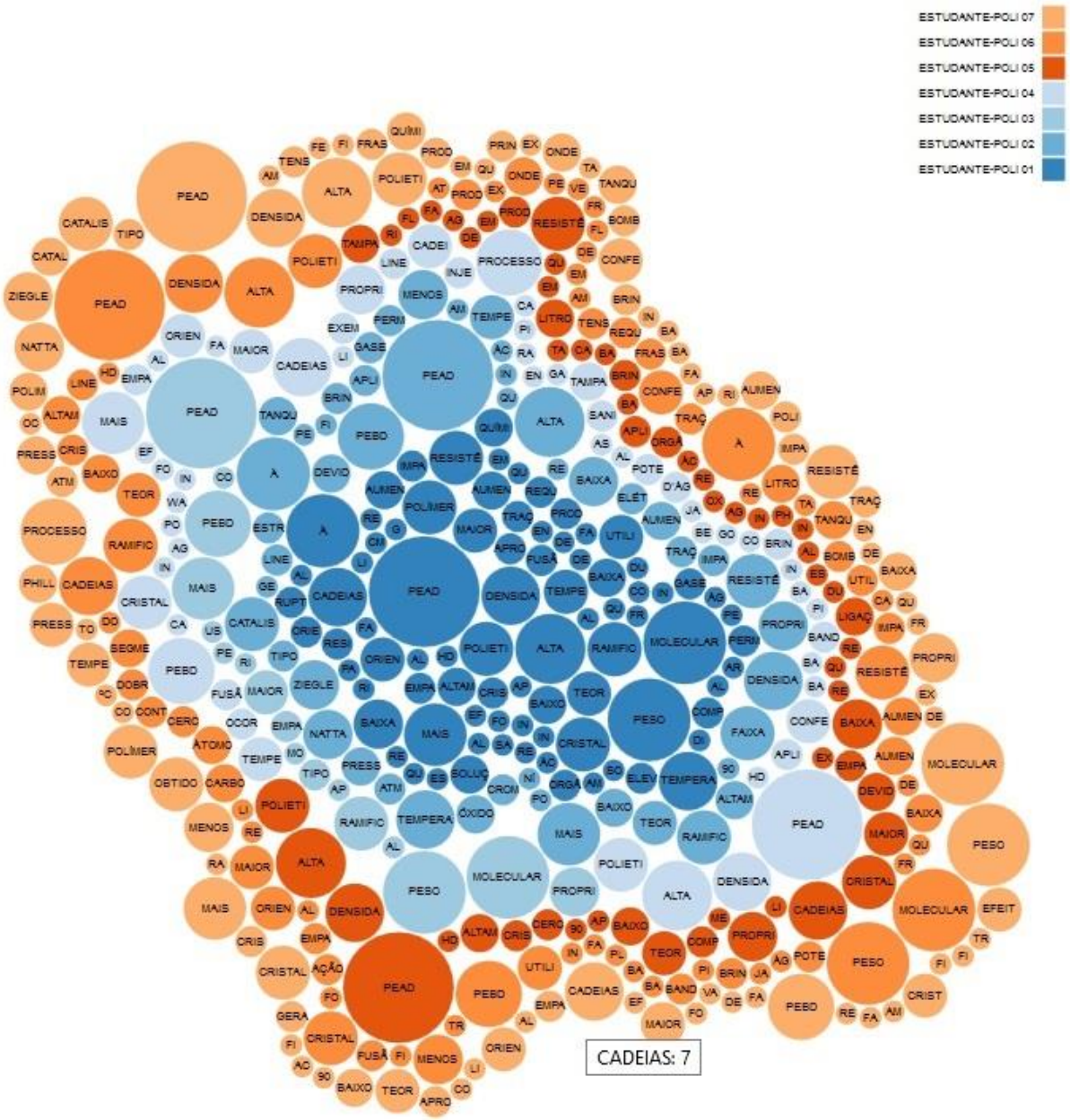
A visualização Collapsible Tree (Árvore Flexível), apresentada na Figura 48, mostra as palavras relevantes utilizadas por cada estudante em suas produções textuais da atividade analisada. Pode-se expandir as ramificações conforme desejado, começando pela turma, atividade e aluno que se quer visualizar, inclusive, visualizando mais de um estudante por vez. Com esta visualização não é possível ter a ideia da importância de cada palavra para o texto.

Figura 48 - Collapsible Tree (Árvore Flexível) para um turma em uma atividade.



A Figura 49 demonstra a visualização Bubble Chart (Gráfico de Bolhas) para uma turma de estudantes em uma atividade. Nesta visualização é possível obter a informação da pontuação de cada palavra para cada produção textual. Ao posicionar o cursor do mouse sobre uma bolha é apresentada a palavra por completo seguida de sua pontuação. Quando maior for a bolha, maior é a pontuação da palavra e, quando menor a bolha, menor a pontuação que esta palavra obteve. As cores representam cada estudante da turma na atividade analisada, para isto, é mostrada uma legenda no canto superior direito.

Figura 49 - Bubble Chart (Gráfico de Bolhas) para uma turma em uma atividade.



A visualização Treemap (Mapa Árvore), para uma turma em uma atividade é apresentada na Figura 50. Esta visualização, assim como a Bubble Chart (Gráfico de Bolhas), tem como objetivo mostrar a pontuação de cada palavra dos textos dos estudantes. Quanto maior for a área da palavra, maior é a pontuação e, quanto menor for a área da palavra, menos pontuação ela obteve. Ao posicionar o cursor do mouse sobre uma área, é apresentada a palavra e a sua pontuação. Cada cor representa um aluno para a atividade analisada, para auxiliar na identificação, há uma legenda no canto superior direito que indica o estudante representado por cada cor. Também há uma funcionalidade de zoom do estudante, conforme demonstrado

na Figura 51. Esta funcionalidade é acionada ao clicar sobre a área colorida do estudante que se deseja detalhar.

Figura 50 - Treemap (Mapa Árvore) para uma turma em uma atividade.

