UNIVERSIDADE DE CAXIAS DO SUL

DOUGLAS JACQUES DE OLIVEIRA

EVOLUÇÃO DE UM SOFTWARE DE APOIO A ANÁLISE TEXTUAL: DETECTANDO FALHAS E REALIZANDO MELHORIAS

CAXIAS DO SUL

2017

DOUGLAS JACQUES DE OLIVEIRA

EVOLUÇÃO DE UM SOFTWARE DE APOIO A ANÁLISE TEXTUAL: DETECTANDO FALHAS E REALIZANDO MELHORIAS

Trabalho de Conclusão de Curso para obtenção do grau de Bacharel em Sistemas de Informação da Universidade de Caxias do Sul.

Orientadora: Profª. Drª. Carine Geltrudes Webber

CAXIAS DO SUL

2017

DOUGLAS JACQUES DE OLIVEIRA

EVOLUÇÃO DE UM SOFTWARE DE APOIO A ANÁLISE TEXTUAL: DETECTANDO FALHAS E REALIZANDO MELHORIAS

Trabalho de Conclusão de Curso para obtenção do grau de Bacharel em Sistemas de Informação da Universidade de Caxias do Sul.

Aprovado em /12/2017.

Banca Examinadora

_____
Profª. Drª. Carine Geltrudes Webber
Universidade de Caxias do Sul – UCS

_____
Prof. Dr. Daniel Luis Notari
Universidade de Caxias do Sul – UCS

_____
Profª. Drª. Helena Graziottin Ribeiro
Universidade de Caxias do Sul – UCS

A todos, que de alguma forma me ajudaram ou apoiaram no decorrer desta jornada, meu agradecimento e gratidão.

AGRADECIMENTOS

Agradeço primeiramente à minha orientadora, Profª. Carine, pelo seu tempo e esforço dedicado aos nossos encontros, pelos momentos de ajuda e instrução, sem os quais seria impossível a realização deste trabalho.

Agradeço aos meus pais, Joacir e Maria Lúcia, pelo apoio, motivação e compreensão pelos momentos que estive ausente durante este período de graduação.

“A mente que se abre a uma nova ideia jamais voltará ao seu tamanho original.”

Albert Einstein

RESUMO

O presente trabalho realiza um estudo sobre aprendizagem automática em dados textuais aplicada na área da educação. No ramo educacional existem muitas produções textuais que precisam ser manualmente analisadas pelos professores. Existe a necessidade do desenvolvimento de softwares que auxiliem nos processos de leitura e revisão de textos. Estes softwares podem auxiliar tanto estudantes quanto professores na percepção de aprendizagem do que foi estudado. Alguns softwares e trabalhos relacionados já foram desenvolvidos, como a ferramenta Análise de Produções Textuais (LOVATO, 2015; GIL, 2016), que foi utilizada neste estudo. Apesar de analisar textos e apresentar visualizações gráficas dos resultados desta análise, esta ferramenta pode ser aprimorada em alguns aspectos. Dentre estes aspectos destacam-se melhorias na precisão de algoritmos e na inserção de novas formas de visualização. Com base nisso, foram implementadas melhorias evolutivas para a ferramenta, refinando as técnicas e algoritmos empregados, de modo a trazer resultados mais completos e úteis para o usuário. Dentre as melhorias, destaca-se a inclusão de novas formas de visualização dos resultados. Testes preliminares indicam que as melhorias vão favorecer a compreensão do professor sobre o desempenho dos estudantes.

Palavras-chave: mineração de textos, análise de textos educacionais, visualização de informações.

ABSTRACT

This paper aims to accomplish a study on automatic learning in textual data applied to the educational area. In the educational field there are many textual productions that need to be manually analyzed by the teachers. There is a need for the development of software that helps in the processes of reading and proofreading texts. These softwares can assist both students and teachers in the perception of learning than has been studied. Some software and related works have already been developed, such as the tool Análise de Produções Textuais (LOVATO, 2015; GIL, 2016), which was used in this study. Although analyzing texts and presenting graphical visualizations of the results of this analysis, this tool can be improved in some aspects. Among these aspects, we highlight improvements in the accuracy of algorithms and the insertion of new forms of visualization. Based on this, evolutionary improvements were developed for the tool, refining the techniques and algorithms used, in order to bring more complete and useful results to the user. Among these improvements, we highlight the inclusion of new forms of data visualization. Preliminary tests indicate that these improvements will enhance the teacher understanding of student performance.

Keywords: text mining, analysis of educational texts, data visualization.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

APT	Análise de Produções Textuais
IDE	*Integrated Development Environment*
KNN	*K-Nearest Neighbor*
JSON	*JavaScript Object Notation*
LSA	*Latent Semantic Analysis*
SVM	*Support Vector Machine*
SVD	*Singular Value Decomposition*
SWOT	*Strengths, Weaknesses, Opportunities, and Threats*
UCS	Universidade de Caxias do Sul

SUMÁRIO

1 INTRODUÇÃO

Por meio da popularização da internet e aumento do número de usuários de computador, surge a oportunidade de automatizar processos do cotidiano, tornando atividades manuais em procedimentos automáticos. Isso também acontece na área da educação, onde há uma grande quantidade de informação disponível, na maioria das vezes de forma textual (SILVA, 2004).

Nos dias de hoje acredita-se que competências relativas à produção e interpretação de textos são ligadas especificamente aos seres humanos. No entanto, em razão das evoluções na área da computação cognitiva, as máquinas também poderão desempenhá-las (IBM, 2017; MICROSOFT, 2017). Realizando uso das tecnologias de aprendizado de máquina, em conjunto com ferramentas para representação do conhecimento, processamento de linguagem natural e raciocínio automático, vários sistemas começam a despontar (e. g. Watson (IBM, 2017)). Um desafio na área em questão compreende em fornecer interpretações para amplos volumes de textos de maneira automática e de fácil entendimento ao ser humano. Isso demanda análise e sumarização de texto, além de realizar a utilização de representações visuais adequadas e de forma correta (LOVATO, GIL, *et al.*, 2016).

A área do aprendizado de máquina pode contribuir com a tarefa de análise textual. O aprendizado de máquina compreende um conjunto de técnicas para classificação de dados e formação de agrupamentos de dados por similaridade. Ela envolve o processo de extrair padrões, classes e grupos através de um conjunto de textos usados para treinamento.

Encontrando padrões nos textos por meio das técnicas de aprendizagem automática, é possível categorizar os textos com maior precisão. Deste modo podem-se realizar comparações e análises através de estatísticas geradas. Estas estatísticas necessitam ser visualizadas com o objetivo de ajudar na compreensão. Assim se fazem necessários meios de visualização dos dados analisados, buscando auxiliar o entendimento de informações que em sua forma bruta são de difícil compreensão.

Lovato (2015) e Gil (2016) desenvolveram a ferramenta Análise de Produções Textuais. Esta ferramenta opera em duas etapas. A primeira etapa é a de treinamento, onde o software realiza o aprendizado a partir de textos fornecidos pelo professor. A segunda etapa consiste na análise dos textos desenvolvidos pelos alunos. Os resultados processados são disponibilizados por meio de diversas formas de visualizações gráficas. O sistema calcula o limiar do documento, que indica e similaridade entre o corpus de texto fornecido pelo professor com as atividades dos estudantes. O software utiliza o algoritmo de análise

semântica latente para realizar a análise dos textos. A análise semântica latente é uma abordagem utilizada para extrair relações entre documentos e os termos neles contidos.

Dentre as várias etapas que compõem o algoritmo implementado está a indexação do documento de treinamento, onde *stopwords* são filtrados e os termos restantes inseridos em uma *HashTable*. Quanto mais precisa essa filtragem, melhor o resultado final do processo. Se por exemplo, forem identificados novos *stopwords* a serem inseridos ou *stopwords* indevidamente implementados no algoritmo forem removidos, obtém-se um resultado mais preciso. As melhorias detectadas poderão afetar igualmente as interfaces do software (interface do usuário e de visualização dos resultados). Neste escopo é que se insere este trabalho, visando identificar otimizações na implementação da ferramenta, trazendo resultados com maior exatidão para o usuário.

1.1 OBJETIVOS

O objetivo geral deste trabalho é refinar as técnicas e algoritmos empregados na implementação de uma ferramenta existente para análise de produções textuais, trazendo resultados mais completos.

Para que seja cumprido o objetivo geral apresentado, o trabalho deverá cumprir os seguintes objetivos específicos:

1. conhecer a implementação do software de apoio à análise textual;
2. identificar funções que possam ser melhoradas e propor melhorias;
3. implementar as modificações propostas;
4. definir meios de comparação e métricas necessárias;
5. coletar textos experimentais de trabalhos desenvolvidos por estudantes;
6. realizar os testes com a inserção dos textos disponibilizados;
7. verificar e comparar os resultados.

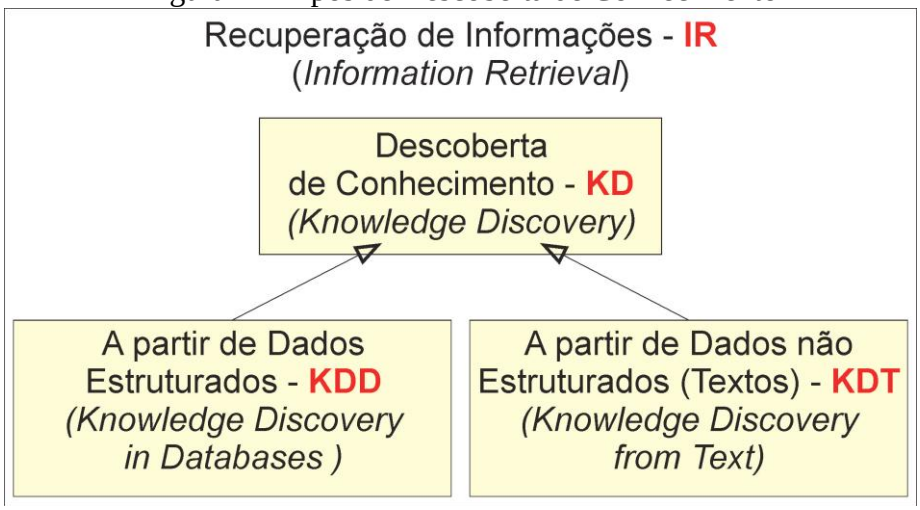
1.2 ESTRUTURA DO TRABALHO

O presente trabalho está estruturado em cinco capítulos. O capítulo 2 trata dos conceitos de mineração de texto, apresentando aplicações e as principais técnicas utilizadas. O terceiro capítulo apresenta a ferramenta de análise de produções textuais, desenvolvida por Lovato (2015) e Gil (2016) e serviu de base para o desenvolvimento deste trabalho. O quarto capítulo apresenta as melhorias implementadas, testes realizados e análise dos resultados obtidos. O último capítulo apresenta as considerações finais, realizando uma síntese do trabalho, mostrando suas contribuições e sugerindo ideias para trabalhos futuros.

2 APRENDIZAGEM AUTOMÁTICA EM DADOS TEXTUAIS

De acordo com Feldman e Sanger (2006), o conceito de aprendizagem automática em dados textuais (*Text Mining* ou Mineração de Textos) pode ser definido como um processo de conhecimento intensivo no qual o usuário interage com uma coleção de documentos através de uma ou mais interfaces de ferramentas de análise. Semelhante à mineração de dados, a mineração de texto busca encontrar informações úteis a partir de textos (dados não estruturados) por meio da identificação e exploração de padrões estabelecidos. Análise de dados em formato não estruturado pode ser considerada uma tarefa mais complexa se comparada à análise de dados estruturados, justamente pelos dados não possuírem a característica de estruturação. Assim se fazem necessárias técnicas e ferramentas específicas para mineração deste tipo de dados. Este conjunto de técnicas e ferramentas faz parte da área de Descoberta do Conhecimento em Textos (*Knowledge Discovery from Text - KDT*) (MORAIS e AMBRÓSIO, 2007).

Figura 1 – Tipos de Descoberta de Conhecimento



Fonte: Morais e Ambrósio (2007).

2.1 ETAPAS DA MINERAÇÃO DE TEXTOS

Segundo Aranha (2007), o processo de mineração de texto é dividido em cinco etapas (Figura 2). Na primeira etapa é realizada a coleta de textos para a formação de bases textuais. Os dados podem ser coletados de diversas fontes dependendo da necessidade, tais como disco local do computador, dispositivos removíveis, bancos de dados e internet.

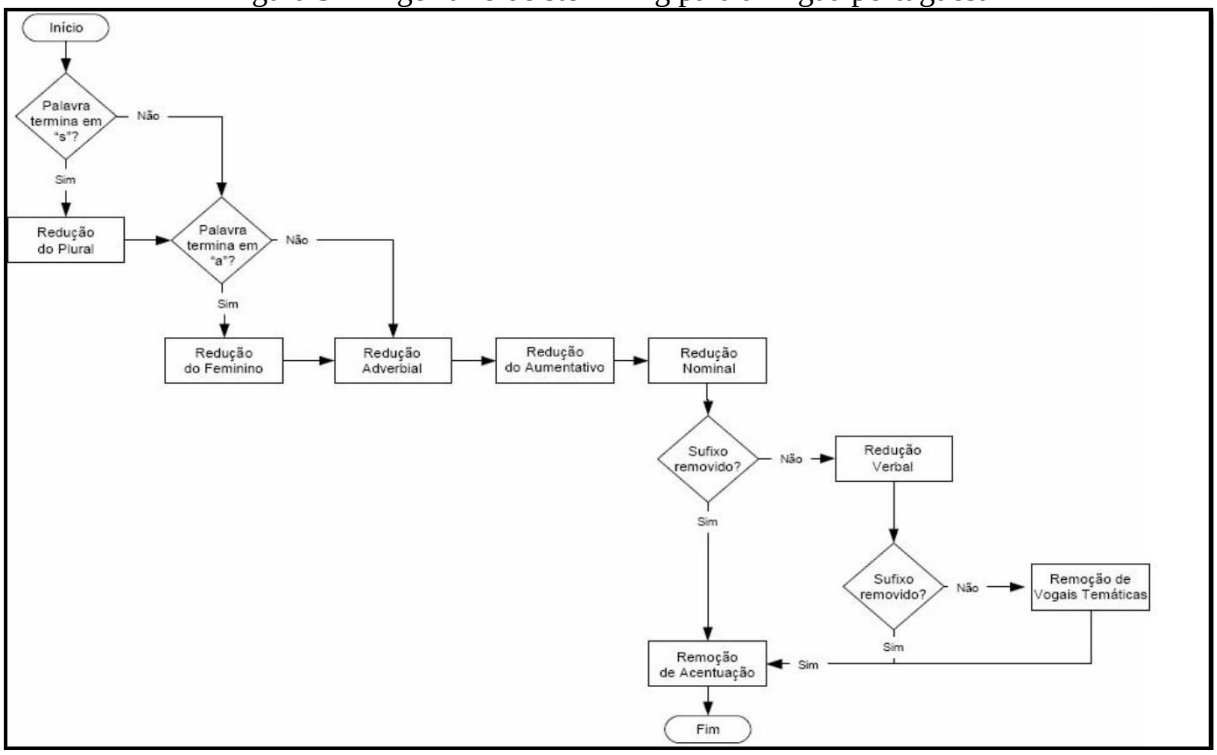
Figura 2 – Processo de Mineração de Textos

Fonte: Aranha (2007).

A segunda etapa do processo é o pré-processamento. Esta etapa busca transformar estes textos em informações mais estruturadas, realizando uma organização e formatação, assim melhorando sua qualidade para as que etapas que se sucedem.

A etapa de indexação é separada em três fases. Na primeira fase se faz a separação dos termos simples dos termos compostos. A segunda fase compõe a remoção das *stopwords*, que são palavras que não possuem relevância para o texto. Na terceira fase acontece a normalização morfológica, que consiste em extrair o radical das palavras através da eliminação do prefixo, sufixo, número, grau e gênero das mesmas. Este processo de redução de uma palavra ao seu radical em inglês é chamado de *stemming*. Um algoritmo de *stemming* para a língua portuguesa foi desenvolvido por Orengo e Huyck (2001). Esse algoritmo foi implementado em linguagem de programação C e é composto por oito etapas (Figura 3).

Figura 3 – Algoritmo de stemming para a língua portuguesa



Fonte: Orengo e Huyck (2001).

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

1. remoção do plural – Consiste na remoção do "s" no final das palavras. Porém, existe uma lista de exceções, como as palavras "lápis" e "tênis", por exemplo;
2. remoção do feminino – As palavras no gênero feminino serão transformadas em sua correspondente no masculino. Por exemplo, "americana" será convertido para "americano";
3. remoção do advérbio – Como a maioria dos advérbios termina com o sufixo "mente", este sufixo será retirado das palavras. Contudo, para esta etapa também existe uma lista de exceções;
4. remoção do aumentativo e diminutivo – Remove os sufixos que indicam aumentativo e diminutivo de substantivos e adjetivos como, por exemplo, "cachorrão" e "espertinho";
5. remoção de sufixos em nomes – Há uma lista com 61 sufixos para substantivos e adjetivos. O objetivo desta etapa é testar a palavra identificando se ela possui algum sufixo presente nesta lista. Caso possua, o sufixo é removido e a execução do algoritmo pula direto para o passo oito. Caso não encontre possível remoção, o algoritmo segue ao passo seis;

6. remoção de sufixos em verbos – A estrutura de um verbo pode ser definida por radical + vogal temática + tempo + pessoa. É importante salientar que a conjugação verbal pode variar conforme tempo, pessoa, modo e número. Portanto, nesta etapa a função é reduzir a forma verbal ao seu radical correspondente;
7. remoção de vogais temáticas - a vogal temática possibilita o recebimento de flexões à palavra, podendo ser classificada como vogal temática verbal ou nominal. As verbais são caracterizadas por "A" na primeira conjugação verbal (ar – ex: caminh**ar**), "E" na segunda conjugação verbal (er – ex: escond**er**) e "I" na terceira conjugação verbal (ir – ex: part**ir**). As nominais são representadas pelas vogais "A" (ex: escol**a**), "E" (ex: test**e**) e "O" (ex: cadern**o**). Nessa etapa será removida a vogal temática das palavras que não foram alteradas pelas etapas cinco e seis;
8. remoção de acentos – Nesta última etapa, será removida a acentuação das palavras.

A quarta etapa do processo de mineração de texto compõe a extração do conhecimento, aplicando algoritmos de acordo com o objetivo que se deseja. Por exemplo, se a aplicação é voltada ao ramo educacional, define-se a relevância dos termos e o objetivo que se deseja chegar nesta área. As palavras possuem diferentes níveis de relevância no texto, por isso é importante o cálculo da relevância dos termos, que é embasado também na frequência com que os termos aparecem no texto. As três frequências mais comuns são a frequência absoluta, a frequência relativa e a frequência inversa de documentos.

A frequência absoluta condiz a quantidade total de vezes que um termo aparece em um documento. A frequência relativa é a frequência absoluta divida pelo número de termos contidos no documento, ou seja, leva em consideração o tamanho do texto. A frequência inversa de documentos utiliza as duas frequências anteriores para ser calculada, podendo assim diminuir a importância de termos que aparecem muitas vezes e aumentar a importância de termos que aparecem poucas vezes, uma vez que, geralmente, termos que aparecem poucas vezes têm a característica de descrever melhor o assunto abordado.

Na quinta e última etapa é realizada a leitura, análise e interpretação dos resultados obtidos. É nesta etapa que serão apresentadas as informações para que sejam obtidas conclusões sobre a mineração de textos realizada. Essa análise não é realizada de forma computacional, mas sim pelos usuários especialistas na área dos textos coletados. É o

especialista que averiguará a compatibilidade dos resultados obtidos com o conhecimento disponível e o usuário irá verificar a aplicabilidade dos resultados (ARANHA, 2007).

2.2 FERRAMENTAS RELACIONADAS

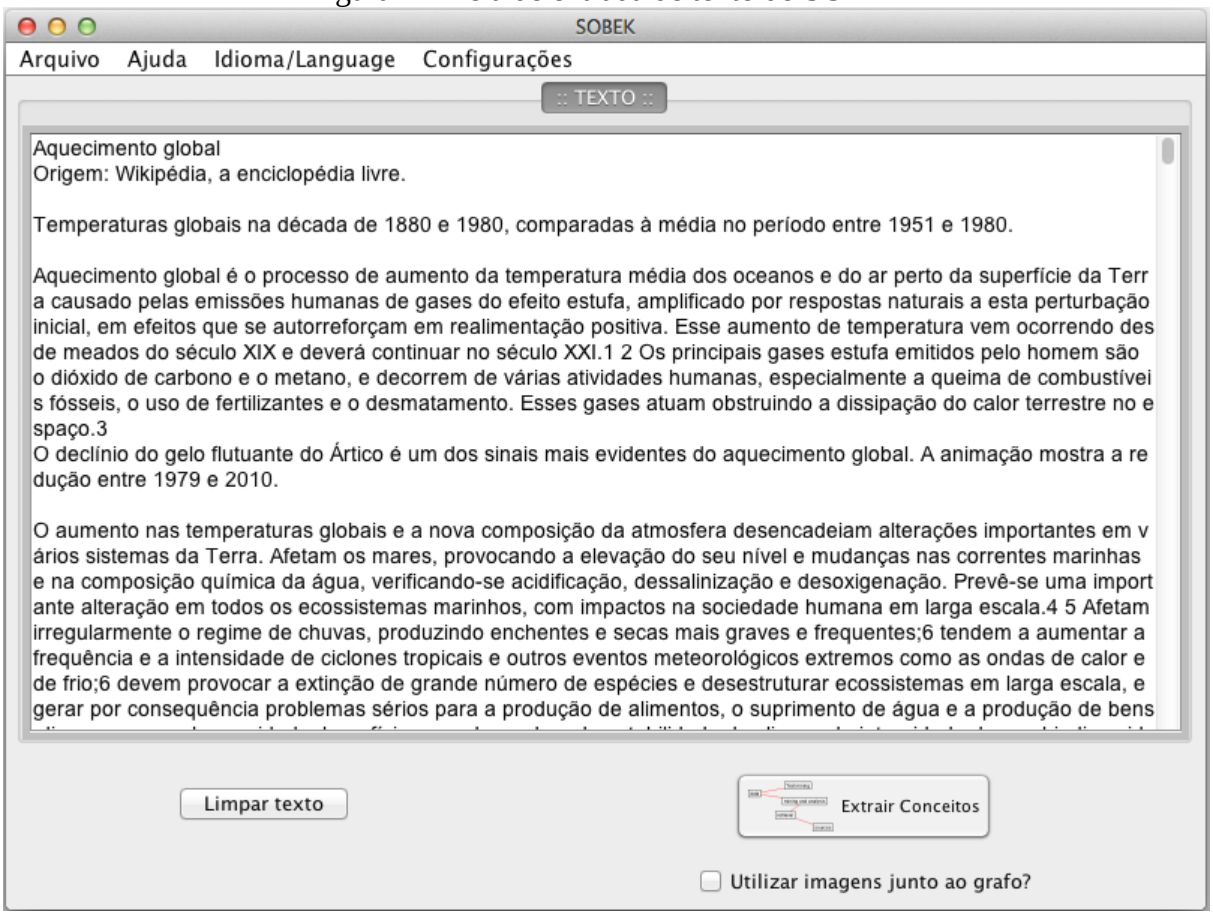
A mineração de textos vem sendo utilizada na área da educação sob forma de ferramentas com o objetivo de auxiliar os estudantes na produção textual. Conforme Silva (2009), os estudantes possuem muita dificuldade em se expressar de forma textual, o que não acontece com a expressão oral, uma vez que a linguagem oral não exige maior formalidade, podendo ser utilizada a linguagem coloquial e gestos para facilitar o entendimento.