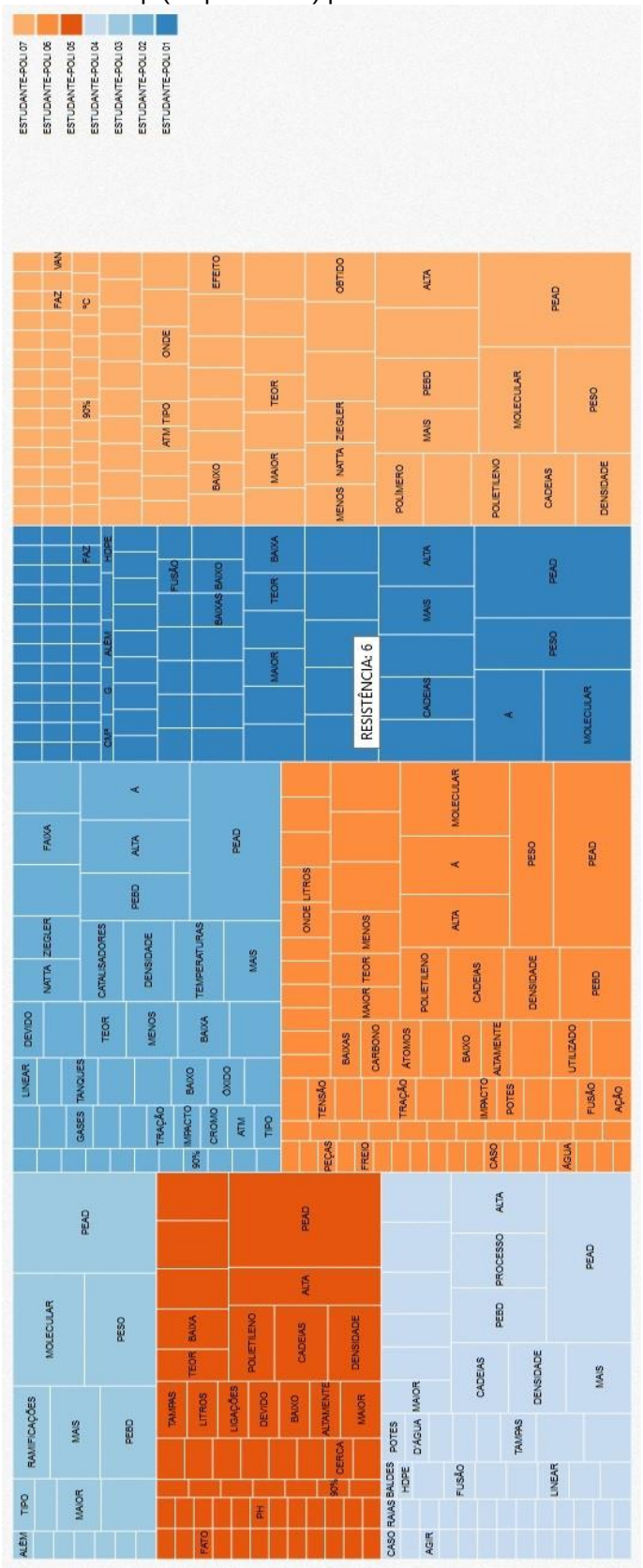
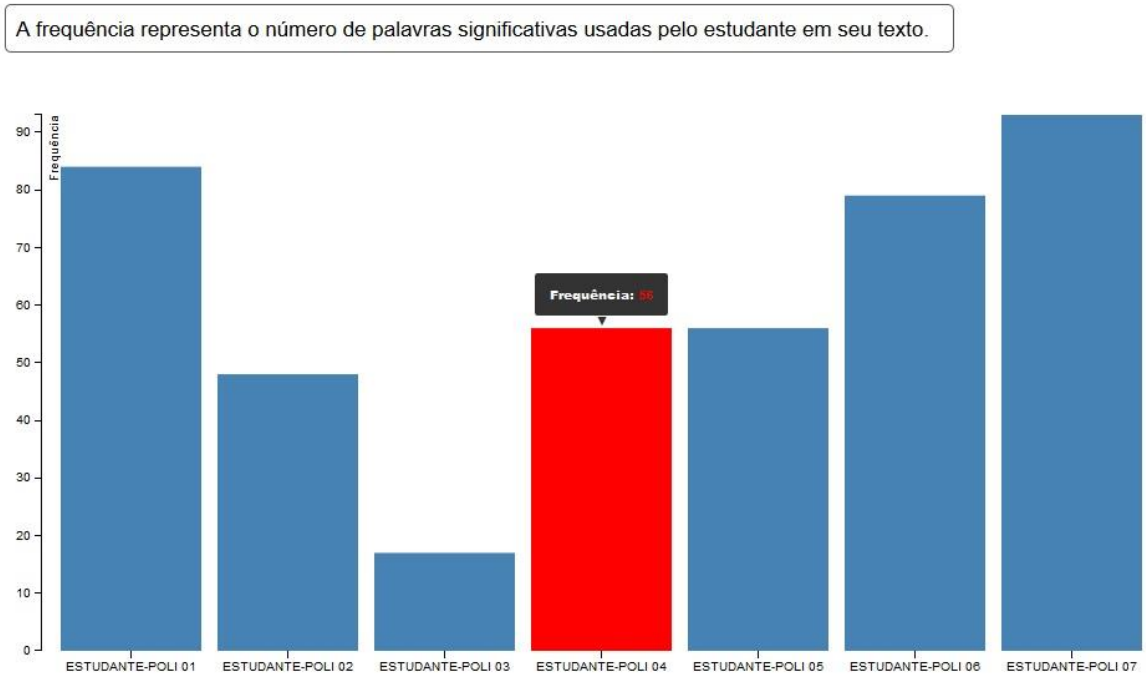


Figura 51 - Treemap (Mapa Árvore) para uma turma em uma atividade. Zoom de um aluno da turma.

O Bar Chart (Gráfico de Barras), é uma visualização que visa a simplicidade da sua interpretação. Pode-se, facilmente, visualizar o estudante que utilizou mais palavras importantes na sua produção textual. A frequência indica a quantidade de palavras significativas utilizadas pelo estudante. Ao posicionar o cursor do mouse sobre cada barra, a informação da frequência é apresentada.

Figura 52 - Bar Chart (Gráfico de Barras) para uma turma em uma atividade.



4.4.2.4 Uma Turma para Todas as Atividades

Esta seção visa apresentar a última forma de visualização da ferramenta Análise de Produções Textuais, onde é possível visualizar uma turma em todas as atividades. Esta é a combinação que mais gera dados, pois podem haver diversos estudantes na turma e estes estudantes podem ter realizado diversas atividades.

Na visualização Collapsible Tree (Árvore Flexível) pode-se ver as ramificações iniciando pela turma, subdividindo entre as atividades e após os estudantes da turma que realizaram a atividade, conforme Figura 53. Clicando em cada estudante tem-se as palavras relevantes utilizadas na produção textual, tendo em vista que é apenas uma listagem e não um comparativo entre os alunos ou entre as atividades.

Figura 53 - Collapsible Tree (Árvore Flexível) para um turma em todas as atividades.

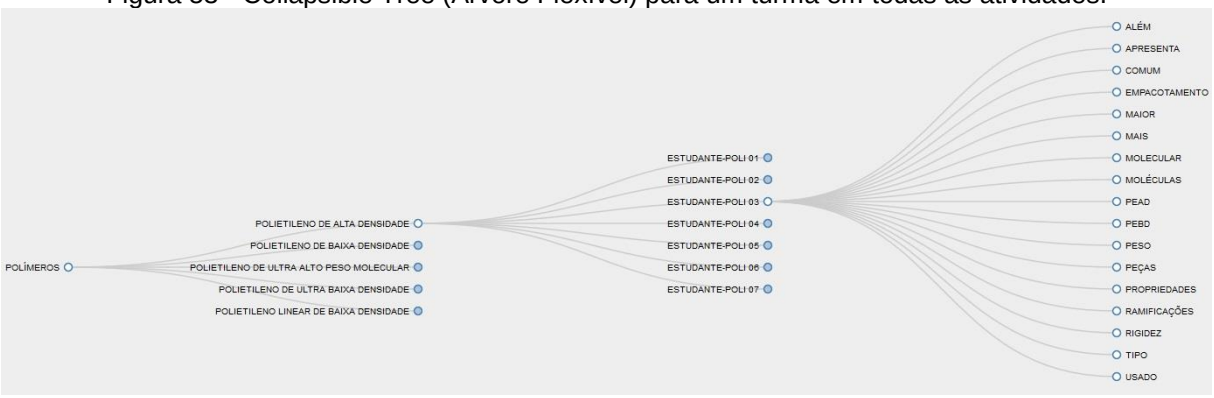
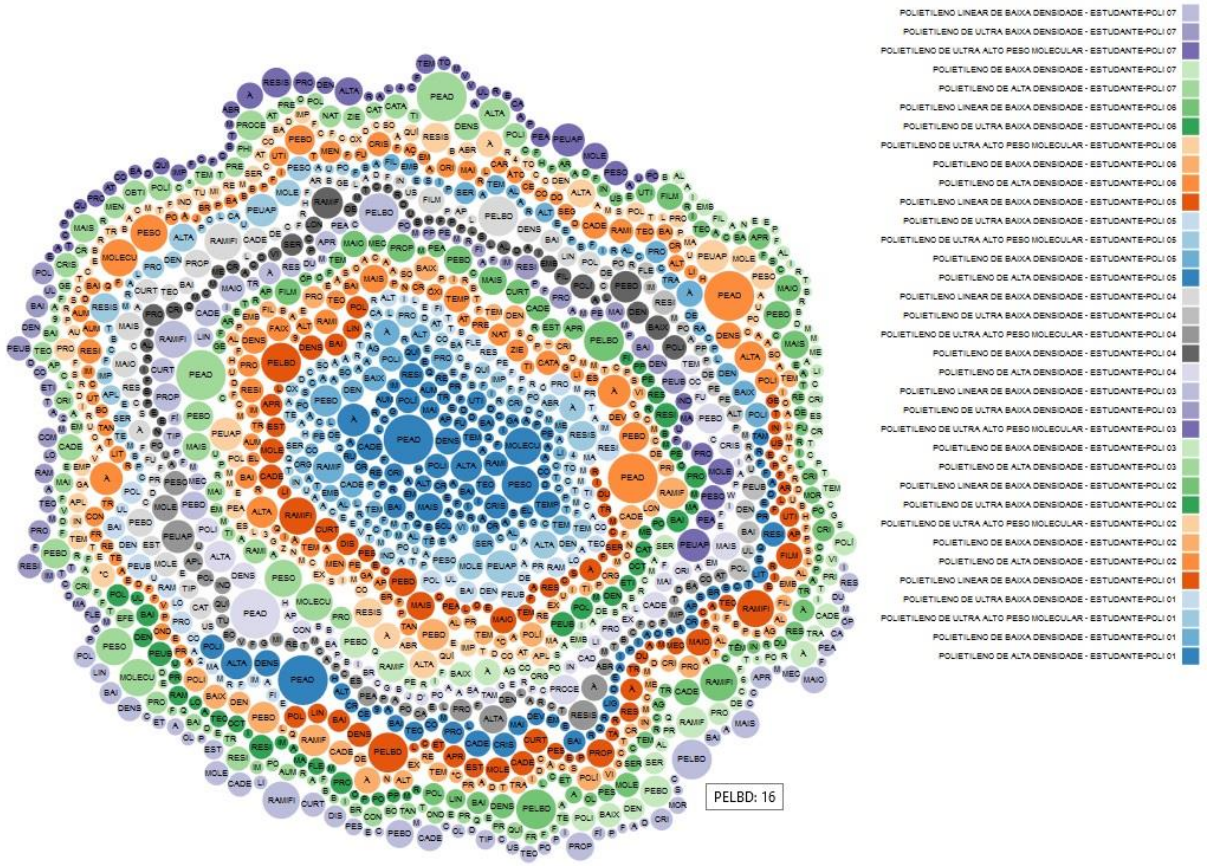


Figura 54 - Bubble Chart (Gráfico de Bolhas) para uma turma em todas as atividades.



A Figura 54 apresenta a visualização Bubble Chart (Gráfico de Bolhas) para todas as produções textuais dos estudantes da turma que está sendo analisada. No canto superior direito é apresentada a legenda que informa que cada cor pertence a um estudante em uma determinada atividade, facilitando a localização no gráfico. Devido a quantidade de informações, algumas bolhas podem ficar muito pequenas, dificultando a identificação da palavra. Pode-se posicionar o cursor do mouse sobre a bolha para ver a palavra e a pontuação na produção textual do aluno. As bolhas

maiores identificam palavras com mais pontuação, e as menores identificam palavras com menos relevância para a atividade.

Figura 55 - Treemap (Mapa Árvore) para uma turma em todas as atividades.

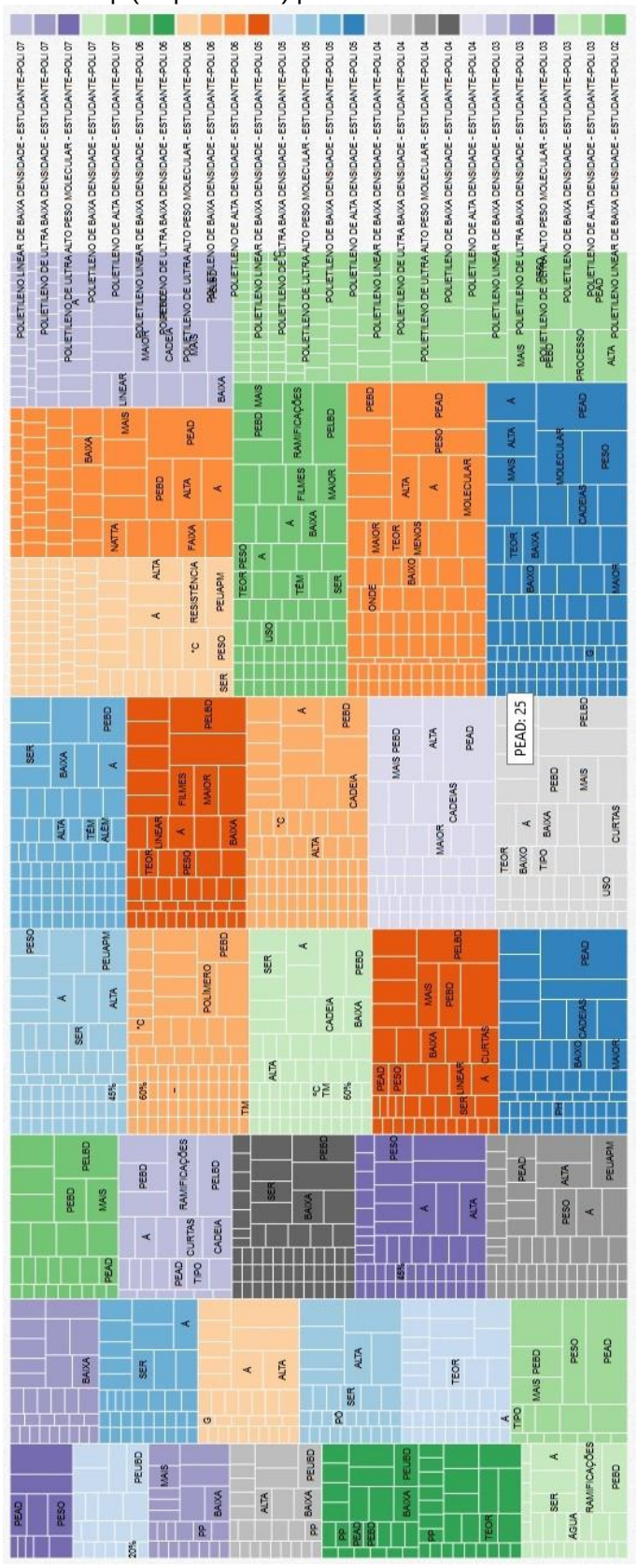


Figura 56 - Treemap (Mapa Árvore) para uma turma em todas as atividades. Zoom de um aluno da turma em uma das atividades.