A seguir seguem alguns exemplos de ferramentas de mineração utilizadas na área da educação:

- SOBEK (SOBEK, 2017)
- TagCrowd (TAGCROWD, 2017)
- WordCounter (WORDCOUNTER, 2017)
- Textalyzer (TEXTALYSER, 2017)

O SOBEK é uma ferramenta de mineração de texto cuja sua função é a extração dos termos mais frequentes em documentos, encontrando relacionamentos entre estes. Desta maneira ele serve como auxílio aos professores na avaliação de trabalhos escritos (MACEDO, REATEGUI, *et al.*, 2009). Esta ferramenta destaca-se por mostrar de forma gráfica as relações entre os termos importantes nos documentos, ajudando os estudantes a obterem suas próprias reflexões sobre os textos lidos.

Figura 4 – Tela de entrada de texto do SOBEK

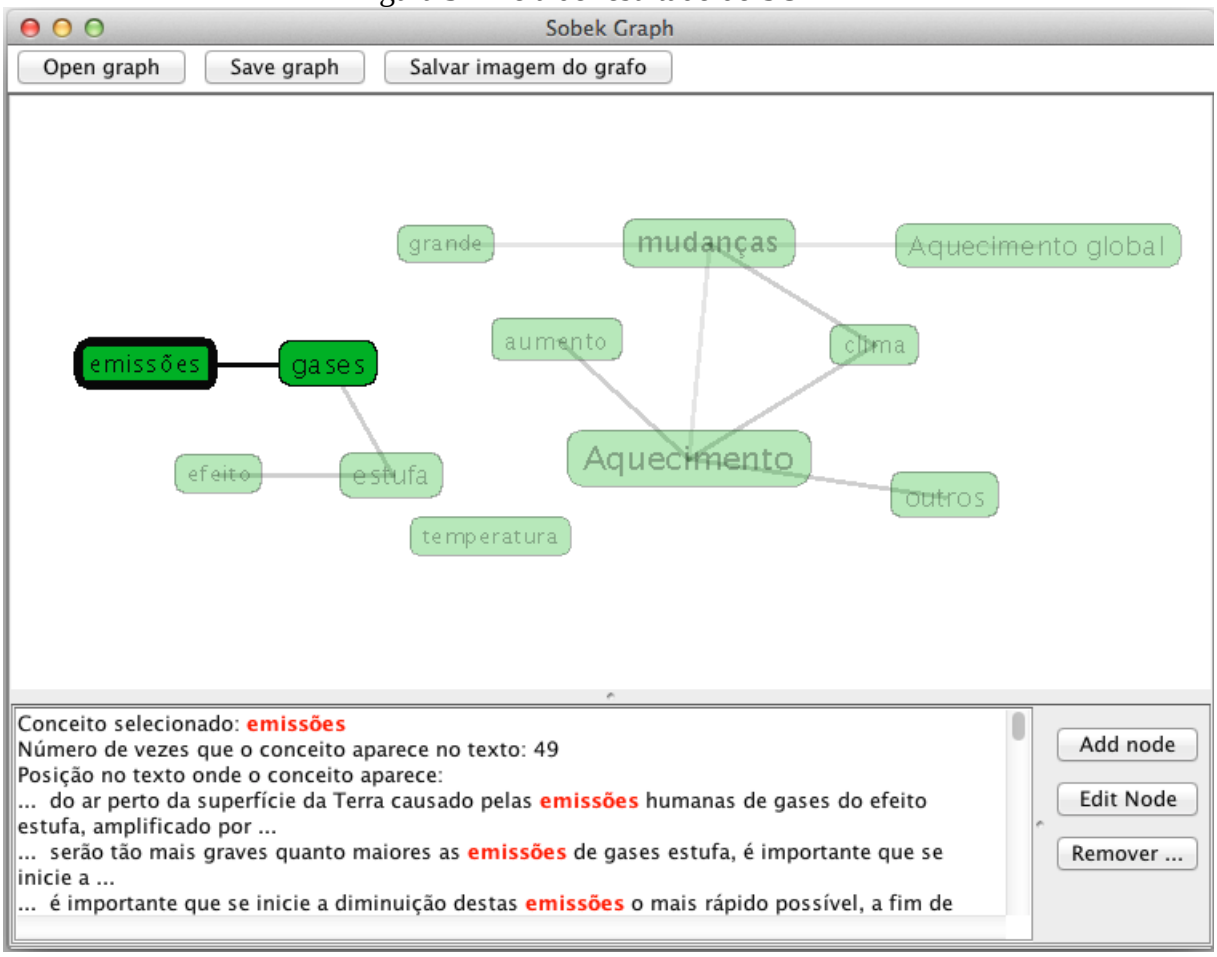


Fonte: http://sobek.ufrgs.br/uploads/2/3/3/9/23394804/sobek_quick_reference_guide_pt.pdf.

Na tela inicial do programa SOBEK (Figura 4) é possível digitar ou colar o texto a ser analisado. Também há possibilidade de importar o texto via arquivo nos formatos *txt*, *doc* ou *pdf* .

Na próxima etapa ocorre a extração dos principais conceitos do texto, gerando um grafo com os principais termos e os relacionamentos entre eles. É possível editar os nodos conforme a necessidade (Figura 5).

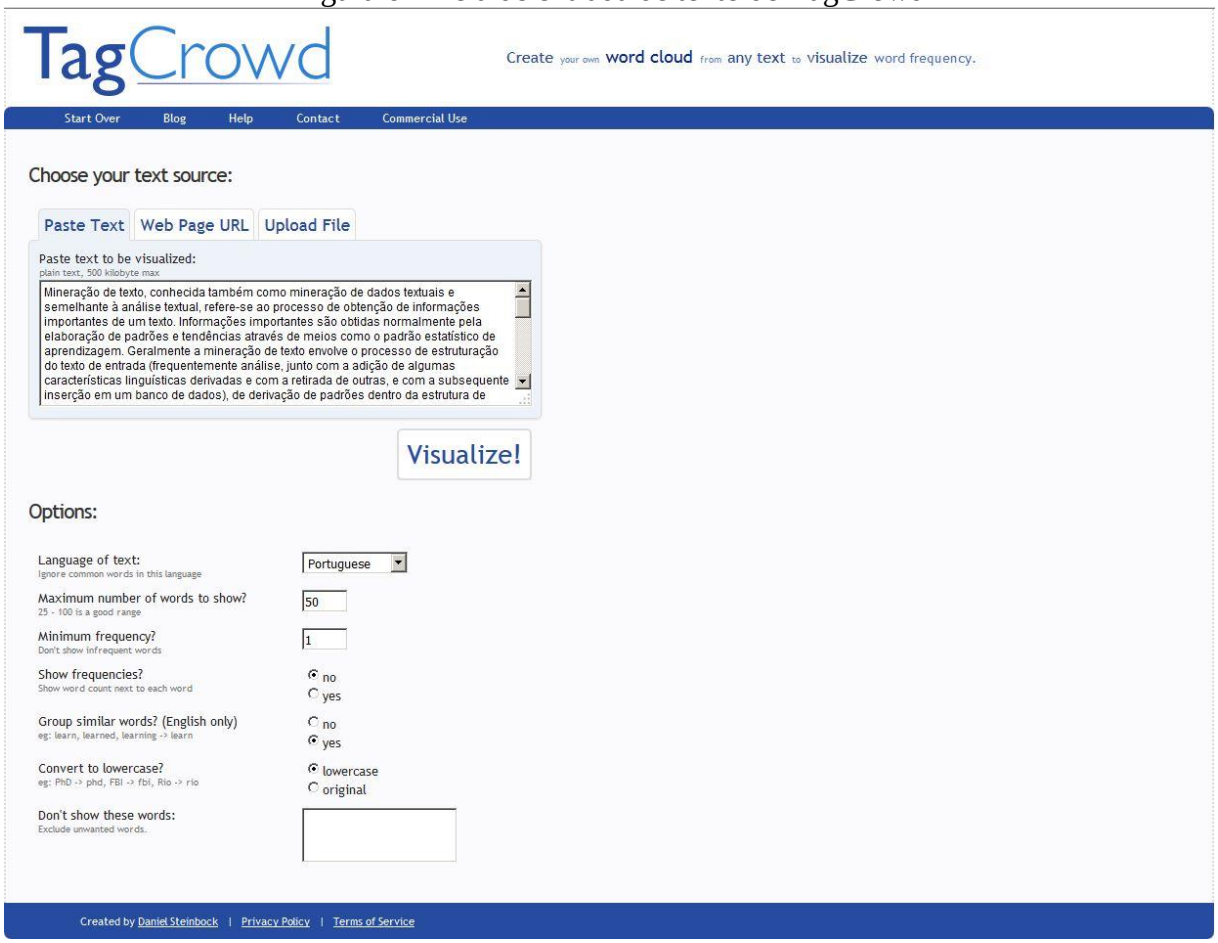
Figura 5 – Tela de resultado do SOBEK



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

O TagCrowd (TAGCROWD, 2017) é uma aplicação web que permite criar uma representação gráfica em estilo de nuvem de palavras, destacando as palavras mais relevantes de acordo com a frequência que aparecem no texto. Os textos podem ser digitados os colados na tela. Também é possível informar uma página web que contenha um texto a ser analisado ou realizar o *upload* de um arquivo de texto (Figura 6). Porém, esta ferramenta não mostra relação entre os termos e não faz a remoção das *stopwords* (Figura 7).

Figura 6 – Tela de entrada de texto do TagCrowd



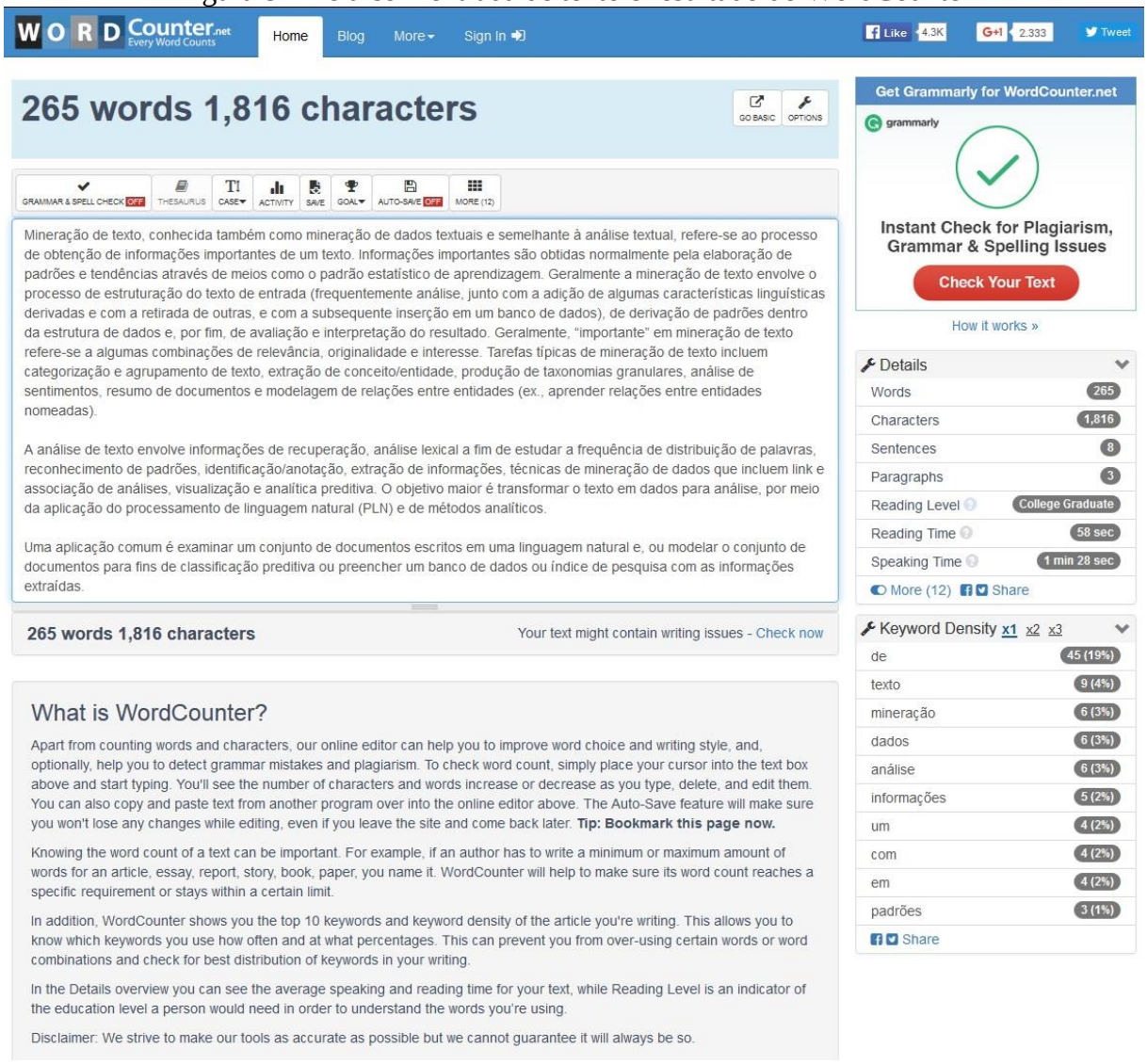
Fonte: http://tagcrowd.com.

Figura 7 – Tela de resultado do TagCrowd

Fonte: http://tagcrowd.com.

Outra ferramenta online é o WordCounter. Esta tem como objetivo apresentar uma relação das palavras mais frequentes em um texto. De acordo com Klemann, Reategui e Rapkiewicz (2011), este programa é muito útil, pois apresenta palavras repetidas ou redundantes no texto em uma listagem (Figura 8).

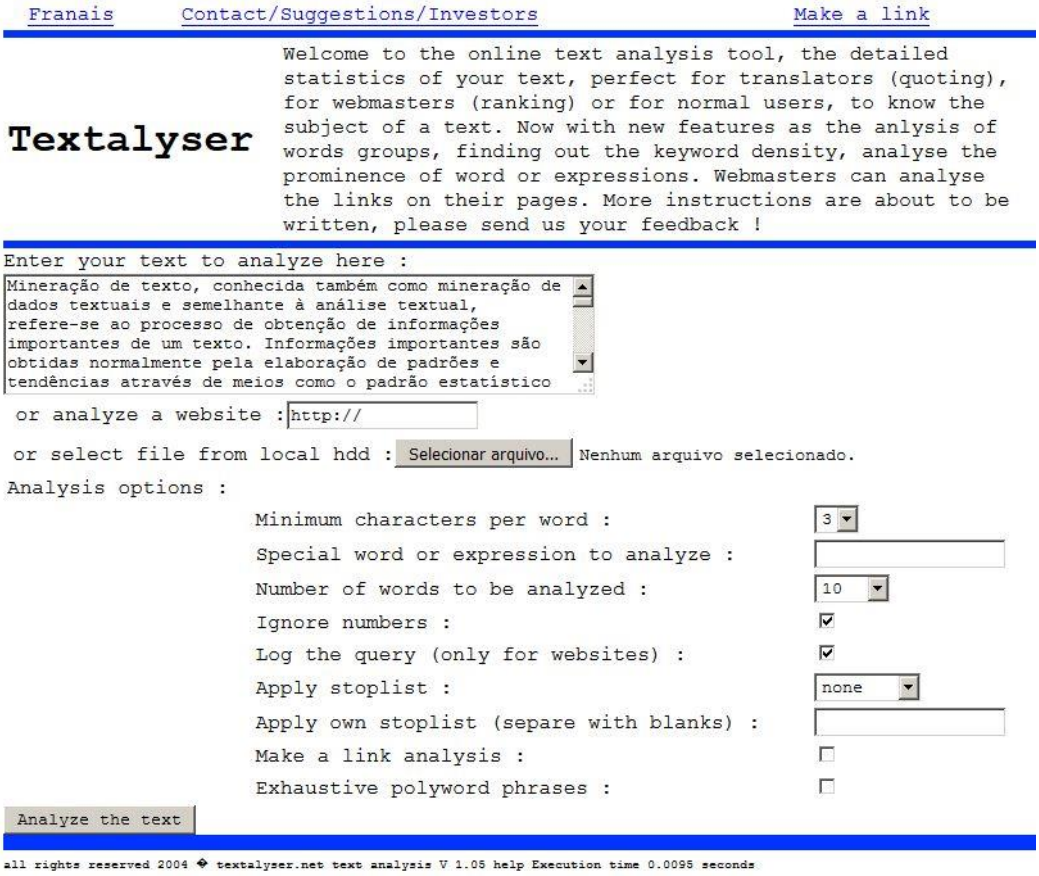
Figura 8 – Tela com entrada de texto e resultado do WordCounter



Fonte: https://wordcounter.net.

De maneira semelhante, o Textalyzer (TEXTALYSER, 2017) é uma ferramenta web que analisa palavras e expressões contidas no texto, mostrando a contagem e algumas estatísticas dos termos e palavras presentes no documento. Esta ferramenta apresenta também um índice de legibilidade do texto analisado, usando como critérios o tamanho das frases e as estatísticas obtidas pelo analisador. Esta aplicação mostra o resultado de dados estatísticos, mas não gera uma interface gráfica elaborada para isto. A seguir na Figura 9 é mostrada a tela de entrada de texto do Textalyzer, enquanto que na Figura 10 é apresentada a tela de resultado, mostrando estatísticas tais como: contagem de palavras, número de palavras diferentes e número de caracteres.

Figura 9 – Tela de entrada de texto do Textalyzer



Fonte: http://textalyser.net.

Figura 10 – Tela de resultado do Textalyzer

Textalyser Results

The complete results, incuding compexity factor, and other features

Total word count :	174
Number of different words :	118
Complexity factor (Lexical Density) :	67.8%
Readability (Gunning-Fog Index) : *(6-easy 20-hard)*	25.7
Total number of characters :	1928
Number of characters without spaces :	1617
Average Syllables per Word :	2.5
Sentence count :	9
Average sentence length (words) :	36.33
Max sentence length (words) :	70
(geralmente a mineração de texto envolve o processo de estruturação do texto de entrada frequentemente análise junto com a adição de algumas características linguísticas derivadas e com a retirada de outras e com a subsequente inserção em um banco de dados de derivação de padrões dentro da estrutura de dados e por fim de avaliação e interpretação do resultado)	
Min sentence length (words) :	6
(aprender relações entre entidades nomeadas)	
Readability (Alternative) beta : *(100-easy 20-hard, optimal 60-70)*	-41.9

Fonte: http://textalyser.net.

Klemann, Reategui e Rapkiewicz (2011) apresentam uma tabela comparando estas ferramentas de mineração de texto, mostrando os objetivos e características de cada uma (Tabela 1). A respeito da disponibilidade via web, somente o SOBEK não possui versão online. Com relação a contagem de termos, as aplicações Textalyzer e WordCounter possuem esta função, enquanto a SOBEK e TagCrowd não a possuem. Quanto a apresentação dos termos analisados, somente o programa WordCounter apresenta todos os termos contidos no texto, enquanto o SOBEK e o TagCrowd apresentam os termos mais relevantes. O software TagCrowd é o único que não apresenta a frequência dos termos. O SOBEK é a única ferramenta capaz de analisar e mostrar o relacionamento entre os termos. Referente à apresentação dos resultados, as ferremantas SOBEK e TagCrowd possuem uma visualização gráfica dos termos, sendo que o SOBEK também possui uma visualização gráfica dos relacionamentos entre eles, enquanto os softwares WordCounter e Textalyzer não possuem nenhuma representação gráfica.

Tabela 1 - Análise comparativa das ferramentas de mineração de texto

	Online	Contagem de termos	Apresenta-ção de todos os termos	Apresenta-ção de termos relevantes	Freqüência dos termos	Relacio-namentos dos termos	Visualiza-ção gráfica dos termos	Visualização gráfica dos termos e relacionamentos
TextAlyser	X	X			X			
Wordcounter	X	X	X		X			
TagCrowd	X			X			X	
Sobek				X	X	X	X	X

Fonte: Klemann, Reategui e Rapkiewicz (2011).

2.3 MÉTODOS DE APRENDIZAGEM AUTOMÁTICA

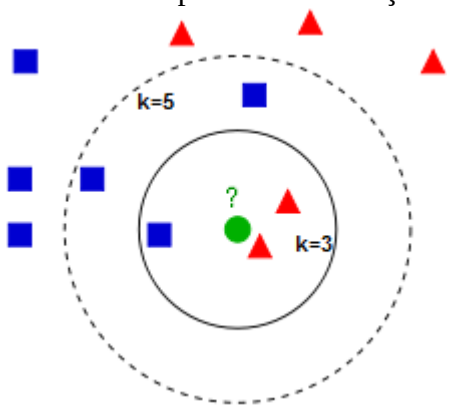
Existem diversos algoritmos utilizados para mineração de textos. Dentre os mais populares, podemos citar o K-Nearest Neighbor (KNN), o Classificador Bayesiano, a Máquina de Suporte Vetorial (SVM) e a Análise de Semântica Latente (LSA).