Como última visualização, é apresentada a Treemap (Mapa Árvore) para visualizar todas as atividades de uma turma, o que demonstra a Figura 55. Conforme legenda no lado direito da imagem, pode-se ver a identificação de cada cor conforme o estudante e a atividade que realizou. Cada quadrante de uma cor representa o estudante em uma atividade. Para visualizar melhor cada palavra, pode-se posicionar o cursor do mouse sobre a área da palavra, podendo visualizar, também, a sua pontuação. As áreas maiores das palavras identificam uma palavra com mais relevância na atividade e as menores identificam uma palavra com menos relevância. Para poder visualizar melhor as palavras de um aluno, pode-se clicar sobre o quadrante do mesmo, expandindo as palavras conforme Figura 56.

4.5 AVALIAÇÃO DOS RESULTADOS

Os resultados foram satisfatórios, o sistema é capaz de gerar visualizações para analisar um estudante em uma atividade, um estudante em diversas atividades, uma turma em uma atividade e uma turma em diversas atividades. Algumas visualizações mostram-se mais adequadas para todas as situações, enquanto outras mostram-se melhores para representar apenas um tipo de texto.

As visualizações Collapsible Tree (Árvore Flexível), Bubble Chart (Gráfico de Bolhas) e Treemap (Mapa Árvore) podem ser utilizadas em todas as combinações, enquanto que visualizações como: Word Cloud (Nuvem de Palavras) pode ser usada apenas quando analisa-se um estudante em uma atividade, isto porque apresenta todas as palavras juntas, não sendo possível identificar de qual texto a palavra pertence. A visualização Rotating Cluster (Agrupamento Rotativo) também pode apenas ser utilizada quando analisa-se um aluno em uma atividade, pois, ao selecionar muitas palavras, a visualização não fica nítida, por sobrepor as palavras. E a visualização Bar Chart (Gráfico de Barras), é utilizada apenas quando analisa-se uma turma em uma atividade, a fim de comparar os alunos em uma atividade em comum.

Os resultados gerados pelo algoritmo de mineração de textos, e posteriormente apresentados em tela pelas visualizações, mostraram-se de acordo com a pontuação atribuída à cada termo na etapa de treinamento. Assim,

conseguindo reproduzir de forma correta a análise dos textos dos estudantes frente ao treinamento efetuado anteriormente.

Desta forma, os objetivos do trabalho, com as visualizações, foram atingidos, ficando a cargo do professor utilizar a visualização que representa de forma mais cognitiva cada caso que deseja-se analisar. Uma avaliação mais completa será realizada no escopo do projeto de pesquisa, UnderMine Text Miner: Software de Mineração de Textos Educacionais, no qual este trabalho está inserido. Para isto, outras técnicas serão implementadas.

5 CONCLUSÃO

Este capítulo visa a apresentação de uma síntese das atividades desenvolvidas durante o desenvolvimento deste trabalho, descrevendo, também, os resultados obtidos e ideias para trabalhos futuros.

5.1 SÍNTESE DO TRABALHO

A Mineração de Textos é um processo importante na extração do conhecimento, identificando palavras, expressões e termos relevantes em uma produção textual. Porém, para ser possível analisar as palavras de um texto minerado, é necessária a apresentação destes dados, objetivando o aumento da cognição humana.

O principal objetivo deste trabalho foi o desenvolvimento de uma ferramenta para a análise de produções textuais no ambiente educacional a partir da reutilização de um algoritmo de mineração de textos. A partir dos dados minerados foi possível realizar a integração com uma ferramenta de geração de visualizações. Desta forma foi possível a combinação de diversas visualizações capazes de representar os resultados da mineração.

Para que o objetivo do trabalho fosse atingido, foram abordados assuntos sobre o processo e os conceitos de Mineração de Textos. Também foram abordados os processos de Visualização de Informação, técnicas para visualização de dados e técnicas para visualização textual. Em seguida foi apresentado o algoritmo de mineração de texto que serviria como base para o desenvolvimento da ferramenta proposta.

Com base no conhecimento adquirido em Mineração de Textos e Visualização de Informação, foi possível o desenvolvimento de um sistema de análise de produções textuais. O funcionamento é dado da seguinte forma: é realizado o treinamento da ferramenta conforme texto base que os estudantes realizarão as atividades. A partir do treinamento, onde o sistema possui a frequência das palavras relevantes, as produções textuais dos estudantes são processadas e

analisadas. Neste momento o sistema está pronto para gerar as visualizações conforme a combinação que se deseja visualizar, podendo ser uma atividade de um estudante, todas as atividades de um estudante, uma atividade de uma turma de estudantes ou todas as atividades de uma turma de estudantes.

Com o sistema desenvolvido, foi realizada uma atividade com uma turma de Polímeros que serviu como amostra alvo do estudo, possibilitando a validação do sistema. A partir da análise destes dados, foi possível concluir que a ferramenta é válida e agrega funcionalidades que aumentam a cognição do processo de Mineração de Textos.

5.2 RESULTADOS E CONTRIBUIÇÕES

A área da ciência da computação está sempre em constante atualização, um profissional, para conseguir acompanhar estas mudanças, tem que estar apto a lidar com novas áreas de estudo, buscando a resolução dos mais diversos problemas computacionais. Desta forma, o esperado é que este profissional desenvolva a habilidade de pesquisa, para que possa sempre buscar novas soluções, acompanhando o mercado e as necessidades dos usuários nas mais diversas áreas.

Ao iniciar este trabalho, foi proposto um projeto de aplicação prática, que inicia uma integração entre dados textuais oriundos da Mineração de Textos com conjuntos de ferramentas visuais que tornassem possível um maior entendimento de quem a utilizar. Para que fosse possível realizar este projeto, muitos estudos foram necessários, expandindo a área de conhecimento além das estudadas no currículo do curso de graduação.

Durante os testes realizados, foi possível verificar que o algoritmo de Mineração de Textos era funcional para os resultados que se esperava, viabilizando a criação do sistema proposto. Assim como a utilização das ferramentas para a geração das visualizações, a qual se mostrou adequada para representar os dados gerados no processo de Mineração de Textos.

Conclui-se, portanto, que o presente trabalho atingiu os objetivos especificados e deixou importantes contribuições para a Mineração de Textos no que diz respeito a apresentação de dados dos textos minerados. Foi factível o

desenvolvimento de uma aplicação que realiza a Mineração de Textos em produções textuais, integrando com ferramentas de visualização. Assim, novos conhecimentos foram adquiridos e agregados aos já obtidos, contribuindo para o enriquecimento de mais uma área da Ciência da Computação.

5.3 TRABALHOS FUTUROS

O presente trabalho abordou a implementação de um sistema de análise de produções textuais no ambiente educacional que reutiliza um algoritmo de Mineração de Texto, integrado com um conjunto de ferramentas para dados textuais.

Como sugestão para trabalhos futuros, o desenvolvimento de novos algoritmos para Mineração de Texto, buscando resultados cada vez mais satisfatórios. Para enriquecer a visualização dos dados minerados, novas pesquisas quanto a visualizações de dados textuais, visando ampliar a interatividade do sistema. Atualmente, com a globalização e o avanço da internet, seria de extrema relevância a transformação deste sistema para uma plataforma totalmente web, permitindo a interação entre pesquisas, com um maior acesso e compartilhamento.