2.3.1 K-Nearest Neighbor

O K-Nearest Neighbor (KNN) é um dos mais simples algoritmos de aprendizado de máquina (WANG e ZHAO, 2012). O objetivo deste algoritmo é classificar objetos dentro de uma classe predefinida de um grupo de amostra criado por aprendizado de máquina. Este método de aprendizagem é baseada em instância, ou aprendizado tardio (*lazy learning*), onde a função é definida apenas localmente, e toda a execução é deferida até a classificação. O

algoritmo KNN não requer dados de treinamento para realizar a classificação, dados de treinamento podem ser usados durante a fase de teste. A Figura 11 mostra um exemplo de classificação KNN. A amostra de teste (círculo verde) deve ser classificada entre a primeira classe (de quadrados azuis) ou a segunda classe (de triângulos vermelhos). Se k = 3 (círculo de linha contínua) então é classificado como segunda classe, pois há dois triângulos e apenas um quadrado dentro deste círculo. Se k = 5 (círculo de linha tracejada) então é classificado como primeira classe, porque há três quadrados contra dois triângulos considerando este círculo externo.

Figura 11 – Exemplo de classificação KNN



Fonte: elaborado pelo autor.

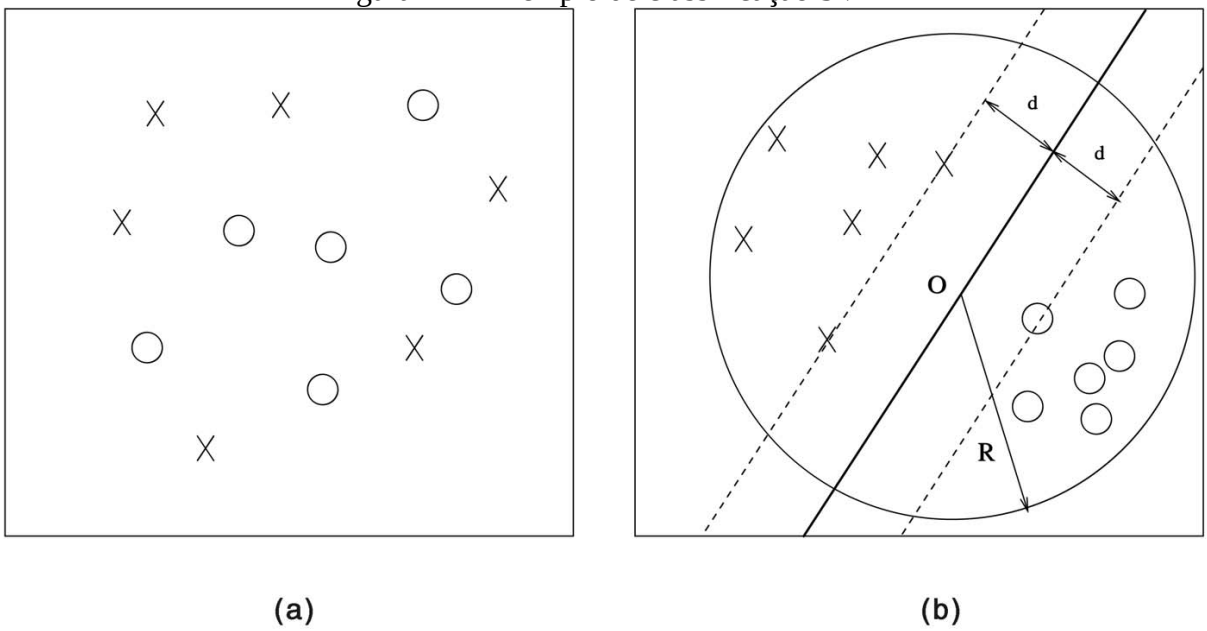
Esta técnica estima a distância entre duas frases, realizando a comparação e classificando o texto conforme a distância, onde X e Y representam as instâncias de dados e d representa a distância entre X e Y. Dentre as vantagens do algoritmo estão a precisão da classificação e implementação simples. De suas desvantagens podemos citar a necessidade de utilização de grande espaço em memória na máquina a ser executado e tempo de execução.

2.3.2 Máquina de Suporte Vetorial

De acordo com Tan e Wang (2004), máquinas de suporte vetorial (SVM) são métodos de aprendizagem supervisionados com algoritmos de aprendizagem associados para análise de dados e reconhecimento de padrões. Esta técnica é utilizada em classificadores e analisadores por regressão. Considerando um conjunto de exemplos de treinamento marcados como pertencendo a uma determinada classe, o algoritmo SVM cria um modelo que indica a qual classe um exemplo novo pertence, o que faz dele um algoritmo classificador linear não probabilístico.

Um padrão SVM é a reprodução dos exemplos como pontos num espaço, mapeados para que as amostras de cada classe sejam afastadas por uma faixa mais larga possível. Então são mapeados novos dados no mesmo espaço e é realizada a previsão para pertencer a uma classe baseada em qual lado da faixa se enquadram A Figura 12 apresenta um conjunto de treinamento, com os símbolos “x” e “o” representando classes diferentes no quadro esquerdo (a). E no quadro direito (b) o conjunto de treinamento transformado, com uma linha de margem máxima separando os dados, estes residindo no espaço de uma esfera de raio R.

Figura 12 – Exemplo de classificação SVM



Fonte: Tan e Wang (2004).

Se porventura os dados não são rotulados, o aprendizado supervisionado não é possível. Neste caso o aprendizado não supervisionado é utilizado, fazendo com que o algoritmo gere o agrupamento natural dos dados, mapeando dados novos nestes agrupamentos.

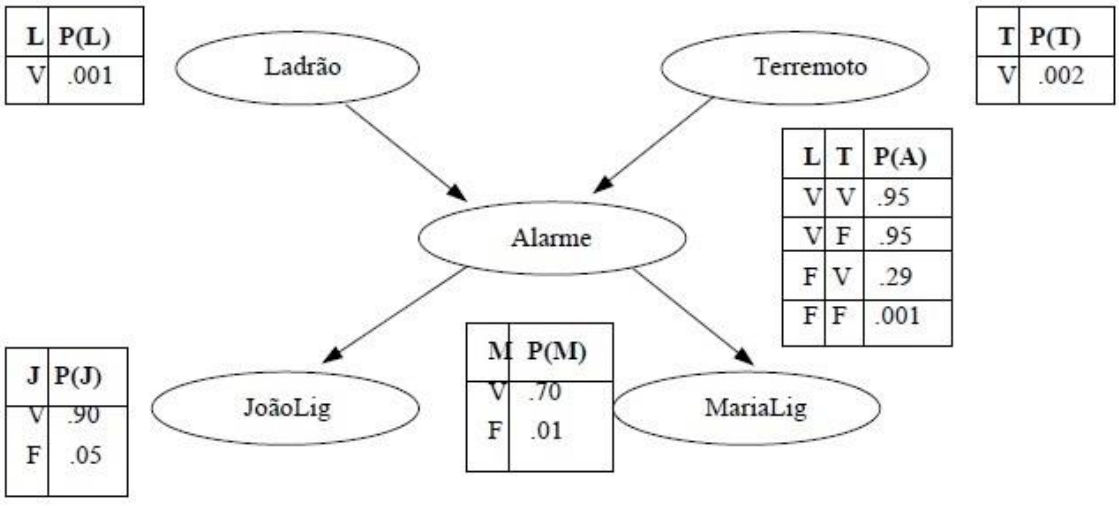
2.3.3 Classificador Bayesiano

Segundo Russell e Norvig (2013), classificadores Bayesianos são classificadores estatísticos que agrupam objetos em uma determinada classe, baseando-se na probabilidade deste objeto pertencer a esta classe.

Os classificadores Bayesianos Simples ou Ingênuos (*Naive Bayes*) se baseiam na hipótese de que o efeito do valor de um atributo não-classe independe dos valores dos demais atributos não-classe. Ou seja, todos os atributos são independentes entre si, sendo que o valor de um atributo não influencia no valor dos outros (TAN, STEINBACH e KUMAR, 2009).

Conforme Han e Kamber (2006), classificadores Bayesianos são largamente utilizados em tarefas de classificação em função do bom desempenho computacional obtido. Na Figura 13 é apresentado um exemplo de rede Bayesiana com suas devidas relações de causa, efeito e dependências. Esta rede Bayesiana apresenta a seguinte situação: há um alarme anti-arrombamento instalado em uma casa. Ele é bastante confiável na detecção de arrombamentos, mas às vezes dispara quando ocorrem pequenos terremotos. O dono desta casa possui dois vizinhos, João e Maria, que concordaram em ligar para o dono caso o alarme toque quando ele não está em casa. João quase sempre liga quando ouve o alarme, mas às vezes ele confunde o som do telefone com o alarme, e liga. Maria ouve música alta com frequência, e às vezes não ouve o alarme tocar. A probabilidade a priori de acontecer um arrombamento, P(L), é 0,1%. Já a de acontecer um terremoto P(T), é 0,2% (terremotos, nesta cidade, são mais frequentes que arrombamentos). A probabilidade de o alarme tocar sem ter acontecido arrombamento ou terremoto é de 0,1% (o alarme pode disparar por alguma falha técnica, por exemplo). P[A|L] = 95% e P[A|T] = 29%. (o alarme é mais sensível a arrombamento que terremotos).

Figura 13 – Rede Bayesiana e probabilidades



Fonte: Marques e Dutra (2002).

2.3.4 Análise de Semântica Latente

Análise Semântica Latente (LSA, *Latent Semantic Analysis* no original) é a técnica utilizada para extrair e representar o significado semântico das palavras em um determinado contexto, obtidos através de cálculos estatísticos aplicados em um grande grupo de textos (LANDAUER, FOLTZ e LAHAM, 1998). LSA é uma variante do modelo espaço vetorial

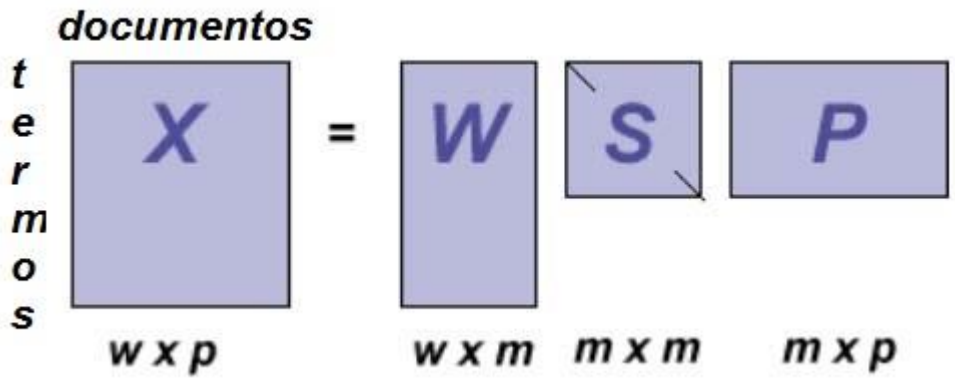
que converte uma amostra representativa de documentos em uma matriz de termo por documento em que cada célula indica a frequência com que cada termo (linhas) ocorre em cada documento (colunas). Deste modo, um documento torna-se um vetor de coluna e pode ser comparado com a consulta de um usuário representada como um vetor de mesma dimensão.

O padrão de indexação semântica do LSA é baseado na coocorrência de palavras em textos. A premissa é que palavras tendem a ocorrer em conjunto em um mesmo documento (frases, parágrafos, textos) são consideradas como amostras de similaridade semântica. Em prática, esta técnica é utilizada para a construção de um espaço semântico, onde não só palavras, mas também sentenças, parágrafos, textos ou qualquer conjunto de termos podem ser representados por vetores.

A LSA herda o modelo de espaço vetorial modelando relações de termo-documento usando uma aproximação reduzida para o espaço de coluna e linha calculado pela decomposição do valor singular do termo por matriz de documento. Pode-se dizer que a LSA consiste na construção de uma matriz X que informa a coocorrência de termos e documentos. Posteriormente, a matriz é algebricamente decomposta seguindo a Decomposição de Valores Singulares.

A Decomposição em Valores Singulares (SDV, *Singular Value Decomposition* no original) é uma forma de análise fatorial, ou mais propriamente, a generalização matemática da qual a análise fatorial é um caso especial (BERRY, DRMAC e JESSUP, 1999). O SVD constrói um espaço semântico abstrato de dimensão *n* em que cada termo original e cada documento original (e qualquer novo) são apresentados como vetores. No SVD uma matriz retangular de termo por documento X é decomposta no produto de três outras matrizes W, S e P', sendo {X} = {W} {S} {P}, conforme a Figura 14.

Figura 14 - Decomposição em Valores Singulares (SVD)



Fonte: elaborado pelo autor.

A matriz W é uma matriz ortonormal e suas linhas correspondem às linhas de X, mas tem *m* colunas correspondentes a novas variáveis especialmente variadas, de modo que não há correlação entre as duas colunas, isto é, cada uma é linearmente independente das outras. P é uma matriz ortonormal e tem colunas correspondentes às colunas originais, mas *m* linhas constituídas por vetores singulares derivados. A matriz S é uma matriz diagonal *m* por *m* com entradas não-zero (denominadas valores singulares) unicamente ao longo de uma diagonal central. O papel desses valores singulares é relacionar a escala dos fatores nas outras duas matrizes uma com a outra, de modo que quando os três componentes são multiplicados por matriz, a matriz original é reconstituída.

2.4 APLICAÇÃO DE ALGORITMOS PARA ANÁLISE TEXTUAL NA EDUCAÇÃO

No ambiente das atividades docentes, tarefas de leitura e interpretação de produções textuais contemplam o dia a dia de grande parcela dos professores. Em determinadas disciplinas as avaliações são realizadas através da resolução de problemas ou provas objetivas. Poucos apontamentos são empregados para avaliar a aprendizagem através de produções textuais. Além de instigar a leitura e a escrita por parte dos estudantes, observa-se que as produções textuais poderiam complementar a avaliação que um docente aplica sobre um estudante (ou uma turma), pois compõem-se de demonstrações da concepção do conhecimento (LOVATO, GIL, *et al.*, 2016).

As técnicas de aprendizado de máquina podem ser utilizadas com o intuito de organizar, categorizar ou descobrir informações que estão em bases de dados de forma rápida e automática. Tecnologias disponíveis possibilitam que o aprendizado seja realizado a partir de dados textuais (FELDMAN e SANGER, 2006). Essas tecnologias proporcionam a identificação de padrões e a subsequente classificação de dados de maneira automática e precisa. Com o propósito de comparações e análises, os dados obtidos podem ser sujeitos a técnicas de visualização quem ajudem na elaboração de *insights* (WEBBER, CINI e LIMA, 2013).

Associado a técnicas de aprendizado de máquina faz-se necessário prover recursos de visualização de dados. A área de visualização de dados leva em conta a elaboração de representações visuais de dados abstratos de maneira a tornar fácil o seu entendimento e a descoberta de novas informações (FREITAS, CHUBACHI, *et al.*, 2001). Após coletados os dados, é realizada uma transformação e organização desses dados em tabelas relacionais. Estas tabelas são convertidas em estruturas visuais que representam os dados utilizando representações gráficas, símbolos e padrões. Operações são realizadas nestas estruturas

visuais com o objetivo de exibir novas informações sobre estes conjuntos de dados por meio de manipulações geométricas, indicações de subconjuntos (regiões) de interesse ou mudança de ponto de observação (CARD, MACKINLAY e SHNEIDERMAN, 1999). Técnicas são empregadas no entendimento e na busca por informações resultantes do que está sendo visualizado. As técnicas se classificam baseadas em três critérios: os dados a serem visualizados, a técnica de visualização utilizada e a técnica de interação aplicada para manipular as representações visuais (KEIM, 2002).

De forma pertinente Judelman (2004) estabelece três categorias de visualização: complexidade, contexto e dinâmica. A categoria de complexidade contempla a visualização de árvores e hierarquias, mostrando topologias ou caminhos. A modalidade de contextos mostra relacionamentos semânticos, exibindo os objetos espacialmente organizados conforme a similaridade ou importância. A visualização de perspectiva dinâmica engloba diagramas de processos mostrando o contexto espaço temporal e comumente um eixo representando o tempo. Para a visualização e comparação entre textos, a categoria que melhor se encaixa é a de contexto, baseado nas propriedades semânticas a serem analisadas.

2.5 CONSIDERAÇÕES FINAIS

A aprendizagem automática aplicada a dados textuais permite extrair padrões, sendo possível sua utilização em diversos domínios de conhecimento. Foram apresentadas algumas ferramentas de mineração de texto já existentes com o objetivo de demonstrar estudos anteriores. Pode-se verificar que existem diferentes técnicas para analisar textos de acordo com o objetivo desejado e informação necessária. Também foi notada que avaliação de produções textuais é uma atividade constante na área da educação e que técnicas de análise de textos para visualização de informações são de grande auxílio para os docentes.

Com o embasamento teórico obtido neste capítulo, é possível avaliar a ferramenta de análise de produções textuais desenvolvida por Lovato (2015) e Gil (2016) para a avaliação dos textos dos estudantes, e apontar melhorias que nela podem ser propostas, que é objetivo deste trabalho. O detalhamento da ferramenta de análise de produções textuais é apresentado no capítulo a seguir.

3 FERRAMENTA DE ANÁLISE DE PRODUÇÕES TEXTUAIS

A ferramenta de análise de produções textuais desenvolvida por Lovato (2015) e Gil (2016) opera em duas etapas. A primeira etapa é a de treinamento, onde o software realiza o aprendizado a partir de textos fornecidos pelo professor. A segunda etapa consiste na análise dos textos desenvolvidos pelos estudantes. Os resultados processados são disponibilizados por meio de diversas formas de visualizações gráficas. O sistema calcula o limiar do documento, que indica a similaridade entre o corpus de texto fornecido pelo professor com as atividades dos estudantes. O software utiliza o algoritmo de análise semântica latente para realizar a análise dos textos. A análise semântica latente é uma abordagem utilizada para extrair relações entre documentos e os termos neles contidos.

Dentre as várias etapas que compõem o algoritmo implementado está a indexação do documento de treinamento, onde *stopwords* são filtrados e os termos restantes inseridos em uma *HashTable*. Quanto mais precisa essa filtragem, melhor o resultado final do processo. Se por exemplo, forem identificados novos *stopwords* a serem inseridos, ou *stopwords* indevidamente implementados forem removidos, obtém-se um resultado mais preciso. As melhorias detectadas poderão afetar igualmente as interfaces do software (interface do usuário e visualização dos resultados). Neste escopo é que se insere este trabalho, trazendo resultados com maior exatidão para o usuário.

3.1 ETAPAS DE FUNCIONAMENTO DA FERRAMENTA

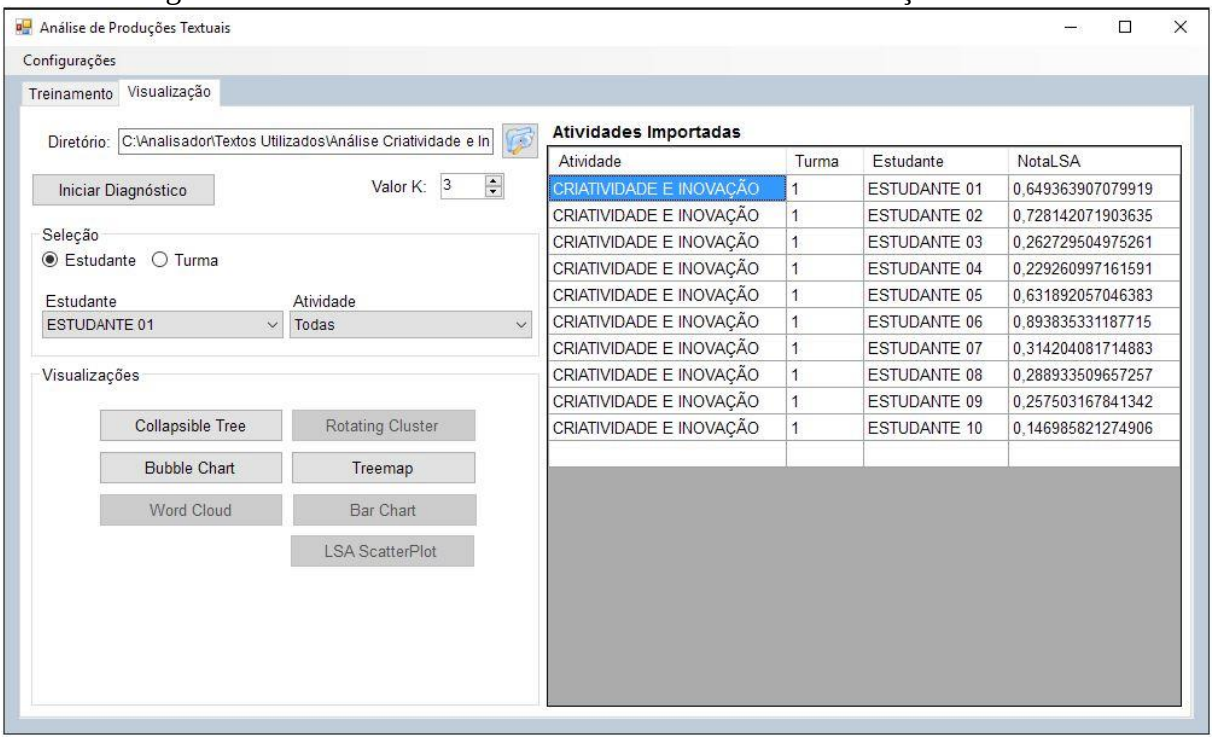
A ferramenta de Análise de Produções Textuais funciona em duas etapas. Primeiro ocorre a etapa de treinamento, onde a aplicação realiza o aprendizado partindo de textos e materiais didáticos do docente. Em seguida, ocorre a segunda etapa, que contempla a análise dos textos escritos pelos estudantes. Os resultados são transformados e disponibilizados através de diversas visualizações gráficas. A interface de usuário possui duas abas. A primeira é a aba de treinamento, a qual permite ao professor informar seu corpus de texto. Na outra aba o professor é capaz de acessar e manipular recursos de visualização. A etapa de treinamento é indispensável para que as análises textuais sejam bem-sucedidas. O docente deve escolher textos e materiais didáticos para serem utilizados, que podem ser em idioma português ou inglês (mas apenas um por vez).

O objetivo da segunda etapa é analisar os textos redigidos pelos estudantes, gerando visualizações que permitem a análise por parte do docente. Primeiramente devem ser informados o diretório onde as produções textuais dos estudantes estão armazenadas e um

valor K para estabelecer a precisão dos resultados do algoritmo LSA. A definição de um valor K é determinante para receber resultados viáveis. Um valor K muito baixo, como 2 por exemplo, melhora o desempenho da execução mas resulta em um cálculo impreciso em razão da falta de informação. Já um valor K exagerado acrescenta muito ruído ao cálculo, também trazendo prejuízo à precisão do resultado final.

Clicando no botão “Iniciar Diagnóstico” (Figura 15), o programa importará todos os arquivos do diretório de origem, mostrando-os como “Atividades Importadas”. O software informa quais arquivos importou, descrevendo a atividade, a turma e o estudante. Na tela da caixa de “Seleção”, ficam acessíveis as visualizações: uma atividade de um estudante (1-1), uma atividade de uma turma (1-n), várias atividades dos estudantes (n-n) e várias atividades de um estudante (n-1).

Figura 15 – Tela de análise da ferramenta Análise de Produções Textuais



Fonte: Lovato, Gil, et al. (2016).

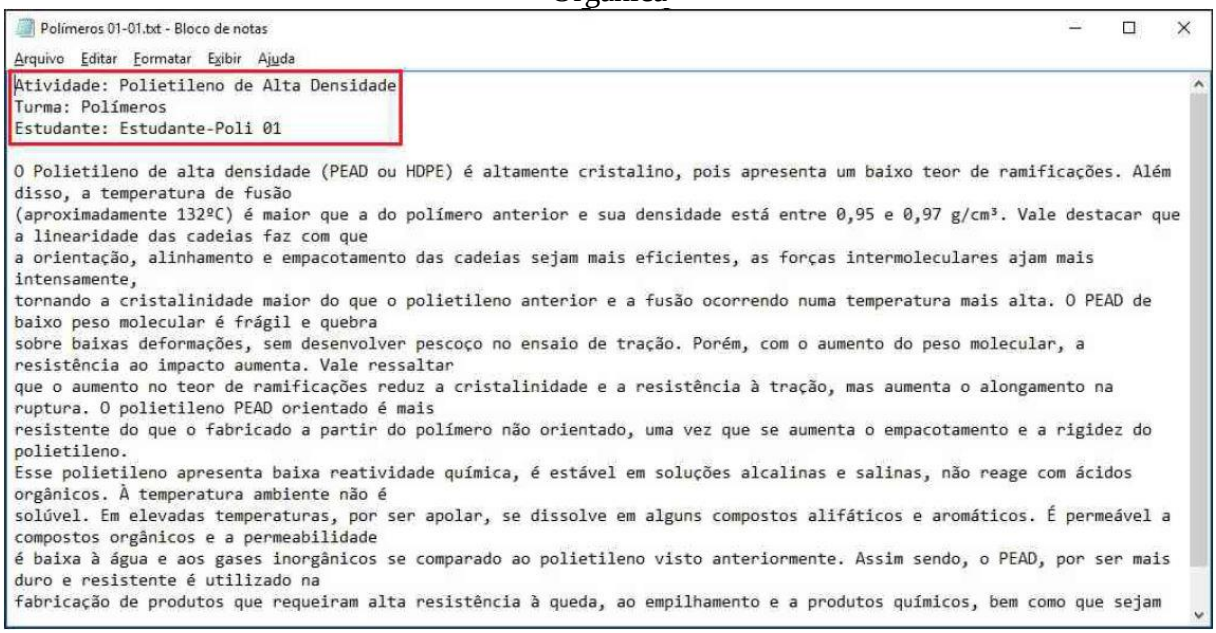
A ferramenta calcula o limiar do documento, que representa sua semelhança em relação ao conteúdo verificado no corpus de texto do professor. O algoritmo LSA foi empregado pela aplicação para efetuar a análise dos textos. A coluna “Nota LSA” (Figura 15) apresenta de forma resumida o resultado da análise. O algoritmo foi implementado em 11 etapas:

1. inicialização das *HashTables* onde são indexados os termos, documentos, e a relação termo/documento. É efetuada a leitura do documento de treinamento, contando a frequência de cada termo;
2. indexação do documento de treinamento: é realizada a filtragem dos *StopWords*, os termos restantes são acrescentados na *HashTable*, é gerada a relação termo/documento perante a frequência de cada termo e o documento de treinamento é inserido;
3. semelhante à segunda etapa, todavia indexa-se os textos a serem analisados;
4. elaboração da matriz termo/documento, em que cada linha corresponde a um termo, a cada coluna corresponde a um documento. Cada célula corresponde a frequência termo/documento (peso local) gerada anteriormente;
5. cálculo do peso global de cada termo, ou seja, quantos documentos dentro da lista apresentam a palavra. A função aplicada foi a frequência inversa de documentos;
6. cálculo dos fatores de normalização dos termos de cada documento, empregando o peso global do termo e o peso local do termo no respectivo documento de todos os termos;
7. formação da matriz termo/documento ponderada, na qual o valor de cada célula considera o peso local de cada termo, o peso global de cada termo e o fator de normalização de cada documento;
8. cálculo do SVD, possibilitando a reconstrução da matriz termo/documento ponderada original;
9. disposição das matrizes reduzidas, aplicadas no cálculo de similaridade dos documentos. As matrizes reduzidas são calculadas com base no valor K (informado na interface de usuário);
10. gravação em disco de diversas variáveis necessárias para o cálculo de similaridade de documentos, aliviando o uso de memória durante a execução, sendo que o conjunto de matrizes pode crescer rapidamente;
11. redução das matrizes, por meio de operações de transposição e multiplicação de matrizes.

A produção textual de um estudante é apresentada na Figura 16, com detalhe no cabeçalho do arquivo. O cabeçalho deve apresentar informações sobre a turma, nome do

estudante e indicador de atividade. Estes dados são necessários para que o programa registre as turmas, estudantes e atividades para gerar as visualizações corretamente.

Figura 16 – Exemplo de produção textual de um estudante de uma turma de Química Orgânica

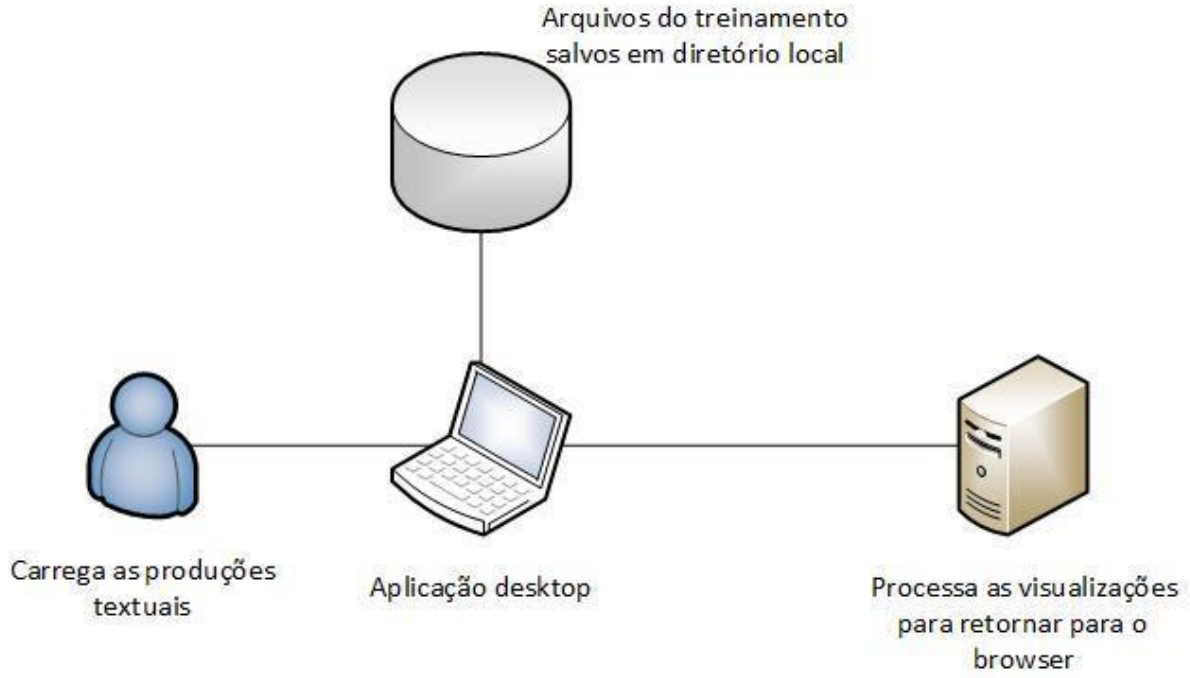


Fonte: Lovato, Gil, et al. (2016).

3.2 TECNOLOGIAS UTILIZADAS NO DESENVOLVIMENTO

O software Análise de Produções Textuais foi programado em linguagem C#. Em complemento, foram integradas visualizações implementadas em *JavaScript* com a biblioteca *D3 Data-Driven Documents* (BOSTOCK, 2013). A comunicação entre o programa e as visualizações *web* é feita através de arquivos em formato *JSON*. Estes arquivos são armazenados no diretório de publicação do servidor *web* e são lidos pelas aplicações desenvolvidas para a camada *web*. A Figura 17 mostra a representação dessa arquitetura.

Figura 17 - Arquitetura do sistema Análise de Produções Textuais



Fonte: Lovato (2015).

A ferramenta Análise de Produções Textuais proporciona sete visualizações diferentes, implementadas usando a tecnologia *D3 Data-Driven Documents* (BOSTOCK, 2013). As classes que estabelecem as visualizações são encarregadas de fazer a chamada da visualização que será aberta no navegador padrão do computador. As visualizações que eram oferecidas até o momento são: Bar Chart (Gráfico de Barras), BubbleChart (Gráfico de Bolhas), CollapsibleTree (Árvore Flexível), RotatingCluster (Agrupamento Rotativo), Treemap (Mapa Árvore) Word Cloud (Nuvem de Palavras) e LSA ScatterPlot. Os nomes das visualizações foram mantidos em inglês na interface de usuário por se tratarem de nomes padronizados.

3.3 ANÁLISE DA FERRAMENTA

A ferramenta Análise de Produções Textuais cumpre seu papel de maneira satisfatória. Dentre os pontos fortes, podemos citar a boa usabilidade, a rapidez na qual a ferramenta executa suas operações (principalmente as etapas de treinamento e análise) e a clareza dos resultados apresentados. Porém também foram encontrados alguns pontos fracos, tais como o número relativamente limitado de visualizações de resultados e divergências na filtragem de *stopwords*. Outro ponto fraco que podemos citar é que pelo fato da ferramenta ser desenvolvida em C# ela só pode ser instalada em versões do sistema operacional

Windows, limitando a possibilidade de uso do usuário pela disponibilidade de um sistema operacional licenciado.

Testes que foram realizados indicaram a presença de termos que não contribuem para a interpretação dos textos dos estudantes (e.g. como, porém) que remetem a necessidade de se aprimorar a lista de palavras a serem removidas do texto (lista de *stopwords*). Por esta razão pesquisou-se por listas de *stopwords* que complementariam a existente.

A necessidade de visualizar dados e informações é permanente. Mesmo que alguns recursos estejam disponíveis, deve-se buscar elementos que melhorem a percepção do professor sobre as produções textuais dos estudantes. Tendo verificado que as visualizações presentes favorecem a compreensão do professor, mas ainda não são completas, deve-se explorar novos recursos de visualização. A própria biblioteca *D3.js* (BOSTOCK, 2013) oferece formas de visualização que complementam as existentes. Por esta razão, optou-se por estudá-las a fim de identificar novas formas relevantes de apresentar os dados visualmente.

Tendo em vista estas observações, o próximo capítulo aborda os desenvolvimentos realizados.

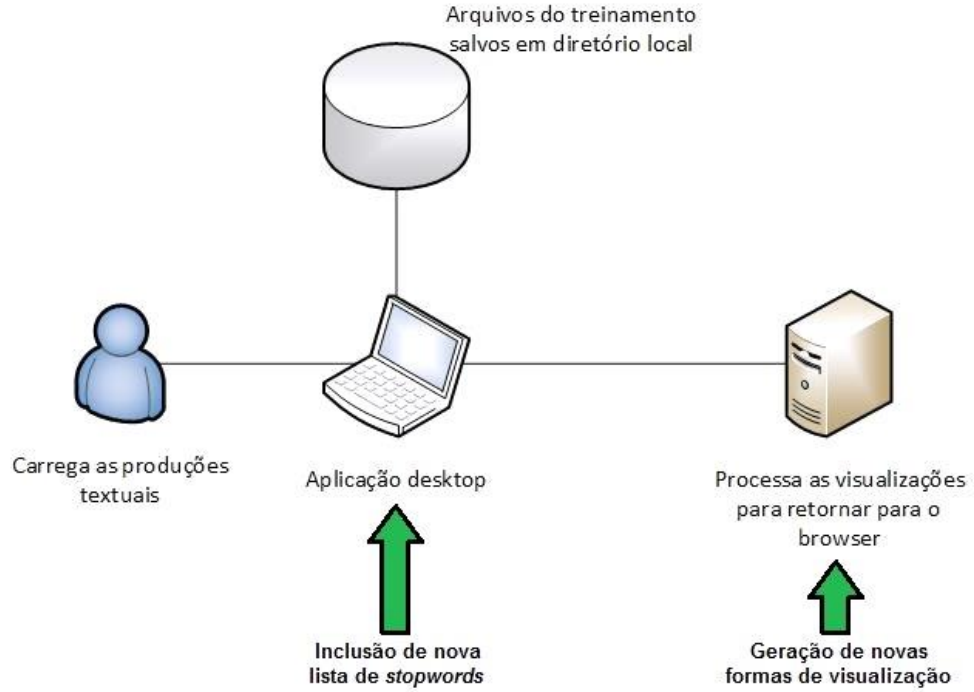
4 IMPLEMENTAÇÃO, TESTES E ANÁLISE DOS RESULTADOS

O capítulo a seguir tem por objetivo apresentar as melhorias desenvolvidas sobre a ferramenta Análise de Produções Textuais. Após isto, serão demonstrados testes envolvendo as evoluções realizadas e em seguida uma análise sobre os resultados obtidos.

4.1 IMPLEMENTAÇÃO

A implementação foi realizada mantendo a estrutura existente e incluindo os novos recursos identificados. Para manter o modelo original de desenvolvimento, decidiu-se por continuar usando a linguagem C# e a biblioteca *JavaScript D3.js* para gerar as visualizações gráficas. A comunicação entre a aplicação e as visualizações *web* também continuou dando-se através da utilização de arquivos no formato *JSON*. A metodologia de uso do software também continua sendo a mesma: necessita-se de texto(s) de referência para realizar o treinamento (corpus do professor) e os textos elaborados pelos estudantes para análise. A Figura 18 apresenta as camadas da ferramenta de Análise de Produções Textuais que foram desenvolvidas, com suas respectivas evoluções realizadas.

Figura 18 – Camadas da ferramenta e evoluções realizadas



Fonte: adaptado de Lovato (2015).

A lista de *stopwords* que continha os termos removidos do texto possuía 256 palavras (Apêndice A). Foi proposto e implementado o uso de uma nova lista de *stopwords* que contém 560 termos, sendo esta lista mais completa que a inicialmente empregada para

remoção de palavras desnecessárias para a análise. Esta nova lista utilizada pode ser visualizada no Apêndice B[1].

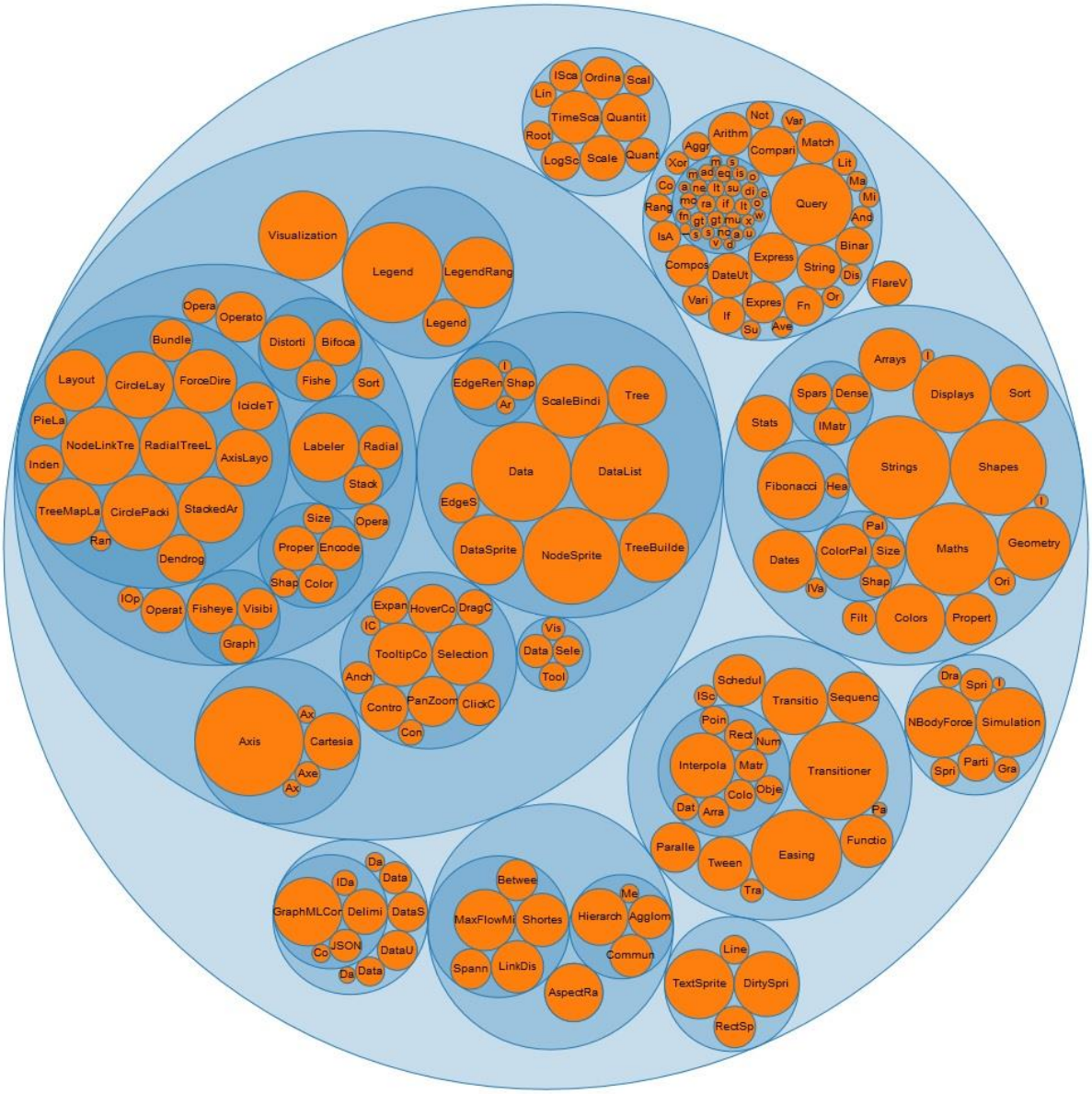
4.2 VISUALIZAÇÕES SELECIONADAS

Para a escolha das visualizações a serem adicionadas na ferramenta, procurou-se selecionar modelos que sejam melhor aplicáveis a casos de análise textual. Ou seja, foi dada preferência a visualizações que apresentem resultados com informações em forma de texto (termos e palavras).

A primeira nova visualização incorporada ao software é a *Static Circle Packing* (Figura 19). Embora não seja tão eficiente em termos de espaço de distribuição de informação em tela quanto a *Treemap*, esta visualização representa bem a hierarquia entre turma, estudantes e atividades. Na implementação deste exemplo (BOSTOCK, 2013), considera-se como cada grupo de círculos menores sendo as palavras analisadas. Cada círculo que os engloba externamente será uma atividade. Estes círculos são agrupados por outro círculo externo que representa um estudante, e o conjunto final (maior círculo externo) representa a turma de estudantes.

[1] Nova lista de *stopwords* utilizada disponível em https://github.com/stopwords-iso/stopwords-pt.

Figura 19 – *Static Circle Packing*



Fonte: Bostock (2013).

Também foi implementada uma variação da *Static Circle Packing* que é a *Zoomable Circle Packing*. Nesta versão é possível alterar a exibição dos dados entre o nome da atividade e dos estudantes, além de ser possível aproximar a visualização para a atividade de um determinado estudante.