REFERÊNCIAS

ARANHA, C. N. **Uma Abordagem de Pré-Processamento Automático para Mineração de Textos em Português:** Sob o Enfoque da Inteligência Computacional. Rio de Janeiro: [s.n.], 2007.

BOSTOCK, M. D3 Data-Driven Documents. **D3 Data-Driven Documents**, 2013. Disponivel em: <http://d3js.org/>. Acesso em: 22 junho 2015.

CABRAL, M. O Texto Escrito. **Brasil Escola**, 2009. Disponivel em: <http://www.brasilescola.com/redacao/texto-escrito.html>. Acesso em: 26 maio 2015.

CARD, S. K.; MACKINLAY, J. D.; SHNEIDERMAN, B. **Information Visualization. Readings in Information Visualization:** Using Vision to Think, Morgan Kaufmann Publishers. San Francisco: [s.n.], 1999.

CINI, G. **Estudo Sobre a Visualização de Informações no Contexto Educacional**. Caxias do Sul: [s.n.], 2013.

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.

CRISTOFOLI, L. **Mineração de Textos Baseada em Algoritmos Imunológicos.** Caxias do Sul: [s.n.], 2012.

DASGUPTA, D. **Artificial Immune Systems and Their Applications**. Springer-Verlag: [s.n.], 1999.

FEKETE, J. D. The InfoVis Toolkit. **The InfoVis Toolkit**, 2005. Disponivel em: <http://ivtk.sourceforge.net/>. Acesso em: 02 Junho 2015.

FELDMAN, R.; SANGER, J. **The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data**. New York: Cambridge University Press, 2007.

FREITAS, C. M. D. S.; AL., E. **Data usability Issues in Information Visualization Applications.** Wageningen: [s.n.], 2001.

GOOGLE. Google Developers. **Google Developers**, 2015. Disponivel em: <https://developers.google.com/chart/?hl=pt-BR>. Acesso em: 22 Junho 2105.

HAUSER, H.; LEDERMANN, F.; DOLEISCH, H. **Angular brushing of extended parallel coordinates. In:** Information Visualization. [S.l.]: [s.n.], 2012.

HOFFMAN, P. E. **Table Visualization:** A formal Model and Its Aplicattions. Massachusetts: [s.n.], 1999.

JUDELMAN, G. B. **Knowledge Visualization:** Problems and Principles for Mapping the Knowledge Space. Germany: [s.n.], 2004.

KEIM, D. A. **Pixel-oriented database visualizations.** [S.l.]: [s.n.], 1996.

KEIM, D. A. **Visual techniques for exploring databases.** [S.l.]: [s.n.], 1997.

KEIM, D. A. **Information Visualization and Visual Data Mining.** [S.l.]: [s.n.], 2002.

KEIM, D. A.; KRIEGEL, H.-P. **Visualization techniques for mining large databases:** A comparison. In: Knowledge and Data Engineering. IEEE Transactions, v. 8, n. 6: [s.n.], 1996.

KLEMANN, M.; REATEGUI, E.; RAPKIEWICZ, C. **Análise de Ferramentas de Mineração de Textos para Apoio à Produção Textual**. Aracaju: [s.n.], 2011.

MACEDO, A. . R. E. . L. A. . B. P. **Using Text-Mining to Support the Evaluation of Texts Produced Collaboratively. Education and Technology for a Better World; Selected papers of the 9th World Conference on Computers in Education**. Bento Gonçalves: [s.n.], 2009.

MANYEYES. IBM Many Eyes. **Many Eyes**, 2015. Disponivel em: <http://www-01.ibm.com/software/analytics/many-eyes/>. Acesso em: 02 Junho 2015.

NASCIMENTO, H. A. D. D.; FERREIRA, C. B. R. **Uma introdução à visualização de informações**. Goiânia: [s.n.], 2011.

ORENGO, V. M.; HUYCK, C. **A stemming algorithm for the portuguese language. In:** EIGHTH INTERNATIONAL SYMPOSIUM ON STRING PROCESSING AND INFORMATION RETRIEVAL (SPIRE 2001). [S.l.]: [s.n.], 2001.

PICKETT, R. M.; GRINSTEIN, G. G. **Iconographic displays for visualizing multidimensional data.** [S.l.]: [s.n.], 1988.

PREFUSE. Prefuse: Information Visualization Toolkit. **Prefuse**, 2012. Disponivel em: <http://prefuse.org/>. Acesso em: 02 Junho 2015.

RAPKIEWICZ, C. E.; MORETTO, M. Z. **Usando Mineração de Textos como Suporte ao Desenvolvimento de Resumos no Ensino Médio**. 3. ed. [S.l.]: [s.n.], v. 11, 2013.

SOBEK. Sobek Mining. **Minerador de Textos Sobek**, 2015. Disponivel em: <http://sobek.ufrgs.br/index.html>. Acesso em: 26 Maio 2015.

TAGCROWD. TagCrowd. **TagCrowd**, 2015. Disponivel em: <http://tagcrowd.com/>. Acesso em: 27 Maio 2015.

TAN, A.-H. **Text mining:** The state of the art and the challenges. In: PAKDD 1999 WORKSHOP ON KNOWLEDGE DISOCOVERY FROM ADVANCED DATABASES. Beijing: [s.n.], 1999.

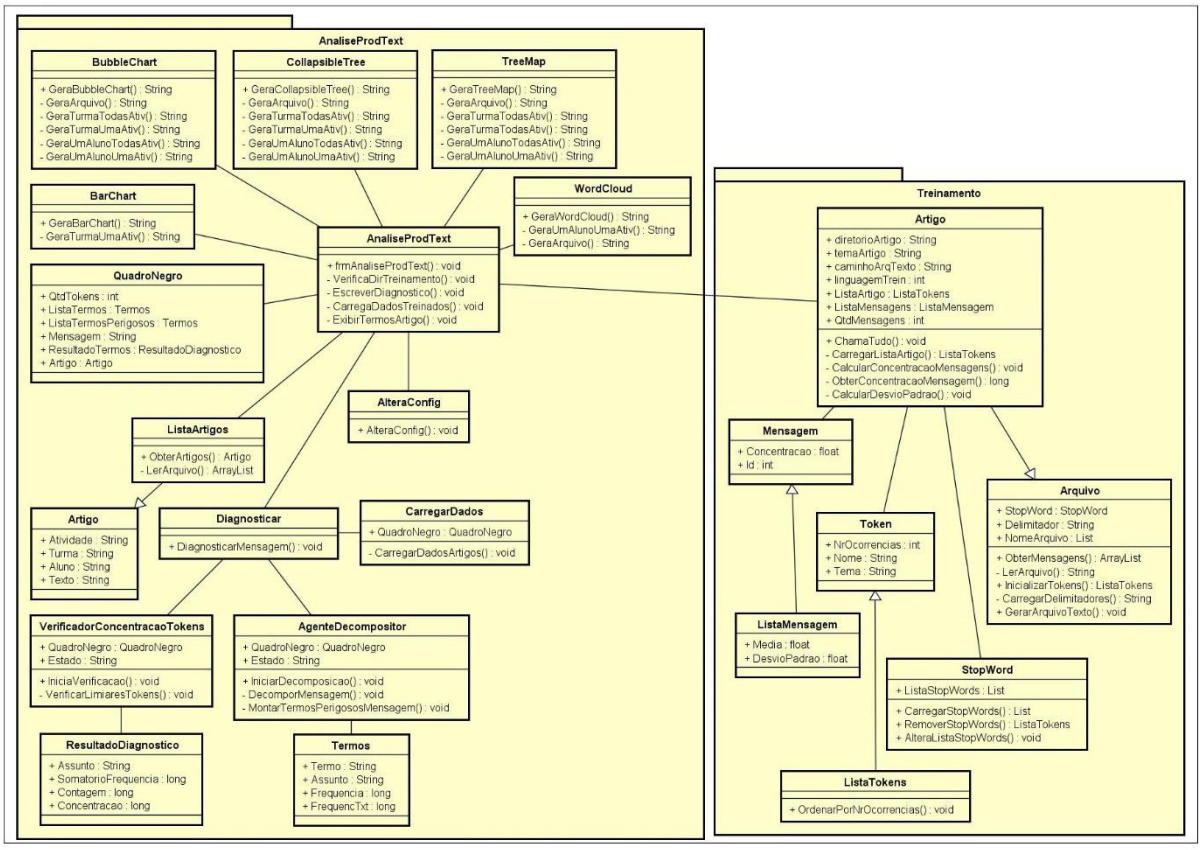
TEXTALYSER. TextAlyser. **TextAlyser**, 2015. Disponivel em: <http://textalyser.net/>. Acesso em: 27 Maio 2015.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining Pratical Machine Learning Tools and Techniques**. Burlington: Elsevier Inc., 2011.

WONG, P. C.; BERGERON, R. D. **30 years of multidimensional multivariate visualization.** Washington: [s.n.], 1997.

WORDCOUNTER. Word Counter. **Word Counter**, 2015. Disponivel em: <http://www.wordcounter.net/>. Acesso em: 27 Maio 2015.

ANEXO A – DIAGRAMA DE CLASSES ANÁLISE DE PRODUÇÕES TEXTUAIS



ANEXO B – MANUAL DO PRODUTO

1 INSTALAÇÃO

Neste capítulo serão descritos os procedimentos necessários para que a ferramenta Análise de Produções Textuais seja utilizada. Também é descrito como pode-se importar o projeto na ferramenta de desenvolvimento, caso algum desenvolvedor tenha o interesse de alterar o sistema.

1.1 PARA DESENVOLVEDOR

Este capítulo é destinado para os desenvolvedores que desejam fazer manutenção no sistema. É descrito os softwares necessário e os passos para importar a aplicação na ferramenta de desenvolvimento.

1.1.1 Softwares necessários

Para a instalação da versão de desenvolvimento serão necessários os softwares abaixo:

a) Os programas liberados para o sistema Análise de Produções Textuais.
b) Visual Studio

   A versão atual do sistema Análise de Produções Textuais foi desenvolvida utilizando a versão 2013 do Visual Studio. Caso seja importado em alguma outra versão, podem haver algumas diferenças de pacotes que terão de ser ajustadas.
c) Apache HTTP Server

   Para a versão atual do sistema Análise de Produções Textuais foi utilizado o Apache HTTP Server 2.2. Os exemplos aqui contidos serão descritos utilizando esta versão. Mas nada impede de utilizar outro software para esta finalidade.

1.1.2 Instalação da Ferramenta

a) A instalação do Servidor Web Apache está descrita no capítulo 1.3. Porém, poderá ser utilizado outro servidor web.
b) O sistema Análise de Produções Textuais está disponibilizado em um arquivo zipado contendo os seguintes diretórios:
    - Implementação
    - Visualizações
c) Dentro do diretório *Implementação* estão os fontes do sistema. O arquivo que deve ser importado no Visual Studio é o (Implementação\AnaliseProdText\AnaliseProdText.sln).
d) Os diretórios contidos dentro de *Visualizações* possuem as implementações das visualizações que devem ser colocadas dentro do diretório do servidor web. Conforme os exemplos desde manual, ficaria conforme a seguir: C:\Webserver\Apache2.2\htdocs\. Abaixo deste diretório deve-se copiar os diretórios contidos em *Visualizações*.

1.2 PARA USUÁRIO

Este capítulo é destinado aos usuários da aplicação. Será descrito os softwares necessários e os procedimentos para instalar a aplicação no computador.

1.2.1 Softwares Necessários

Para a instalação da versão de desenvolvimento serão necessários os softwares abaixo:

a) Os programas liberados para o sistema Análise de Produções Textuais.

   Junto com os programas do sistema há um instalador do sistema, como o sistema foi desenvolvido utilizando a linguagem C#, só poderá ser instalado no sistema operacional Windows.

e) Apache HTTP Server

   Para a versão atual do sistema Análise de Produções Textuais foi utilizado o Apache HTTP Server 2.2. Os exemplos aqui contidos

serão utilizando esta versão. Mas nada impede de utilizar outro software para esta finalidade.

1.2.2 Instalação da Ferramenta

a) A instalação do Servidor Web Apache está descrita no capítulo 1.3. Porém, poderá ser utilizado outro servidor web.
b) O sistema Análise de Produções Textuais está disponibilizado em um arquivo zipado contendo os seguintes diretórios:
    - Implementação
    - Visualizações
c) Dentro do diretório *Implementação**Instalador* está o instalador da aplicação. Executando o arquivo *setup.exe* a aplicação é instalada no computador.
d) Os diretórios contidos dentro de *Visualizações* possuem as implementações das visualizações que devem ser colocadas dentro do diretório do servidor web. Conforme os exemplos desde manual, ficaria conforme a seguir: C:\Webserver\Apache2.2\htdocs\. Abaixo deste diretório deve-se copiar os diretórios contidos em *Visualizações*.

1.3 INSTALAÇÃO SERVIDOR APACHE

O servidor apache deverá ser configurado conforme padrão. Apenas deve-se cuidar com o caminho que se disponibilizara os arquivos que ficarão publicados pelo servidor. Na instalação deste manual foi configurado o caminho (C:\Webserver\Apache2.2\htdocs). Neste caminho ficarão publicados os programas responsáveis pelas visualizações da aplicação.

A Figura 57 apresenta como ficará a estrutura de diretórios após configurado conforme escrito no capítulo 1.1.2 e 1.2.2.

Figura 57 - Servidor web.