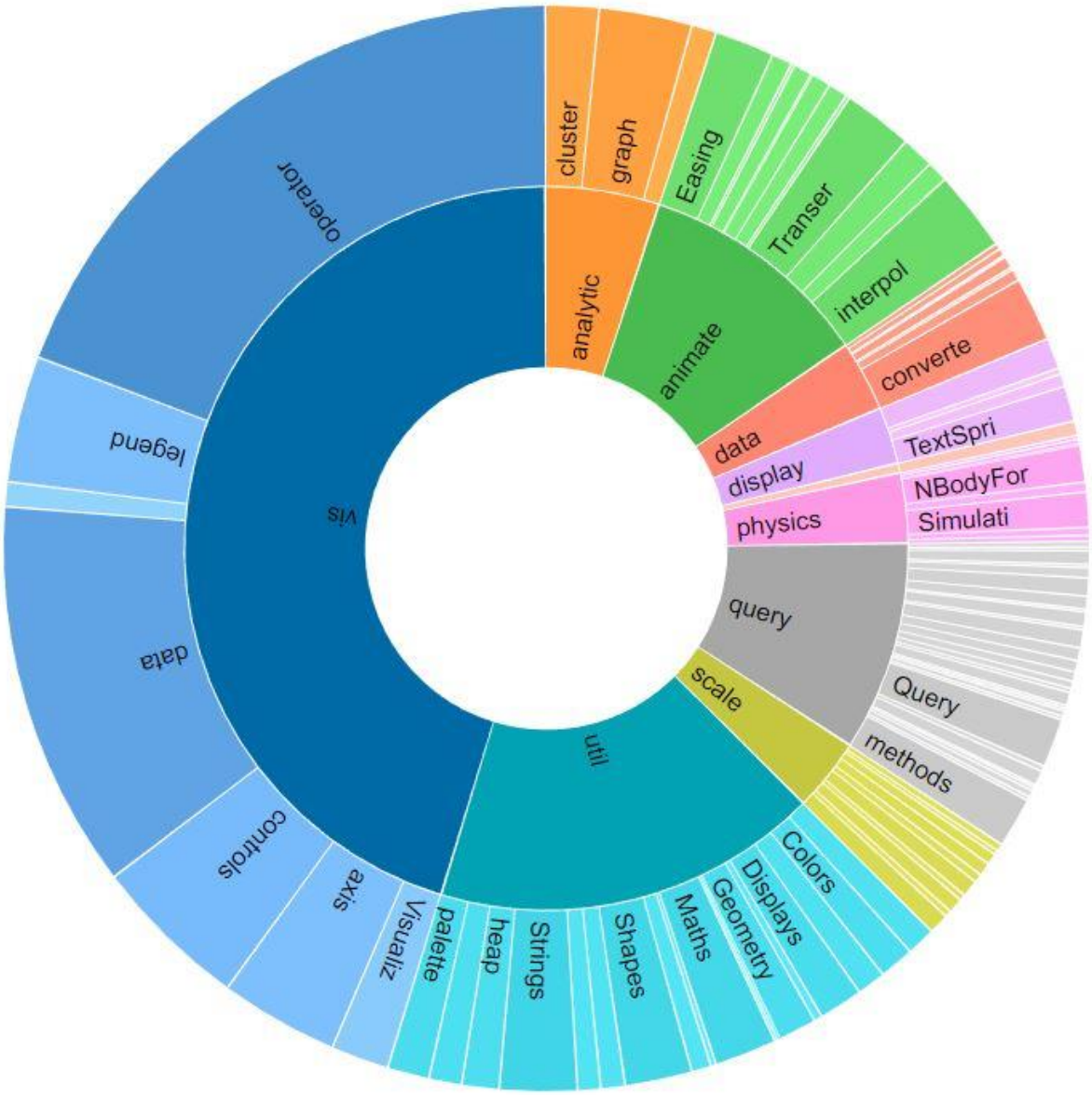
Outra visualização implementada foi a *Sunburst Partition* (Figura 20). Esta visualização é semelhante à *Treemap*, porém ela apresenta os dados de forma radial (BOSTOCK, 2013).

Figura 20 - *Sunburst Partition*

Fonte: Bostock (2013).

A última visualização incorporada ao software foi a *Bilevel Partition* (Figura 21). Esta visualização é semelhante à *Sunburst Partition*. No entanto, ela apresenta as palavras analisadas e possibilita expandir dinamicamente os subníveis da hierarquia entre turma, estudantes e atividades (BOSTOCK, 2013).

Figura 21 - *Bilevel Partition*



Fonte: Bostock (2013).

4.3 MATERIAIS E MÉTODO

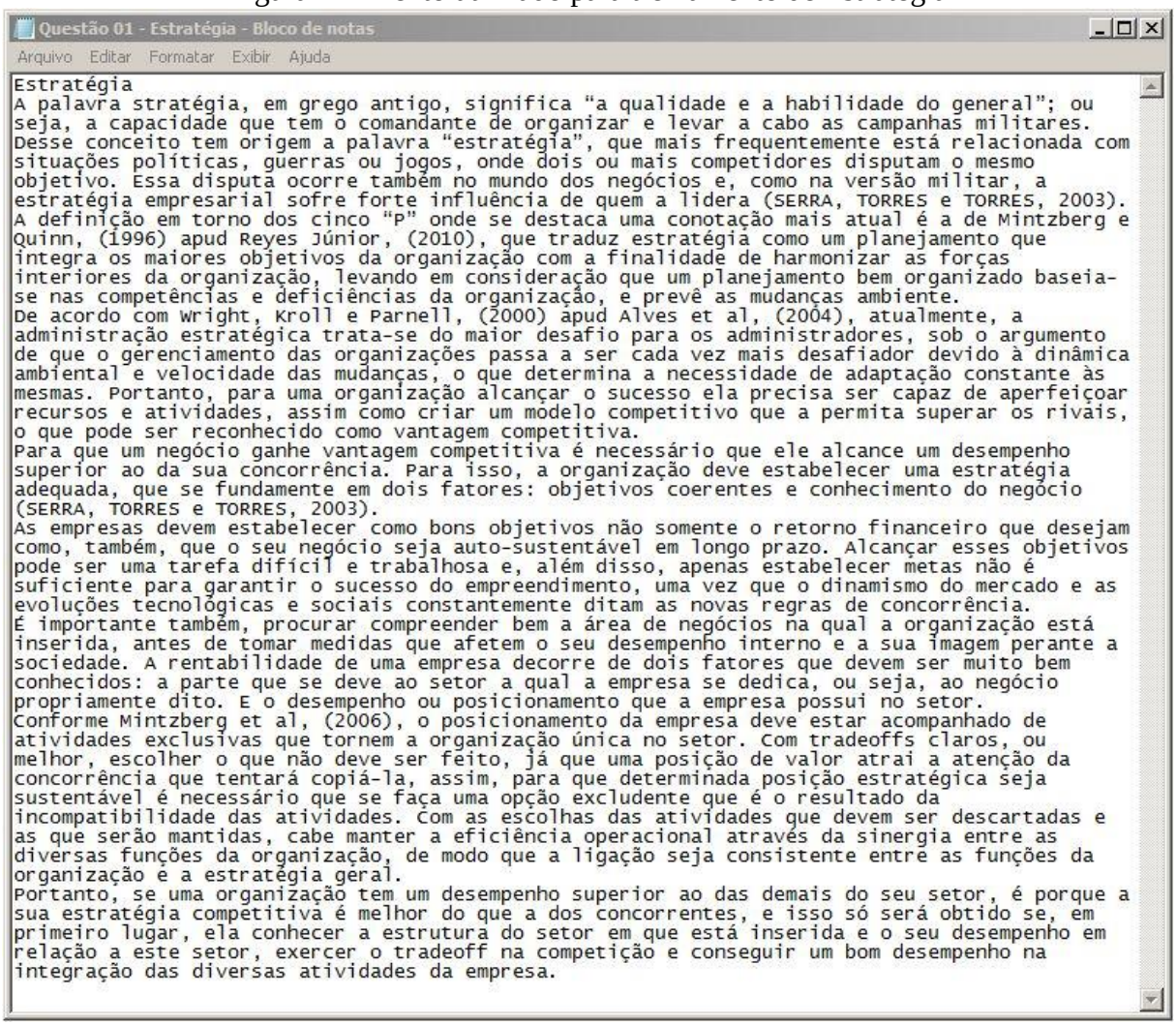
Este capítulo visa apresentar os materiais e método utilizados no desenvolvimento do trabalho. Dentre estes, constam os arquivos de texto elaborados para treinamento e análise da ferramenta APT e os testes realizados com os textos selecionados.

4.3.1 Materiais

Para elaboração do cenário de teste foram utilizadas seções do estudo de caso "Análise estratégica: um estudo na lanchonete e pizzaria 'Bom Te Ver'" (OESTREICHER, SANTO e JÚNIOR, 2010). O texto em questão trata do conhecimento do ambiente interno de

uma empresa e do setor ao qual ela pertence, mostrando o quanto essas informações são importantes para a sobrevivência e crescimento de negócios voltados a pequenos restaurantes e lanchonetes, diante de um mercado de grande concorrência. As seções utilizadas abordam conceitos do referencial teórico do estudo de caso utilizado, tais como estratégia, planejamento e análise SWOT. A Figura 22 apresenta um arquivo elaborado para utilização na etapa de treinamento da ferramenta, neste caso abordando o conceito de estratégia.

Figura 22 - Texto utilizado para treinamento de Estratégia


Questão 01 - Estratégia - Bloco de notas

Arquivo Editar Formatar Exibir Ajuda

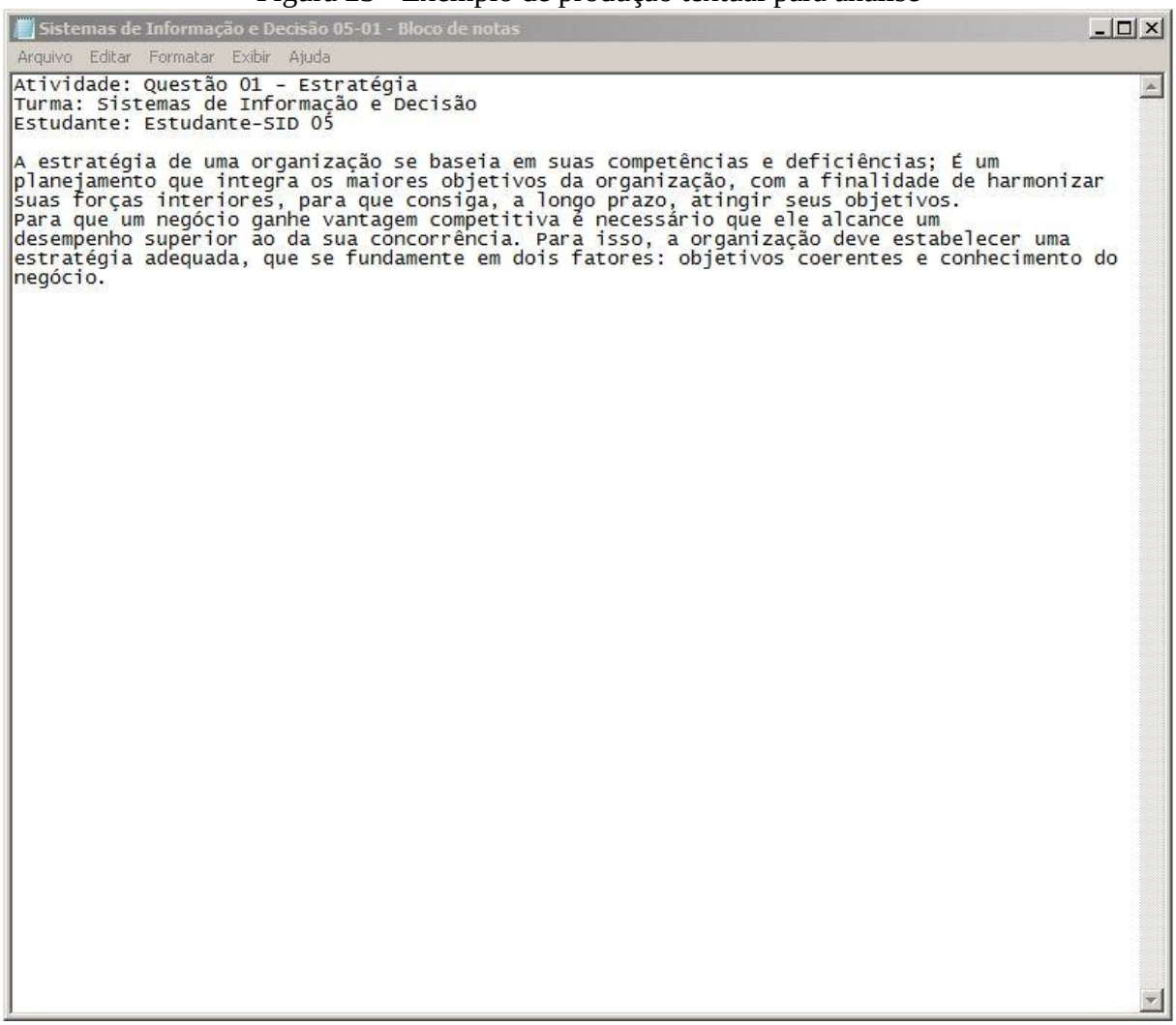
Estratégia
A palavra stratégia, em grego antigo, significa "a qualidade e a habilidade do general"; ou seja, a capacidade que tem o comandante de organizar e levar a cabo as campanhas militares. Desse conceito tem origem a palavra "estratégia", que mais frequentemente está relacionada com situações políticas, guerras ou jogos, onde dois ou mais competidores disputam o mesmo objetivo. Essa disputa ocorre também no mundo dos negócios e, como na versão militar, a estratégia empresarial sofre forte influência de quem a lidera (SERRA, TORRES e TORRES, 2003).
A definição em torno dos cinco "P" onde se destaca uma conotação mais atual é a de Mintzberg e Quinn, (1996) apud Reyes Júnior, (2010), que traduz estratégia como um planejamento que integra os maiores objetivos da organização com a finalidade de harmonizar as forças interiores da organização, levando em consideração que um planejamento bem organizado baseia-se nas competências e deficiências da organização, e prevê as mudanças ambiente.
De acordo com Wright, Kroll e Parnell, (2000) apud Alves et al, (2004), atualmente, a administração estratégica trata-se do maior desafio para os administradores, sob o argumento de que o gerenciamento das organizações passa a ser cada vez mais desafiador devido à dinâmica ambiental e velocidade das mudanças, o que determina a necessidade de adaptação constante às mesmas. Portanto, para uma organização alcançar o sucesso ela precisa ser capaz de aperfeiçoar recursos e atividades, assim como criar um modelo competitivo que a permita superar os rivais, o que pode ser reconhecido como vantagem competitiva.
Para que um negócio ganhe vantagem competitiva é necessário que ele alcance um desempenho superior ao da sua concorrência. Para isso, a organização deve estabelecer uma estratégia adequada, que se fundamente em dois fatores: objetivos coerentes e conhecimento do negócio (SERRA, TORRES e TORRES, 2003).
As empresas devem estabelecer como bons objetivos não somente o retorno financeiro que desejam como, também, que o seu negócio seja auto-sustentável em longo prazo. Alcançar esses objetivos pode ser uma tarefa difícil e trabalhosa e, além disso, apenas estabelecer metas não é suficiente para garantir o sucesso do empreendimento, uma vez que o dinamismo do mercado e as evoluções tecnológicas e sociais constantemente ditam as novas regras de concorrência.
É importante também, procurar compreender bem a área de negócios na qual a organização está inserida, antes de tomar medidas que afetem o seu desempenho interno e a sua imagem perante a sociedade. A rentabilidade de uma empresa decorre de dois fatores que devem ser muito bem conhecidos: a parte que se deve ao setor a qual a empresa se dedica, ou seja, ao negócio propriamente dito. E o desempenho ou posicionamento que a empresa possui no setor.
Conforme Mintzberg et al, (2006), o posicionamento da empresa deve estar acompanhado de atividades exclusivas que tornem a organização única no setor. Com tradeoffs claros, ou melhor, escolher o que não deve ser feito, já que uma posição de valor atrai a atenção da concorrência que tentará copiá-la, assim, para que determinada posição estratégica seja sustentável é necessário que se faça uma opção excludente que é o resultado da incompatibilidade das atividades. Com as escolhas das atividades que devem ser descartadas e as que serão mantidas, cabe manter a eficiência operacional através da sinergia entre as diversas funções da organização, de modo que a ligação seja consistente entre as funções da organização e a estratégia geral.
Portanto, se uma organização tem um desempenho superior ao das demais do seu setor, é porque a sua estratégia competitiva é melhor do que a dos concorrentes, e isso só será obtido se, em primeiro lugar, ela conhecer a estrutura do setor em que está inserida e o seu desempenho em relação a este setor, exercer o tradeoff na competição e conseguir um bom desempenho na integração das diversas atividades da empresa.

Fonte: elaborado pelo autor.

A partir da leitura e interpretação do estudo de caso mencionado, 14 (catorze) estudantes de uma turma da disciplina de Sistemas de Informação e Decisão responderam 14 (catorze) questões sobre os assuntos abordados no texto. Esta lista de questões pode ser visualizada no Apêndice C. Desta lista foram selecionadas 3 (três) questões para realizar a análise pela ferramenta. Estas questões (14 estudantes x 3 questões = 42 textos) tratavam

respectivamente dos mesmos tópicos das seções do texto utilizadas para treinamento (estratégia, planejamento e análise SWOT). A Figura 23 apresenta um exemplo de produção textual desenvolvida por um estudante e utilizada para ser submetida à análise da ferramenta.

Figura 23 - Exemplo de produção textual para análise



Fonte: elaborado pelo autor.

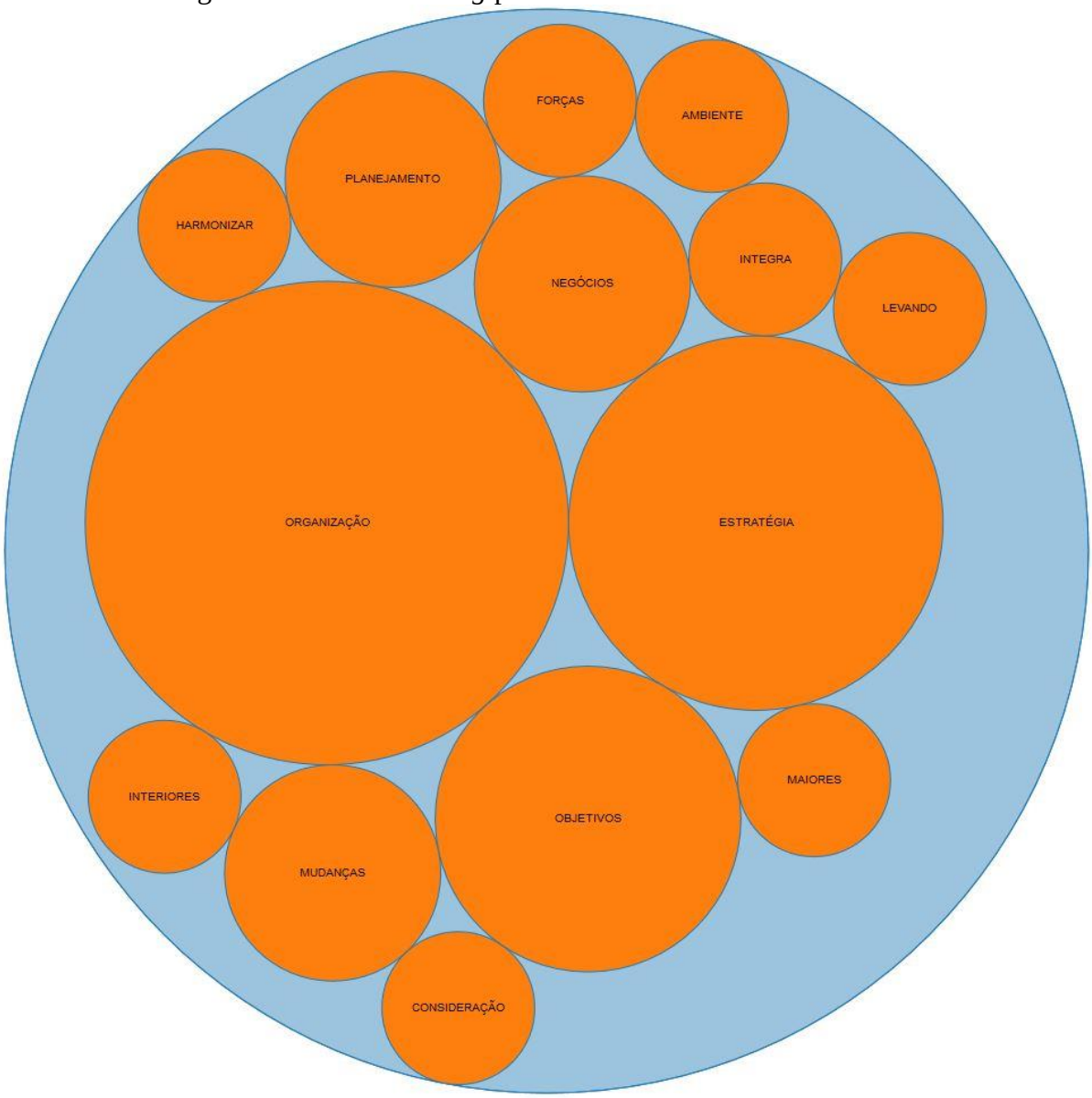
4.3.2 Método

Esta subseção tem por objetivo apresentar os testes realizados sobre as melhorias desenvolvidas. Para isto, será apresentado um comparativo com a versão antiga da ferramenta APT em relação à filtragem de *stopwords* e exemplos gerados com as novas visualizações implementadas utilizando os materiais coletados.

4.3.2.1 Testes para um estudante

A Figura 24 apresenta o teste da visualização *Static Circle Packing* analisando o texto de uma atividade de um estudante. Esta visualização tem por objetivo mostrar as palavras relevantes usadas pelo estudante no seu texto. Além de apresentar as palavras, é possível saber a relevância de cada uma delas. Proporcionalmente, as esferas maiores são os termos com o maior número de ocorrências no texto original. Também é possível saber o número exato de ocorrências posicionando o cursor do mouse sobre a palavra desejada.

Figura 24 - *Circle Packing* para um estudante em uma atividade

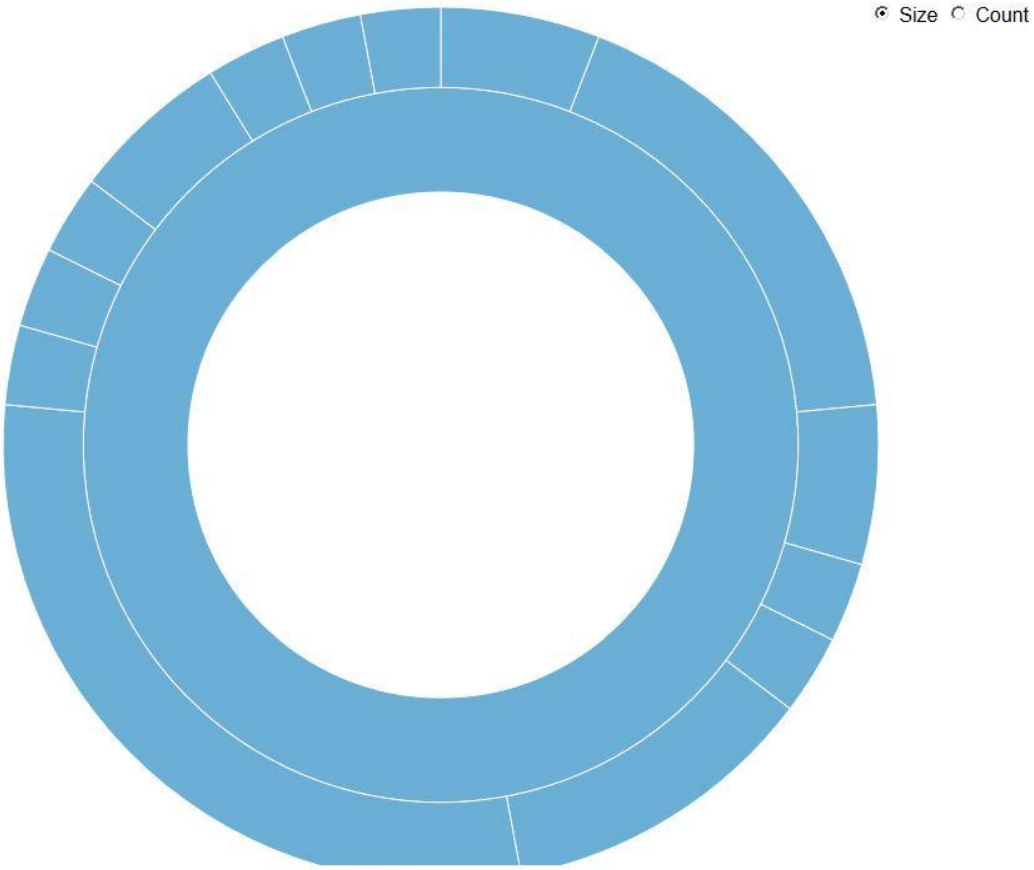


Fonte: elaborado pelo autor.

Na Figura 25 é demonstrada a visualização *Sunburst Partition*. O objetivo desta visualização é apresentar uma proporção da contagem das palavras relevantes encontradas na

atividade do estudante. É possível selecionar na apresentação da visualização no navegador se deseja visualizar o gráfico apenas com a contagem de palavras relevantes encontradas ou também com a proporção do número de ocorrências de cada palavra.

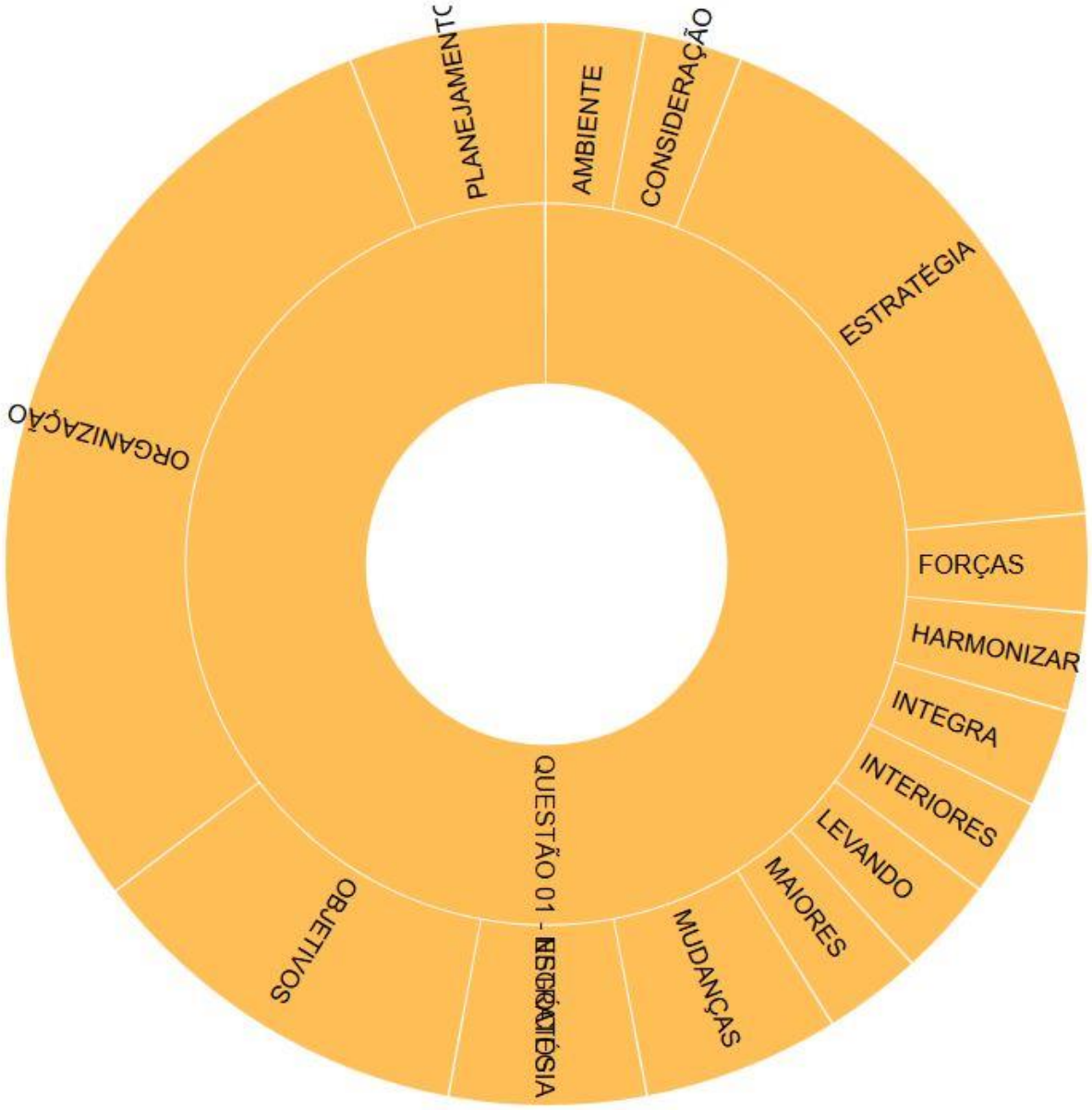
Figura 25 - *Sunburst Partition* para um estudante em uma atividade



Fonte: elaborado pelo autor.

A Figura 26 apresenta a visualização Bilevel Partition. Esta visualização tem a função de demonstrar um comparativo entre as palavras. Quanto maior a fatia da palavra, mais relevância ela tem na produção textual. Em contrapartida, quanto menor a fatia da palavra, menos relevância ela tem dentro da produção textual. Algumas palavras podem aparecer cortadas ou sobrepostas, mas basta posicionar o cursor do mouse sobre a área desejada para visualizar uma descrição completa da palavra e seu número de ocorrências.

Figura 26 - *Bilevel Partition* para um estudante em uma atividade

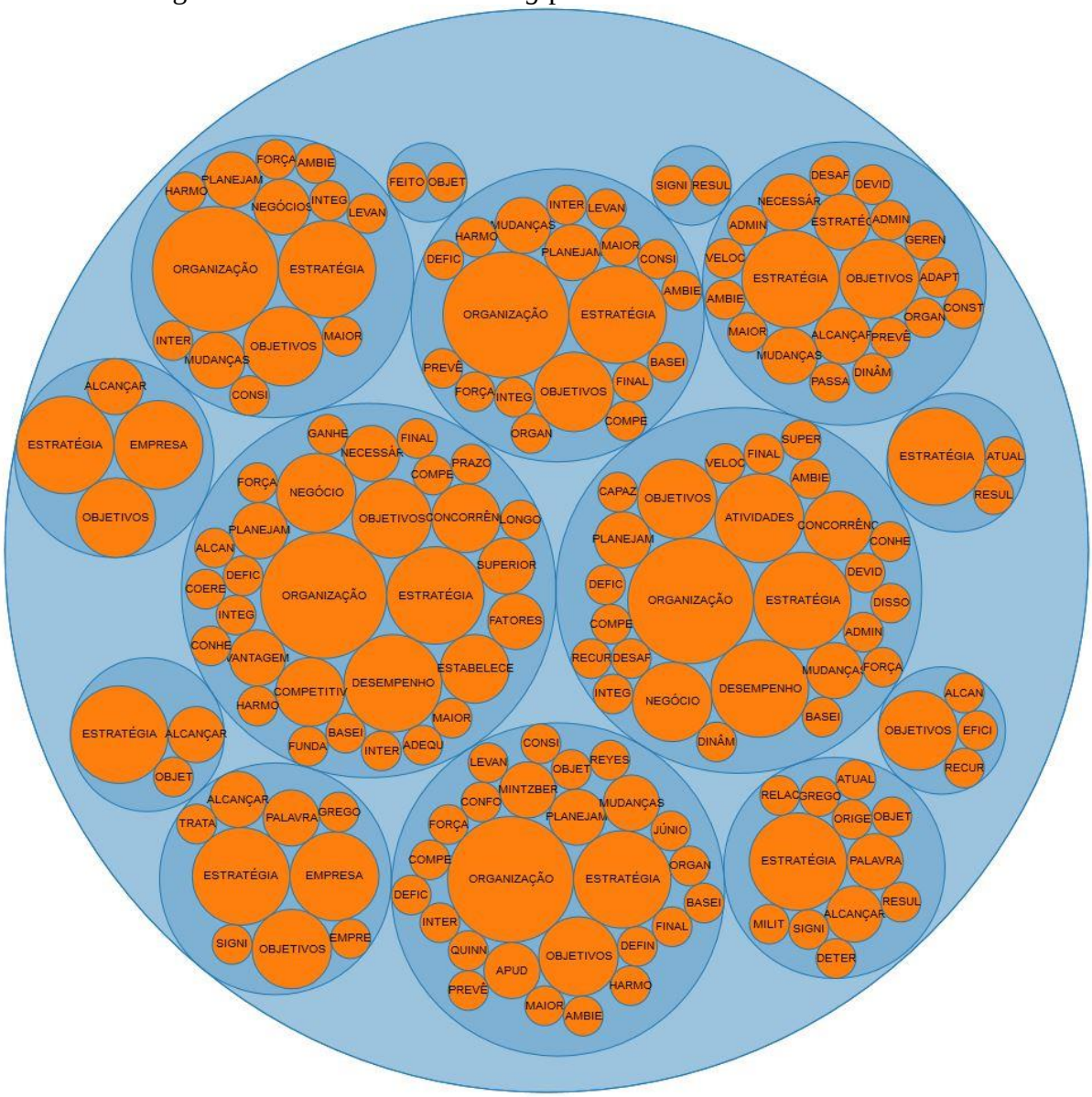


Fonte: elaborado pelo autor.

4.3.2.2 Testes para uma turma

Quando analisamos a visualização *Static Circle Packing* em uma turma de estudantes para uma atividade, consideramos como a turma o círculo maior externo que engloba todo o conjunto. Cada subconjunto dentro deste representa uma atividade de um estudante e as esferas menores são as palavras relevantes da atividade analisada daquele respectivo estudante. Algumas palavras apresentam-se abreviadas devido ao tamanho reduzido das esferas (termos com menos ocorrências), mas é possível visualizar a palavra completa posicionando o ponteiro do mouse sobre a palavra desejada (Figura 27).

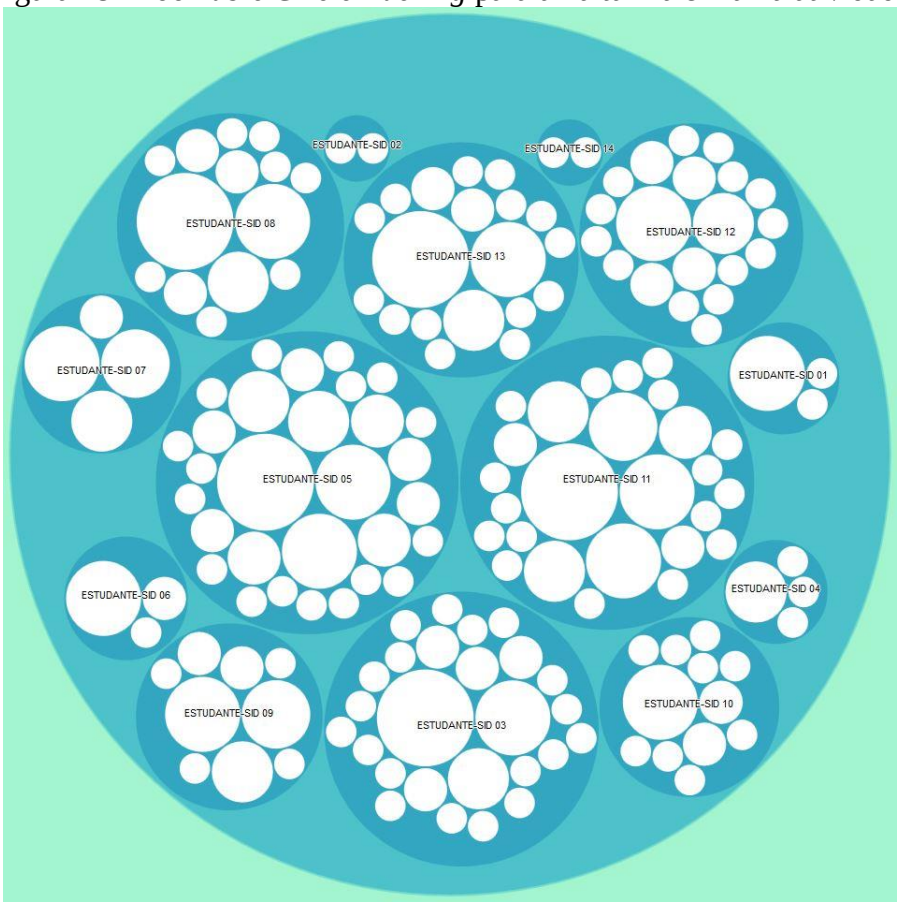
Figura 27 – *Static Circle Packing* para uma turma em uma atividade



Fonte: elaborado pelo autor.

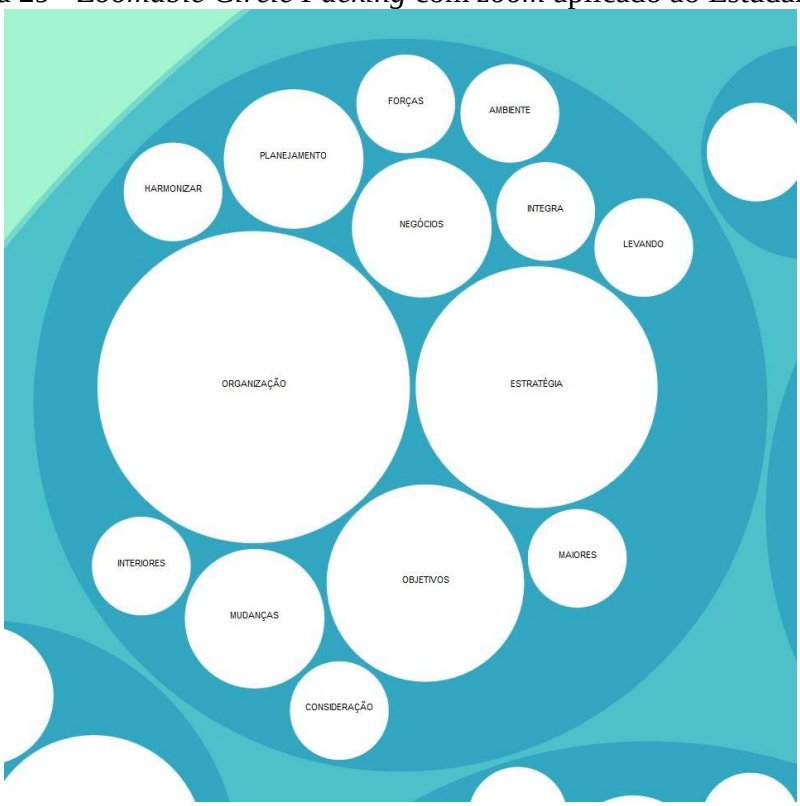
Na visualização *Zoomable Circle Packing* (Figura 28) há a opção de escolher as informações apresentadas no navegador. Ao clicar com o mouse na área do círculo mais externo, é possível alternar entre a apresentação do nome da atividade ou o nome dos estudantes. Realizando o clique de seleção em um determinado estudante, a visualização aproxima para este, apresentando as palavras relevantes analisadas para este estudante nesta atividade. Neste teste aproximamos a visualização para o "Estudante 08" (Figura 29).

Figura 28 - *Zoomable Circle Packing* para uma turma em uma atividade



Fonte: elaborado pelo autor.

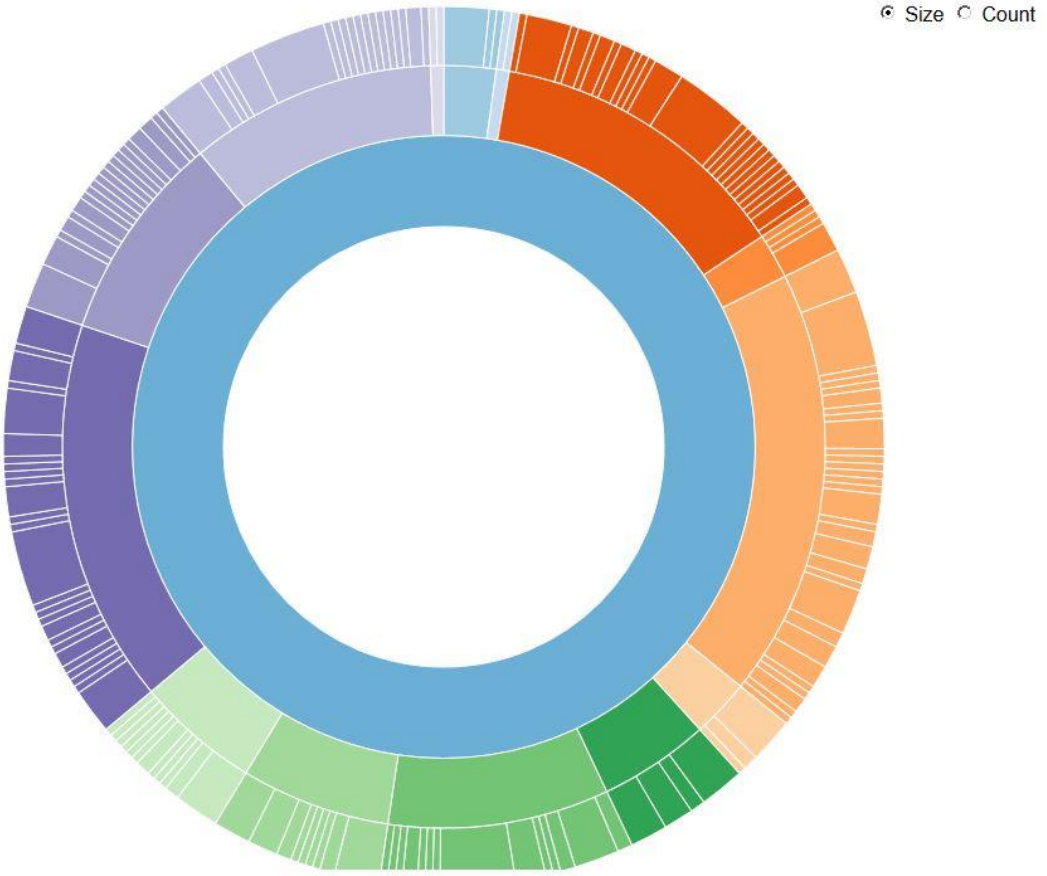
Figura 29 - *Zoomable Circle Packing* com *zoom* aplicado ao Estudante 08



Fonte: elaborado pelo autor.

Para a visualização Sunburst Partition (Figura 30), as cores separam cada estudante que realizou a atividade que está sendo analisada. Assim como na visualização para um estudante, na visualização para a turma é possível escolher a opção no navegador para apresentar a proporção das palavras relevantes encontradas ou apenas a contagem.

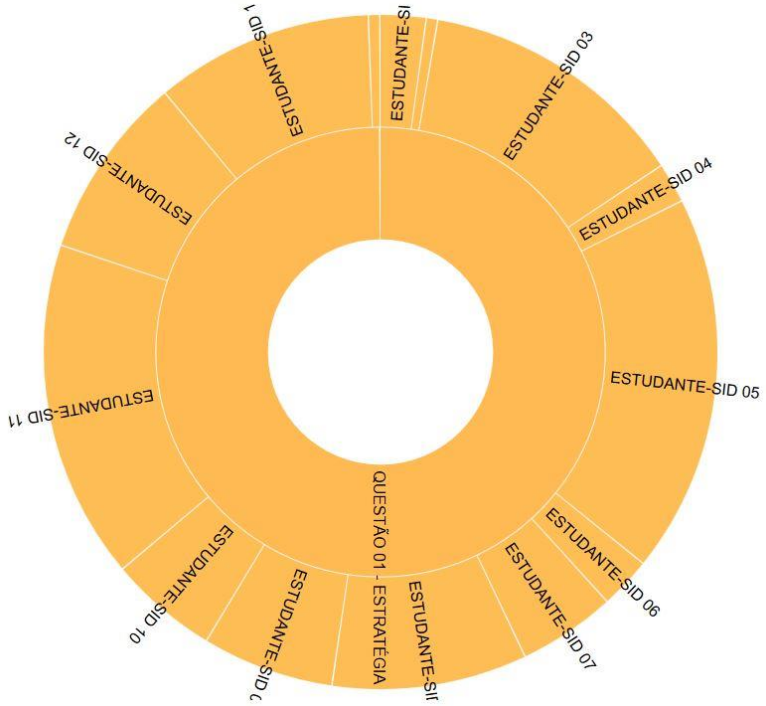
Figura 30 - *Sunburst Partition* para uma turma em uma atividade



Fonte: elaborado pelo autor.