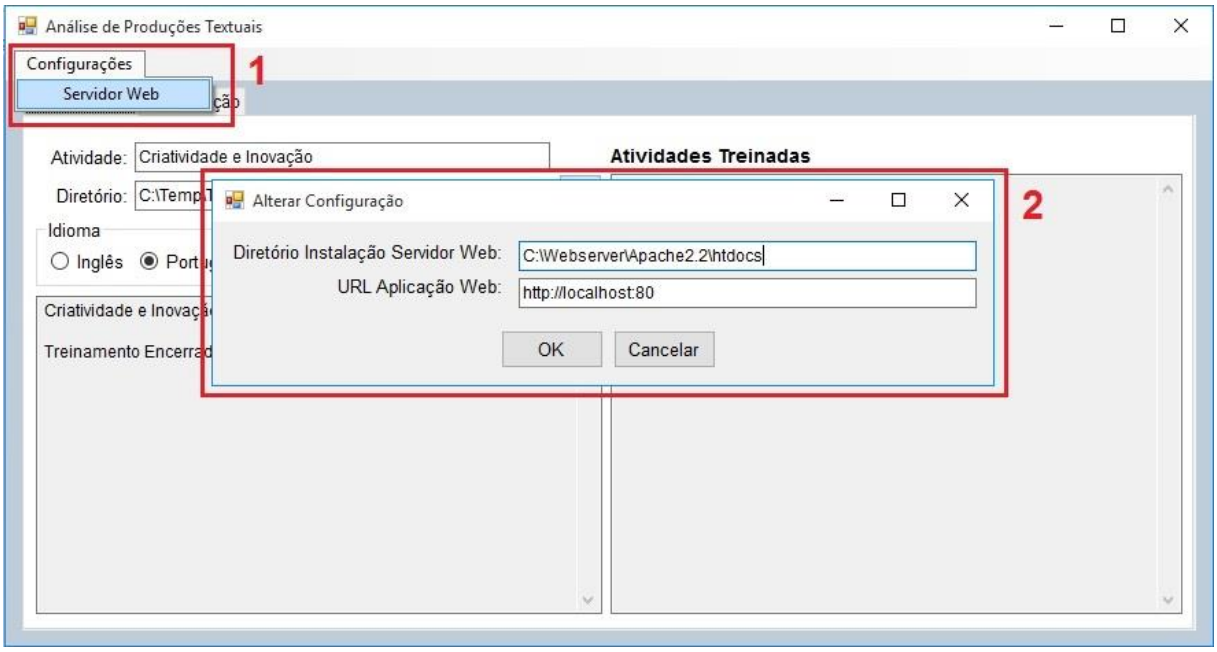
2 UTILIZAÇÃO

Neste capítulo são descritas as funcionalidades do sistema Análise de Produções Textuais, configurações iniciais do uso e detalhamento de cada item da tela. Também são descritos como devem ser disponibilizados os textos para o treinamento da ferramenta e para a análise.

2.1 CONFIGURAÇÔES PARA PRIMEIRO USO

São necessárias duas configurações para o primeiro uso caso o servidor de web tenha sido instalado com alguma parametrização diferente do que o descrito neste manual. Estas configurações são necessárias devido a integração do sistema com as visualizações que serão apresentadas no Navegador padrão do computador.

Figura 58 - Configurações Análise de Produções Textuais.



Conforme Figura 58, é necessário configurar o diretório de instalação do servidor web. O que está na imagem é o caminho padrão (C:\Webserver\Apache2.2\htdocs). Também é necessário configurar a URL da aplicação web. Por padrão é configurado o caminho (http://localhost:80), porta 80 instalada por padrão na instalação do servidor Apache. Após a configuração inicial, o sistema instalado sempre iniciará com o que foi salvo.

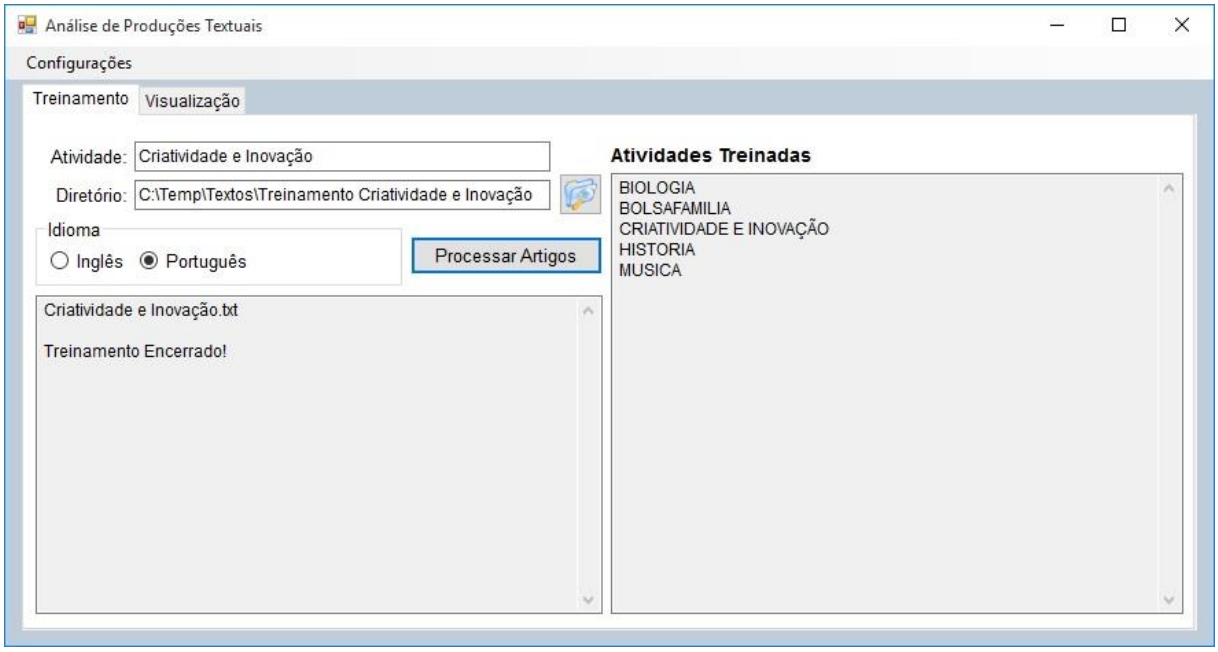
2.2 TREINAMENTO

Descrição das funcionalidades da tela:

a) Atividade: É a descrição da atividade que será treinado o sistema. Esta descrição é importante, porque os textos analisados posteriormente deverão conter o respectivo assunto ao serem importados.
b) Diretório: Deverá ser informado o diretório que contém os artigos em formado *txt* que serão importados no treinamento.
c) Idioma: O sistema está preparado para aceitar textos no idioma Inglês e Português. Importante informar o idioma correto, pois os textos analisados deverão estar no mesmo idioma.

d) Processar Artigos: Irá importar os artigos conforme o diretório informado para a importação. Esta ação irá criar um diretório (C:\listasTermos) caso seja a primeira vez que o sistema seja treinado. Este diretório armazenará os treinamentos.
e) Abaixo do Idioma é apresentada uma listagem com os artigos importados no treinamento que está sendo realizado, assim como uma mensagem informando que o treinamento foi efetuado.
f) Atividades Treinadas: Apresenta uma listagem das atividades que já foram treinadas no sistema.

Figura 59 - Treinamento Análise de Produções Textuais.