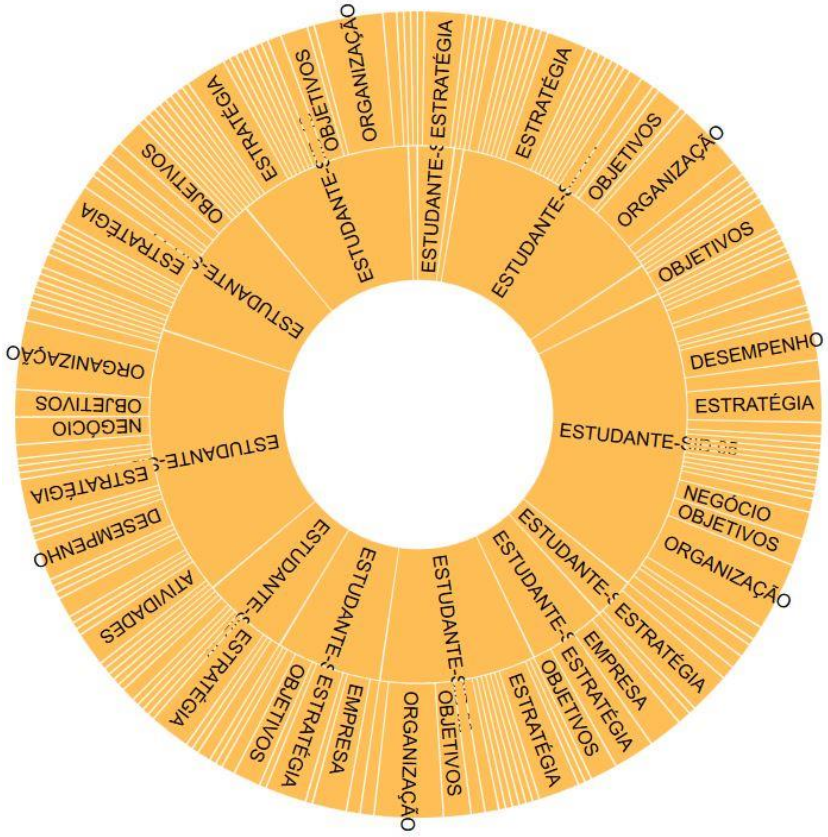
A visualização *Bilevel Partition* (Figura 31) apresenta a atividade com os estudantes separados por fatias. Ao clicar na área do estudante, um primeiro nível de aproximação é apresentado, mostrando também as palavras da atividade de cada estudante (Figura 32). Ao clicar novamente na área de algum estudante específico ou em alguma das palavras de sua atividade analisada, mais um nível de aproximação é gerado, desta vez mostrando somente as palavras deste estudante para esta atividade (Figura 33). Algumas palavras analisadas podem ficar ocultas quando as fatias forem muito finas ao ponto de não caber a expressão dentro da área destinada. Neste caso, basta posicionar o ponteiro do mouse sobre a área para visualizar a palavra a ser apresentada.

Figura 31 - *Bilevel Partition* para uma turma em uma atividade



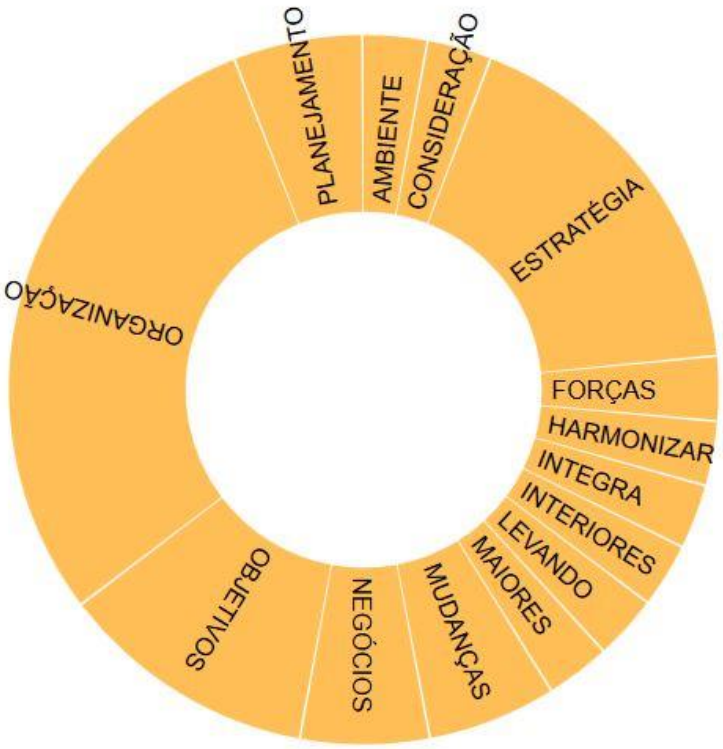
Fonte: elaborado pelo autor.

Figura 32 - *Bilevel Partition* com primeiro nível de aproximação



Fonte: elaborado pelo autor.

Figura 33 - *Bilevel Partition* com *zoom* aplicado ao Estudante 08



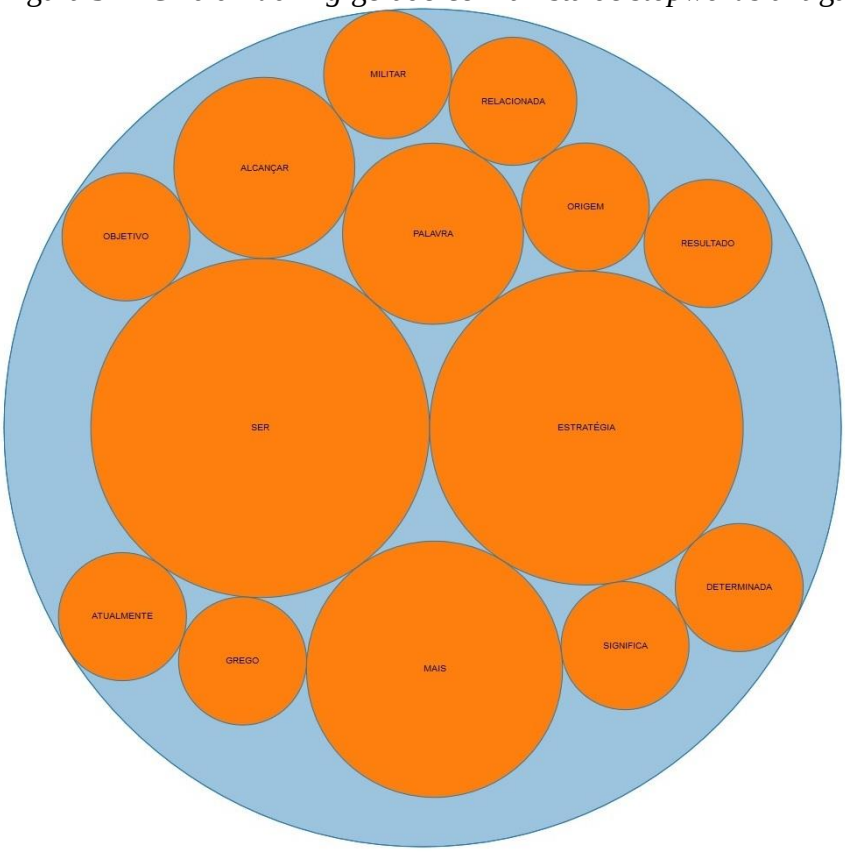
Fonte: Elaborado pelo autor.

4.4 ANÁLISE DOS RESULTADOS

Este capítulo apresenta a análise dos resultados obtidos através das melhorias desenvolvidas na ferramenta Análise de Produções Textuais. Também será realizado um comparativo entre as novas visualizações implementadas a fim de identificar os pontos fortes e fracos de cada uma, identificando se ela é apropriada ou não para determinados cenários de análise.

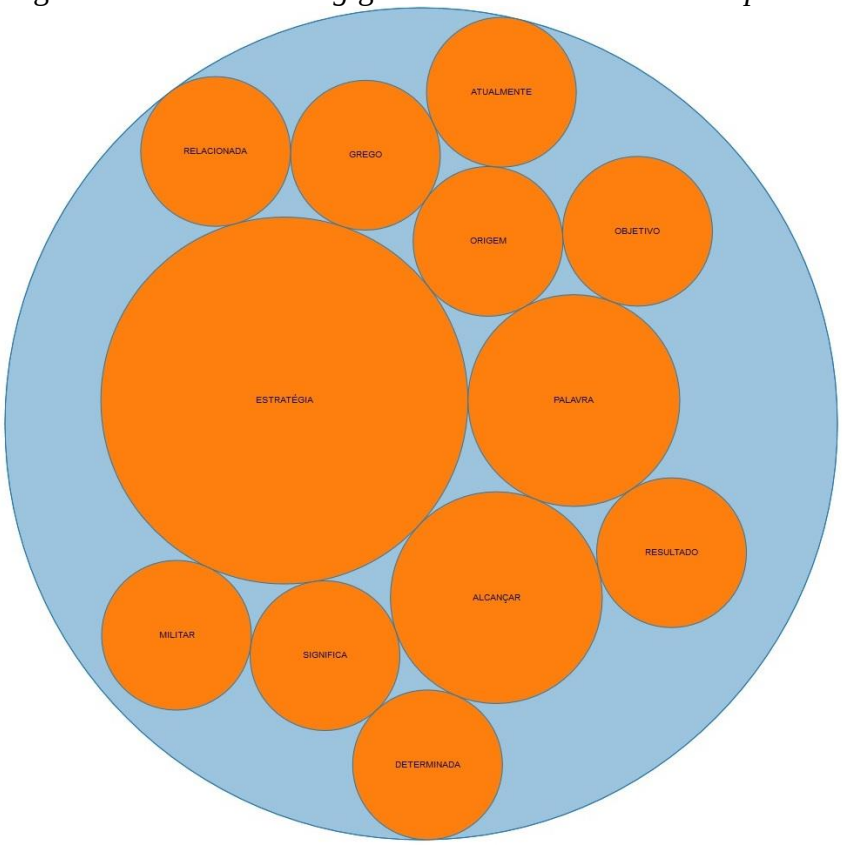
A implementação da nova lista de *stopwords* apresentou uma melhoria significativa em relação aos resultados obtidos nas visualizações. Palavras que antes eram consideradas relevantes (Figura 34) como o verbo “ser” e o advérbio “mais”, mas que isoladas não tem nenhum sentido para a análise, deixaram de ser apresentadas na nova versão (Figura 35).

Figura 34 - *Circle Packing* gerado com a lista de *stopwords* antiga



Fonte: elaborado pelo autor.

Figura 35 - *Circle Packing* gerado com a nova lista de *stopwords*



Fonte: elaborado pelo autor.

4.4.1 Static Circle Packing

A visualização *Static Circle Packing* apresenta os dados de forma estática na tela. Entre os pontos fortes, pode-se destacar a possibilidade de uma melhor comparação entre vários estudantes por parte do professor, visto que todos os termos relevantes são mostrados simultaneamente para cada estudante quando a análise é realizada para toda a turma. Outro ponto forte é a possibilidade de saber a quantidade de ocorrências de cada termo no texto original, bastando posicionar o ponteiro do mouse sobre a palavra desejada. Já como pontos fracos, deve-se considerar o fato de que esta visualização não é adequada quando se está analisando uma turma muito grande (muitos estudantes), visto que a distribuição dos dados na tela por esta forma de visualização não é tão otimizada, resultando em abreviação de algumas palavras. No entanto, pode-se visualizar as palavras completas apontando o cursor do mouse sobre elas. Outro ponto fraco é que esta visualização não mostra explicitamente o nome de cada estudante, porém pode-se visualizar o nome de cada estudante posicionando o ponteiro do mouse sobre seu círculo.

4.4.2 Zoomable Circle Packing

A visualização *Zoomable Circle Packing* se assemelha a *Static Circle Packing*, porém apresenta os dados de forma mais dinâmica. Como pontos fortes desta visualização pode-se citar a possibilidade de aproximação (*zoom*) em cada estudante quando estamos realizando a análise para uma turma, sendo possível assim ter uma melhor percepção para um estudante específico. Outro ponto forte é a possibilidade de alternar as informações utilizando o clique do mouse. Por exemplo, quando se realiza a análise para uma turma, a primeira informação mostrada é o nome da atividade, sendo que ao realizar um clique com o mouse na visualização o nome da atividade desaparece, dando lugar ao nome de cada estudante. Por outro lado, um ponto fraco desta visualização é a impossibilidade de visualizar todas as palavras analisadas de todos os estudantes em simultâneo, visto que quando se analisa uma turma inteira as palavras relevantes só aparecem ao realizar a aproximação no aluno.

4.4.3 Sunburst Partition

A Sunburst Partition é uma forma de visualização dinâmica. Como ponto forte desta visualização pode-se citar a possibilidade de ter uma noção geral de número de termos relevantes encontrados nas atividades analisadas, ou também do número de termos com mais ocorrências, visto que esta visualização possui uma opção para alternar entre a contagem e a

proporção de ocorrências das palavras analisadas. Entre os pontos fracos, pode-se citar o fato de que esta visualização não mostra informações textuais, tais como nome dos estudantes ou as palavras de cada atividade, sendo indicada apenas para uma comparação geral do número de palavras relevantes encontradas em cada atividade, mas não quando se deseja saber quais palavras estão presentes.

4.4.4 Bilevel Partition

A visualização Bilevel Partition apresenta os dados de forma dinâmica, incluindo recursos multiníveis de aproximação. Entre os pontos fortes desta visualização, pode-se citar o fato de ela ser adequada para praticamente qualquer cenário de análise, visto que seus recursos de *zoom* possibilitam apresentar somente os estudantes da turma, ou os estudantes com as respectivas palavras relevantes encontradas no texto, ou somente um estudante selecionado e as palavras de sua atividade. Outo ponto forte desta visualização é que ela apresenta o número de ocorrências de cada palavra colocando o ponteiro do mouse sobre as mesmas. Um ponto fraco desta visualização é que dependendo do número ou tamanho de palavras, algumas podem aparecer cortadas ou omitidas, porém pode-se visualizar a palavra completa posicionando o cursor do mouse sobre a área em questão.

4.4.5 Análise comparativa

A fim de melhorar a compreensão sobre os recursos oferecidos em cada uma das novas visualizações implementadas, foi elaborada uma tabela comparativa (Tabela 2). Definiu-se um conjunto de critérios, redigidos em forma de questões, para avaliar o desempenho de cada visualização na tarefa de auxílio ao professor. Para todos os critérios, foram feitos testes em um computador com o navegador Mozilla Firefox versão 57.0 instalado e apresentando a imagem em um monitor de 24” (polegadas) em resolução 1920x1080 (1080p / *Full HD*).

Tabela 2 - Análise comparativa das novas visualizações implementadas

	Static Circle Packing	**Zoomable Circle Packing**	**Sunburst Partition**	**Bilevel Partition**
As informações cabem na tela?	Sim	Sim	Sim	Sim
A análise individual fica visível e legível?	Sim	Sim	Não	Sim
A análise para a turma fica visível e legível?	Sim	Sim	Não	Sim
A visualização oferece recursos dinâmicos?	Não	Sim	Sim	Sim
Tem recursos aproximação (*zoom*)?	Não	Sim	Não	Sim
Fornece contagem das palavras?	Sim	Não	Não	Sim
Usa cores para a representação de palavras, estudantes ou atividades?	Sim	Sim	Sim	Sim
Total de critérios atendidos	**5**	**6**	**3**	**7**

Fonte: elaborado pelo autor.

Conforme mostra a Tabela 2, a visualização *Bilevel Partition* atende o maior número de critérios, sendo recomendada para praticamente qualquer cenário de análise. A visualização *Sunburst Partition* foi a que atendeu o menor número de critérios, sendo indicada apenas para se obter uma noção geral do número de palavras relevantes ou proporção de ocorrências de palavras nas atividades. A visualização *Static Circle Packing* é recomendada caso o usuário deseje visualizar todos os termos analisados ao mesmo tempo para uma turma ou para mais de uma atividade. A visualização *Zoomable Circle Packing* é indicada caso a análise seja feita em uma turma com muitos alunos, pois fornece o recurso de *zoom*.

Para finalizar, considera-se que cada visualização tem recursos que possibilitam uma percepção diferente sobre os dados. Como Judelman (2004) afirma, os recursos que permitem a análise semântica de textos em geral necessitam de suporte na forma de aplicação de várias técnicas de visualização conjuntamente. Cada recurso é complementar aos outros.

5 CONCLUSÃO

Este capítulo apresenta uma síntese das atividades desenvolvidas durante a elaboração deste trabalho. Também serão apresentados os resultados obtidos e sugestões para trabalhos futuros.

5.1 SÍNTESE DO TRABALHO

A aprendizagem automática em dados textuais é um processo importante na extração do conhecimento, identificando palavras, termos e expressões relevantes em uma produção textual. Todavia, é necessário utilizar técnicas de análise textual e métodos de visualização gráfica dos dados para que o usuário possa ter melhor proveito desta análise.

O objetivo geral deste trabalho foi aprimorar uma ferramenta de análise de produções textuais no ambiente educacional a partir da melhoria de técnicas e algoritmos já existentes. Por este motivo, buscou-se trazer novas formas de visualização e melhorar processo de remoção de *stopwords*.

Para que o objetivo do trabalho fosse atingido, foram abordados conceitos sobre o processo de aprendizagem automática em dados textuais. Também foi obtida uma visão geral sobre a ferramenta Análise de Produções Textuais para saber como esta atua e encontrar pontos que pudessem ser trabalhados.

Após a conclusão das implementações das melhorias propostas, foram realizadas atividades com uma turma de Sistemas de Informação e Decisão de uma universidade no segundo período letivo de 2017, que serviram como amostra do alvo de estudo, possibilitando a validação das modificações realizadas no sistema. A partir da análise dos dados fornecidos pela ferramenta, foi possível concluir que as alterações eram válidas e apresentaram resultados mais refinados ao usuário. Vale ressaltar que de acordo com Card, Mackinlay e Shneiderman (1999) o modelo de referência para construção de visualização de informações requer a participação do usuário para que este possa moldar e sugerir formas de visualização conforme sua necessidade. Portanto, testes e estudos de caso em conjunto com o(s) professor(es) ainda são necessários.

5.2 RESULTADOS E CONTRIBUIÇÕES

A área de Sistemas de Informação passa por constantes atualizações e evoluções. Um profissional, para conseguir acompanhar estas mudanças, deve sempre estar apto a lidar com novas áreas de estudo, buscando a resolução de problemas computacionais dos mais diversos

tipos. Desta maneira, o esperado é que este profissional desenvolva a capacidade de pesquisa, para que possa sempre buscar novos conhecimentos, acompanhando as necessidades do mercado e de usuários de diversas áreas.

Neste trabalho, foi proposto um projeto de aplicação prática, que visou evoluir um sistema de análise de produções textuais, aprimorando técnicas e algoritmos existentes. Como produto deste trabalho obteve-se uma ferramenta que traz resultados úteis, precisos e completos para o usuário (professor). Com isso, novos conhecimentos adquiridos foram agregados aos já obtidos, contribuindo para o enriquecimento da área de aprendizagem automática em dados textuais.

Com os testes realizados, foi possível verificar que as modificações fornecem ao professor mecanismos visuais para que ele possa realizar uma análise textual automática das produções dos estudantes de forma individual e coletiva pela comparação entre os textos dos estudantes. Uma contribuição para a área de visualização de informações constituiu em identificar, selecionar, aplicar e avaliar formas de visualização de textos para produções textuais de estudantes.

5.3 TRABALHOS FUTUROS

O presente trabalho abordou a evolução de uma ferramenta de análise de produções textuais para o ambiente educacional, aprimorando técnicas já existentes. Para isto, foram feitas melhorias na filtragem de *stopwords* e implementadas novas formas de visualizações gráficas.

A análise textual a partir de recursos visuais pode ser redutora e simplificar demais as ideias ou conceitos que o autor do texto tentou expressar. Estas análises não foram realizadas neste estudo, mas identifica-se que seria importante complementar este trabalho comparando as visualizações com as produções dos estudantes.

Conforme explicado anteriormente, o processo de construção de visualização de informações requer a interação com o usuário a fim de testar a efetividade das estruturas visuais para a cognição humana. Desta forma, como sugestão para trabalhos futuros, testes e geração de novas formas de visualização com a participação do professor se fazem necessários. Estes testes devem incluir tarefas de análise textual em sala de aula, com produções textuais dos estudantes.

Outra sugestão para trabalhos futuros sobre a ferramenta é substituir a versão implementada da visualização Sunburst Partition por uma versão que apresente os dados textuais (*labels*), a fim de obter-se uma visualização que atenda mais cenários de análise.

REFERÊNCIAS

ARANHA, C. N. **Uma abordagem de pré-processamento automático para mineração de textos em português: sob o enfoque da inteligência computacional**. Pontifácia Universidade Católica do Rio de Janeiro. Rio de Janeiro, p. 144. 2007.

BERRY, M. W.; DRMAC, Z.; JESSUP, E. R. Matrices, Vector Spaces, and Information Retrieval. **SIAM Review**, Philadelphia, v. 41, n. 2, p. 335-362, 1999.

BOSTOCK, M. D3.js - Data-Driven Documents. **D3 Data-Driven Documents**, 2013. Disponivel em: <https://d3js.org/>. Acesso em: 13 maio 2017.

CARD, S. K.; MACKINLAY, J. D.; SHNEIDERMAN, B. **Readings in Information Visualization:** Using Vision to Think. São Francisco: Morgan Kaufmann, 1999.

FELDMAN, R.; SANGER, J. **The text mining handbook:** advanced approaches in analyzing unstructured data. New York: Cambridge University Press, 2006.

FREITAS, C. M. D. S. et al. Introdução à Visualização de Informações. **RITA – Revista de Informática Teórica e Aplicada**, Porto Alegre, v. 8, n. 2, p. 143-158, 2001.

GIL, R. D. **Desenvolvimento de um Sistema de Análise de Semântica Latente para Avaliar Produções Textuais**. Universidade de Caxias do Sul. Caxias do Sul. 2016.

HAN, J.; KAMBER, M. **Data Mining:** Concepts and Techniques. 2ª. ed. São Francisco: Morgan Kaufmann Publishers, 2006.

IBM. IBM Watson - Build Your Cognitive Business with IBM. **IBM Watson**, 2017. Disponivel em: <https://www.ibm.com/watson/>. Acesso em: 29 maio 2017.

JUDELMAN, G. B. **Knowledge Visualization: Problems and Principles for Mapping the Knowledge Space**. International School of New Media. Lübeck. 2004.

KEIM, D. A. Information Visualization and Visual Data Mining. **IEEE Transactions on Visualization and Computer Graphics**, v. 8, n. 1, p. 1-8, Jan/Mar 2002.

KLEMANN, M.; REATEGUI, E.; RAPKIEWICZ, C. Análise de Ferramentas de Mineração de Textos para Apoio à Produção Textual. **Anais do XXII SBIE - XVII WIE**, Aracaju, p. 1100-1103, novembro 2011.