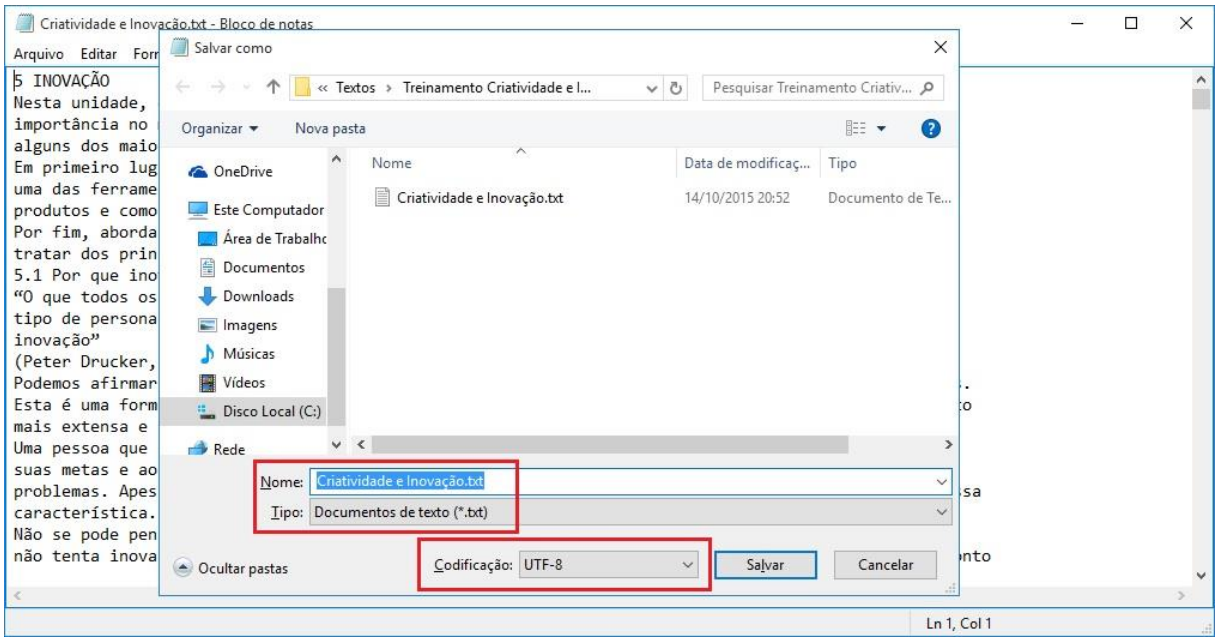
O treinamento de uma atividade não é incremental, ou seja, sempre que se deseja adicionar um texto a uma atividade já treinada a fim de aumentar o nível de conhecimento do sistema sobre determinada atividade, deve-se treinar novamente com todos os textos desta atividade.

2.2.1 Arquivos de Entrada

Tanto os arquivos de entrada para o treinamento quanto para a análise têm que estar no formato TXT e tem que estar salvos utilizando o padrão UTF8.

A Figura 60 representa um arquivo de treinamento sendo salvo conforme os formatos corretos.

Figura 60 - Arquivo Treinamento.



2.3 ANÁLISE

Descrição das funcionalidades da tela:

a) Diretório: Deverá ser informado o diretório que contém os artigos em formado *txt* que serão importados para análise.

b) Iniciar Diagnóstico: Irá importar os artigos conforme o diretório informado para a importação. Após esta ação o sistema está pronto para ser manipulado com o objetivo de analisar os textos importados.

c) Atividades Importadas: Apresenta um resumo dos textos que foram importados, contendo informações sobre a atividade, a turma e os estudante de cada artigo.

d) Seleção: A combinação dos textos para análise é representada na Figura 62. Pode-se selecionar para visualização o texto de um estudante em todas as atividades ou em uma atividade. Ou pode-se selecionar os textos de uma turma inteira em uma atividade ou em todas as atividades.

e) Visualizações: Cada botão representa uma visualização. As visualizações abrirão no Navegador padrão do computador. As visualizações estão disponíveis conforme seleção abaixo:

- Collapsible Tree – Todas as seleções.
- Bubble Chart – Todas as seleções.
- Word Cloud – Apenas um estudante em uma atividade.
- Rotating Cluster – Todas as seleções.
- Treemap – Todas as seleções.
- Bar Chart – Apenas uma turma em uma atividade.

Figura 61 - Análise de Produções Textuais.

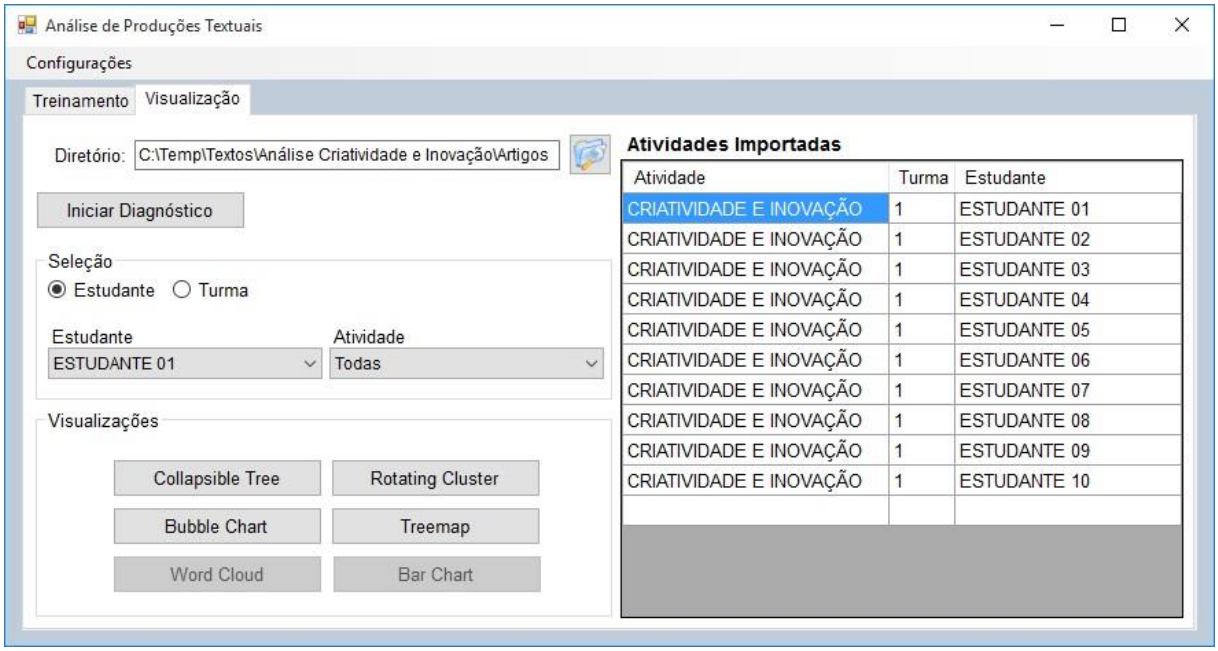
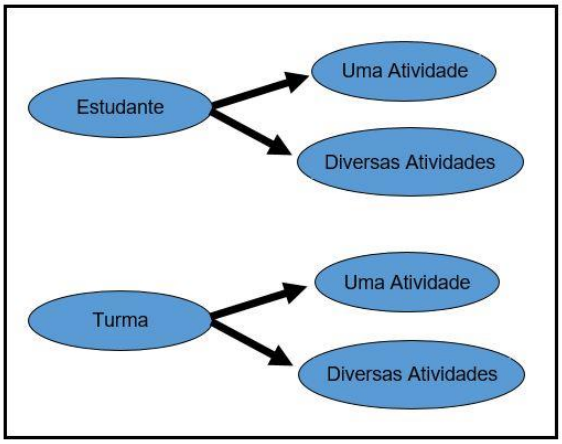


Figura 62 - Combinação para seleção de textos para análise.



2.3.1 Arquivos de Entrada

Os textos para análise, além de estarem salvos em TXT e no formato UTF8 conforme imagem da Figura 60, devem possuir um cabeçalho que os identifique. O cabeçalho deve ser composto conforme abaixo:

Atividade: Criatividade e Inovação

Turma: 1

Estudante: Estudante 01

Importante verificar que a atividade precisa ter exatamente o mesmo nome da atividade que foi colocada no treinamento conforme Figura 59, inclusive com a mesma acentuação, caso utilizado. Um exemplo de cabeçalho e arquivo de entrada para análise é demonstrado na Figura 63.

Figura 63 - Arquivo para análise.