LANDAUER, T. K.; FOLTZ, P. W.; LAHAM, D. An Introduction to Latent Semantic Analysis. **Discourse Processes**, 25, 1998. 259-284.

LOVATO, G. **Aplicação da Mineração de Textos na Análise de Produções Textuais**. Univeridade de Caxias do Sul. Caxias do Sul. 2015.

LOVATO, G. et al. Uma Ferramenta Visual para Análise de Produções Textuais dos Estudantes. **RENOTE - Revista Novas Tecnologias na Educação**, Porto Alegre, v. 14, n. 2, 2016.

MACEDO, A. L. et al. **Using Text-Mining to Support the Evaluation of Texts**. Education and Technology for a Better World. Berlim: Springer. 2009. p. 368-377.

MARQUES, R. L.; DUTRA, I. **Redes Bayesianas: o que são, para que servem, algoritmos e exemplos de aplicações**. Universidade Federal do Rio de Janeiro. Rio de Janeiro. 2002.

MICROSOFT. Serviços Cognitivos - Aplicativos de Inteligência | Microsoft Azure. **Microsoft Azure**, 2017. Disponivel em: <https://azure.microsoft.com/pt-br/services/cognitive-services/>. Acesso em: 29 maio 2017.

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. **Mineração de textos**. Universidade Federal de Goiás. [S.l.], p. 1-29. 2007.

OESTREICHER, E. B.; SANTO, L. S. D. E.; JÚNIOR, E. R. Análise estratégica: um estudo na lanchonete e pizzaria "Bom Te Ver". **VII Convibra Administração – Congresso Virtual Brasileiro de Administração**, São Paulo, 19 a 21 novembro 2010. 17.

ORENGO, V. M.; HUYCK, C. **A stemming algorithm for the portuguese language**. EIGHTH INTERNATIONAL SYMPOSIUM ON STRING PROCESSING AND INFORMATION RETRIEVAL. London: SPIRE. 2001. p. 186-193.

RUSSELL, S.; NORVIG, P. **Inteligência Artificial**. 3ª. ed. Rio de Janeiro: Elsevier, 2013.

SILVA, E. F. D. A. E. **Um sistema para extração de informação em referências bibliográficas baseado em aprendizagem de máquina**. Universidade Federal do Pernambuco. Recife. 2004.

SILVA, M. C. D. O texto escrito. **Brasil Escola**, 2009. Disponivel em: <http://brasilescola.uol.com.br/redacao/texto-escrito.html>. Acesso em: 10 abril 2017.

SOBEK. Minerador de Textos Sobek. **Sobek Mining**, 2017. Disponivel em: <http://sobek.ufrgs.br/index.html>. Acesso em: 17 abril 2017.

TAGCROWD. TagCrowd: create your own word cloud from any text. **TagCrowd**, 2017. Disponivel em: <http://tagcrowd.com/>. Acesso em: 17 abril 2017.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introdução ao Data Mining:** Mineração de Dados. 1ª. ed. Rio de Janeiro: Ciência Moderna, 2009.

TAN, Y.; WANG, J. A Support Vector Machine with a Hybrid Kernel and Minimal Vapnik-Chervonenkis Dimension. **IEEE Transactions on Knowledge and Data Engineering**, v. 16, n. 4, p. 385-395, abril 2004.

TEXTALYSER. Text analysis, wordcount, keyword density analyzer, prominence analysis. **Textalyser**, 2017. Disponivel em: <http://textalyser.net/>. Acesso em: 17 abril 2017.

WANG, L.; ZHAO, X. **Improved KNN Classification Algorithms Research in Text Categorization**. 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet). Yichang: [s.n.]. 2012. p. 1848-1852.

WEBBER, C. G.; CINI, G.; LIMA, M. D. F. W. D. P. Facilitando a Análise se Dados Educacionais através de Ferramentas de Visualização. **RENOTE - Revista Novas Tecnologias na Educação**, Porto Alegre, v. 11, n. 3, 2013.

WORDCOUNTER. WordCounter - Count Words & Correct Writing. **WordCounter**, 2017. Disponivel em: <https://wordcounter.net/>. Acesso em: 17 abril 2017.

APÊNDICE A – LISTA DE STOPWORDS LEGADO

Lista de *stopwords*					
A	DESSA	ESTAMOS	MUITAS	PODER	SOBRE
À	DESSAS	ESTÃO	MUITO	PODERIA	SUA
AGORA	DESSE	ESTAS	MUITOS	PODERIAM	SUAS
AINDA	DESSES	ESTAVA	NA	PODIA	TALVEZ
ALGUÉM	DESTA	ESTAVAM	NÃO	PODIAM	TAMBÉM
ALGUM	DESTAS	ESTÁVAMOS	NAS	POIS	TAMPOUCO
ALGUMA	DESTE	ESTE	NEM	POR	TE
ALGUMAS	DESTE	ESTES	NENHUM	PORÉM	TEM
ALGUNS	DESTES	ESTOU	NESSA	PORQUE	TENDO
AMPLA	DEVE	EU	NESSAS	POSSO	TENHA
AMPLAS	DEVEM	FAZENDO	NESTA	POUCA	TER
AMPLO	DEVENDO	FAZER	NESTAS	POUCAS	TEU
AMPLOS	DEVER	FEITA	NINGUÉM	POUCO	TEUS
ANTE	DEVERÁ	FEITAS	NO	POUCOS	TI
ANTES	DEVERÃO	FEITO	NOS	PRIMEIRO	TIDO
AO	DEVERIA	FEITOS	NÓS	PRIMEIROS	TINHA
AOS	DEVERIAM	FOI	NOSSA	PRÓPRIA	TINHAM
APÓS	DEVIA	FOR	NOSSAS	PRÓPRIAS	TODA
AQUELA	DEVIAM	FORAM	NOSSO	PRÓPRIO	TODAS
AQUELAS	DISSE	FOSSE	NOSSOS	PRÓPRIOS	TODAVIA
AQUELE	DISSO	FOSSEM	NUM	PUDE	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	PODENDO	SOB	

Fonte: Gil (2016).

APÊNDICE B – LISTA DE STOPWORDS NOVA

Lista de *stopwords*					
a	caminho	destas	esse	faco	houveríamos
acerca	catorze	deste	esses	fez	houvesse
adeus	cedo	destes	esta	fim	houvessem
agora	cento	deve	estado	final	houvéramos
ainda	certamente	devem	estamos	foi	houvéssemos
alem	certeza	deverá	estar	fomos	há
algo	cima	dez	estará	for	hão
algum	cinco	dezanove	estas	fora	iniciar
algumas	coisa	dezasseis	estava	foram	inicio
alguns	com	dezassete	estavam	forem	ir
ali	como	dezoito	este	forma	irá
além	comprido	dia	esteja	formos	isso
ambas	conhecido	diante	estejam	fosse	ista
ambos	conselho	direita	estejamos	fossem	iste
ano	contra	dispoe	estes	foste	isto
anos	contudo	dispoem	esteve	fostes	já
antes	corrente	diversa	estive	fui	lado
ao	cuja	diversas	estivemos	fôramos	lhe
aonde	cujas	diversos	estiver	fôssemos	lhes
aos	cujo	diz	estivera	geral	ligado
apenas	cujos	dizem	estiveram	grande	local
apoio	custa	dizer	estiverem	grandes	logo
apontar	cá	do	estivermos	grupo	longe
apos	da	dois	estivesse	ha	lugar
após	daquela	dos	estivessem	haja	lá
aquela	daquelas	doze	estiveste	hajam	maior
aquelas	daquele	duas	estivestes	hajamos	maioria
aquele	daqueles	durante	estivéramos	havemos	maiorias
aqueles	dar	dá	estivéssemos	havia	mais
aqui	das	dão	estou	hei	mal
aquilo	de	dúvida	está	hoje	mas
as	debaixo	e	estás	hora	me
assim	dela	ela	estávamos	horas	mediante
através	delas	elas	estão	houve	meio
atrás	dele	ele	eu	houvemos	menor
até	deles	eles	exemplo	houver	menos
aí	demais	em	falta	houvera	meses
baixo	dentro	embora	fará	houveram	mesma
bastante	depois	enquanto	favor	houverei	mesmas
bem	desde	entao	faz	houverem	mesmo
boa	desligado	entre	fazeis	houveremos	mesmos
boas	dessa	então	fazem	houveria	meu
bom	dessas	era	fazemos	houveriam	meus
bons	desse	eram	fazer	houvermos	mil
breve	desses	essa	fazes	houverá	minha
cada	desta	essas	fazia	houverão	minhas

momento	oitava	promeiro	sejamos	tentar	um
muito	oitavo	propios	sem	tentaram	uma
muitos	oito	proprio	sempre	tente	umas
máximo	onde	própria	sendo	tentei	uns
mês	ontem	próprias	ser	ter	usa
na	onze	próprio	serei	terceira	usar
nada	os	próprios	seremos	terceiro	vai
nao	ou	próxima	seria	terei	vais
naquela	outra	próximas	seriam	teremos	valor
naquelas	outras	próximo	será	teria	veja
naquele	outro	próximos	serão	teriam	vem
naqueles	outros	puderam	seríamos	terá	vens
nas	para	pôde	sete	terão	ver
nem	parece	põe	seu	teríamos	verdade
nenhuma	parte	põem	seus	teu	verdadeiro
nessa	partir	quais	sexta	teus	vez
nessas	paucas	qual	sexto	teve	vezes
nesse	pegar	qualquer	sim	tinha	viagem
nesses	pela	quando	sistema	tinham	vindo
nesta	pelas	quanto	sob	tipo	vinte
nestas	pelo	quarta	sobre	tive	você
neste	pelos	quarto	sois	tivemos	vocês
nestes	perante	quatro	somente	tiver	vos
no	perto	que	somos	tivera	vossa
noite	pessoas	quem	sou	tiveram	vossas
nome	pode	quer	sua	tiverem	vosso
nos	podem	quereis	suas	tivermos	vossos
nossa	poder	querem	são	tivesse	vários
nossas	poderá	queremas	sétima	tivessem	vão
nosso	podia	queres	sétimo	tiveste	vêm
nossos	pois	quero	só	tivestes	vós
nova	ponto	questão	tal	tivéramos	zero
novas	pontos	quieto	talvez	tivéssemos	à
nove	por	quinta	tambem	toda	às
novo	porque	quinto	também	todas	área
novos	porquê	quinze	tanta	todo	é
num	portanto	quáis	tantas	todos	éramos
numa	posicão	quê	tanto	trabalhar	és
numas	possivelmente	relacão	tarde	trabalho	último
nunca	posso	sabe	te	treze	
nuns	possível	sabem	tem	três	
não	pouca	saber	temos	tu	
nível	pouco	se	tempo	tua	
nós	poucos	segunda	tendes	tuas	
número	povo	segundo	tenha	tudo	
o	primeira	sei	tenham	tão	
obra	primeiras	seis	tenhamos	tém	
obrigada	primeiro	seja	tenho	têm	
obrigado	primeiros	sejam	tens	tínhamos	

APÊNDICE C – QUESTÕES APLICADAS AOS ESTUDANTES

Questões sobre o artigo “Análise Estratégica: um estudo na lanchonete e pizzaria ‘Bom Te Ver’”:

1. Defina o conceito de estratégia.

2. Defina o conceito de planejamento.

3. Qual a relação entre estratégia e planejamento?

4. Como se chama o procedimento para avaliar estrategicamente uma organização, identificando seus pontos fortes e fracos?

5. O que é uma ameaça para uma organização?

6. O que é uma oportunidade para uma organização?

7. O que é o ambiente interno da organização?

8. O que é o ambiente externo da organização?

9. Para que serve a análise de dados?

10. Apresente brevemente a empresa alvo do estudo de caso: Lanchonete e Pizzaria Bom Te Ver.

11. Quadrante 1 – o que foi identificado?

12. Quadrante 2 – o que foi identificado?

13. Quadrante 3 – o que foi identificado?

14. Quadrante 4 – o que foi identificado?

APÊNDICE D – MANUAL DO PRODUTO

1. INSTALAÇÃO

Neste capítulo serão explicados os procedimentos necessários para utilização da ferramenta Análise de Produções Textuais. Também é descritos como proceder para importar o projeto no ambiente de desenvolvimento (IDE), caso algum desenvolvedor precise alterar o sistema futuramente.

1.1 PARA O DESENVOLVEDOR

Este capítulo é direcionado aos desenvolvedores que desejam realizar alterações no sistema. São descritos os softwares necessários e as etapas para importar o projeto no ambiente de desenvolvimento.

1.1.1 Softwares necessários

Para importação dos arquivos de desenvolvimento serão necessários os recursos abaixo:

a) o código-fonte para o sistema Análise de Produções Textuais;
b) Microsoft Visual Studio: a versão atual do sistema Análise de Produções Textuais foi desenvolvida utilizando a versão 2017 do Visual Studio. Caso seja importado em alguma versão mais antiga, 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 versão 2.2. Os exemplos aqui contidos foram descritos utilizando esta versão, mas nada impede de utilizar outro programa para esta finalidade.

1.1.2 Instalação da versão de desenvolvimento

Para importação dos arquivos de desenvolvimento, deve-se seguir os seguintes passos:

a) instalar o servidor *web* Apache: a instalação do Apache HTTP Server está descrita no capítulo 1.3. Poderá ser utilizado outro servidor *web* conforme a preferência do desenvolvedor;

b) o sistema Análise de Produções Textuais está disponibilizado em um arquivo compactado 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 colocados dentro do diretório do servidor *web*. Conforme os exemplos deste manual ficaria conforme a seguir: “C:\Webserver\Apache2.2\htdocs”. Abaixo deste diretório deve-se copiar as pastas contidas em “Visualizações”.

1.2 PARA O USUÁRIO

Este capítulo é destinado aos usuários da ferramenta. Serão descritos os softwares e os procedimentos para instalar o programa no computador.

1.2.1 Softwares necessários

Para instalação e uso da aplicação no computador serão necessários os recursos abaixo:

a) o instalador executável (.exe) da ferramenta Análise de Produções Textuais: junto com os programas do sistema há um instalador do sistema, como o sistema foi desenvolvido utilizado a linguagem C# , só poderá ser instalado em uma máquina com alguma versão do sistema operacional Windows;
b) Apache HTTP Server: para a versão atual da ferramenta Análise de Produções Textuais foi utilizado o Apache HTTP Server versão 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

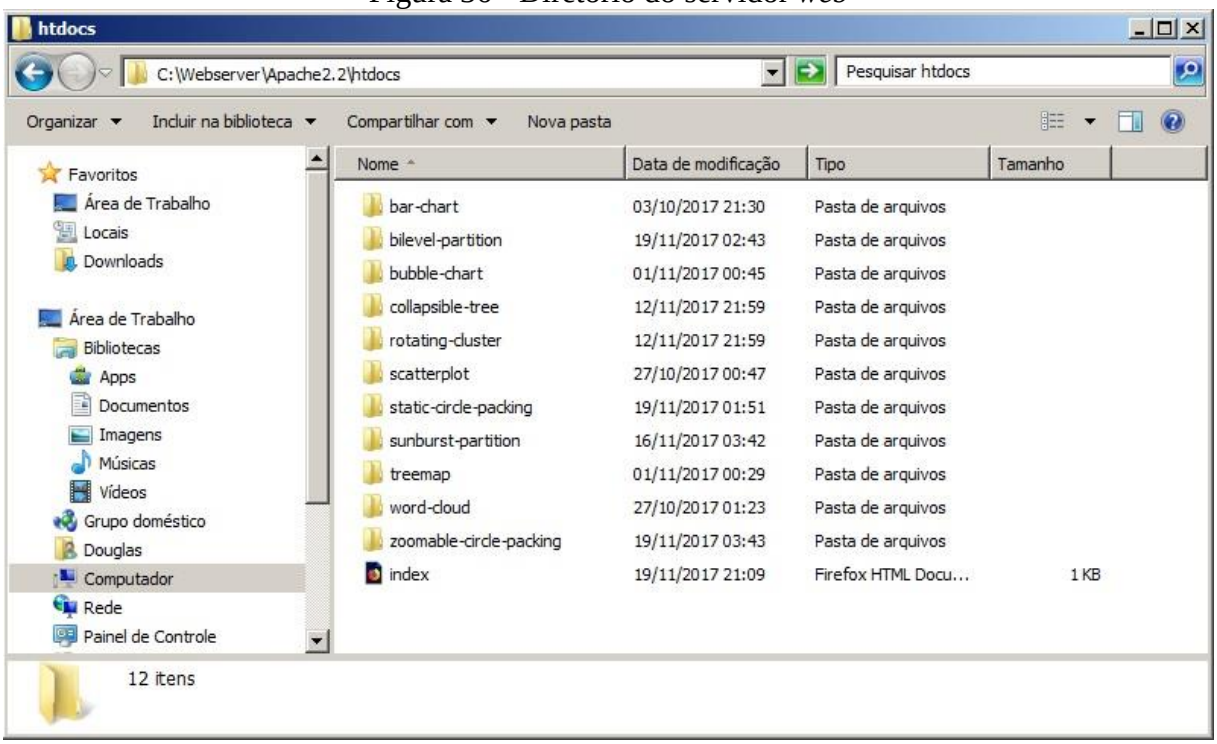
Para realizar a instalação da ferramenta Análise de Produções Textuais, deve-se seguir os seguintes passos:

a) instalar o servidor *web* Apache: a instalação do Apache HTTP Server está descrita no capítulo 1.3. Poderá ser utilizado outro servidor *web* conforme a preferência do usuário;
b) o sistema Análise de Produções Textuais está disponibilizado em um arquivo compactado contendo os seguintes diretórios:
    - Implementação;
    - Visualizações;
c) dentro do diretório "Implementação\Instalador" está o instalador da ferramenta. Executando o arquivo "setup.exe" a aplicação é instalada no computador;
d) as pastas contidas dentro de "Visualizações" possuem as implementações das visualizações que devem ser colocadas dentro do diretório do servidor web. Conforme os exemplos deste manual ficaria conforme a seguir: "C:\Webserver\Apache2.2\htdocs". Abaixo deste diretório deve-se copiar as pastas contidas em "Visualizações".

1.3 INSTALAÇÃO DO SERVIDOR WEB APACHE

O servidor *web* Apache deverá ser configurado conforme padrão. Apenas deve-se ficar atento com o caminho que se disponibilizará os arquivos que ficarão publicados pelo servidor. Na elaboração deste manual foi configurado o caminho "C:\Webserver\Apache2.2\htdocs". Neste caminho ficarão publicados os arquivos responsáveis pelas visualizações da aplicação. A Figura 36 apresenta como ficará a estrutura de diretórios após configurado conforme descrito nos capítulos 1.1.2 e 1.2.2.

Figura 36 - Diretório do servidor *web*



Fonte: elaborado pelo autor.

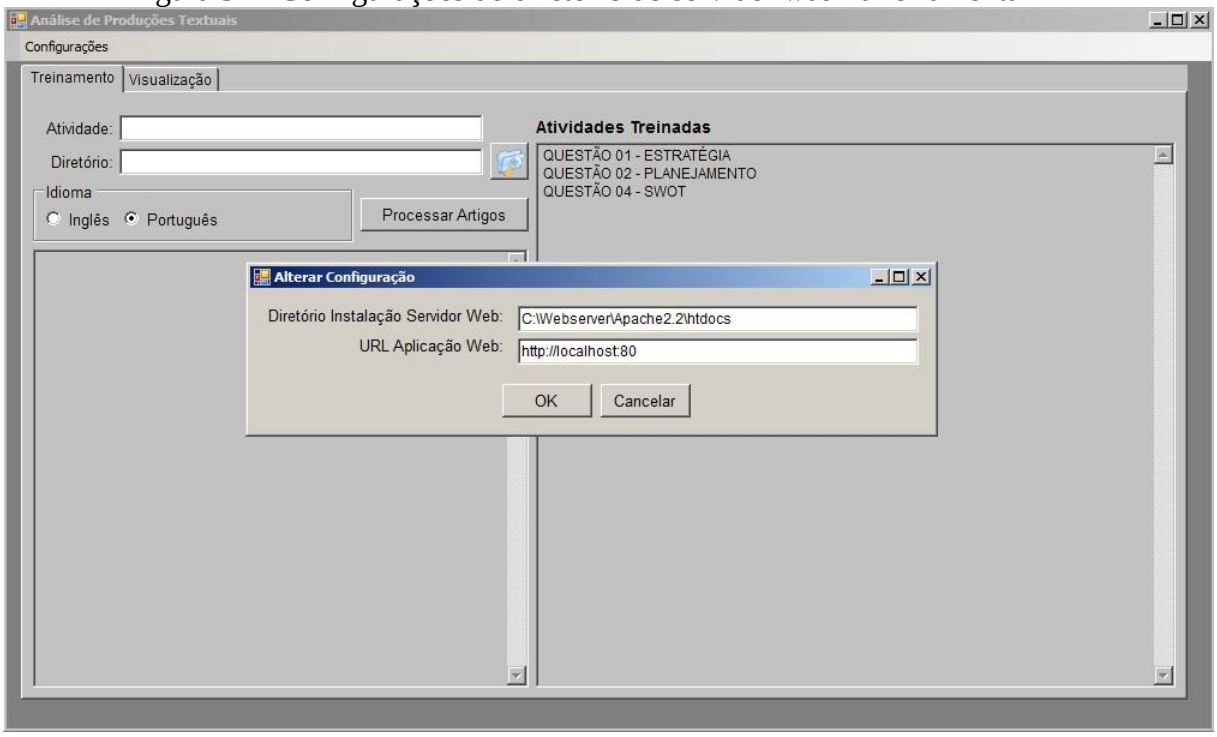
2. UTILIZAÇÃO

Neste capítulo são descritas as funcionalidades da ferramenta 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 treinamento e análise do software.

2.1 CONFIGURAÇÕES PARA PRIMEIRO USO

São necessárias duas configurações para o primeiro uso caso o servidor web tenha sido instalado com alguma parametrização diferente da qual descrita neste manual. Estas configurações são necessárias devido à integração da ferramenta com as visualizações que serão apresentadas no navegador padrão do computador.

Figura 37 - Configurações do diretório do servidor *web* na ferramenta APT



Fonte: elaborado pelo autor.

Conforme a Figura 37 é necessário configurar o diretório de instalação do servidor *web*. O que consta 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, a ferramenta instalada sempre iniciará com a última configuração parametrizada.

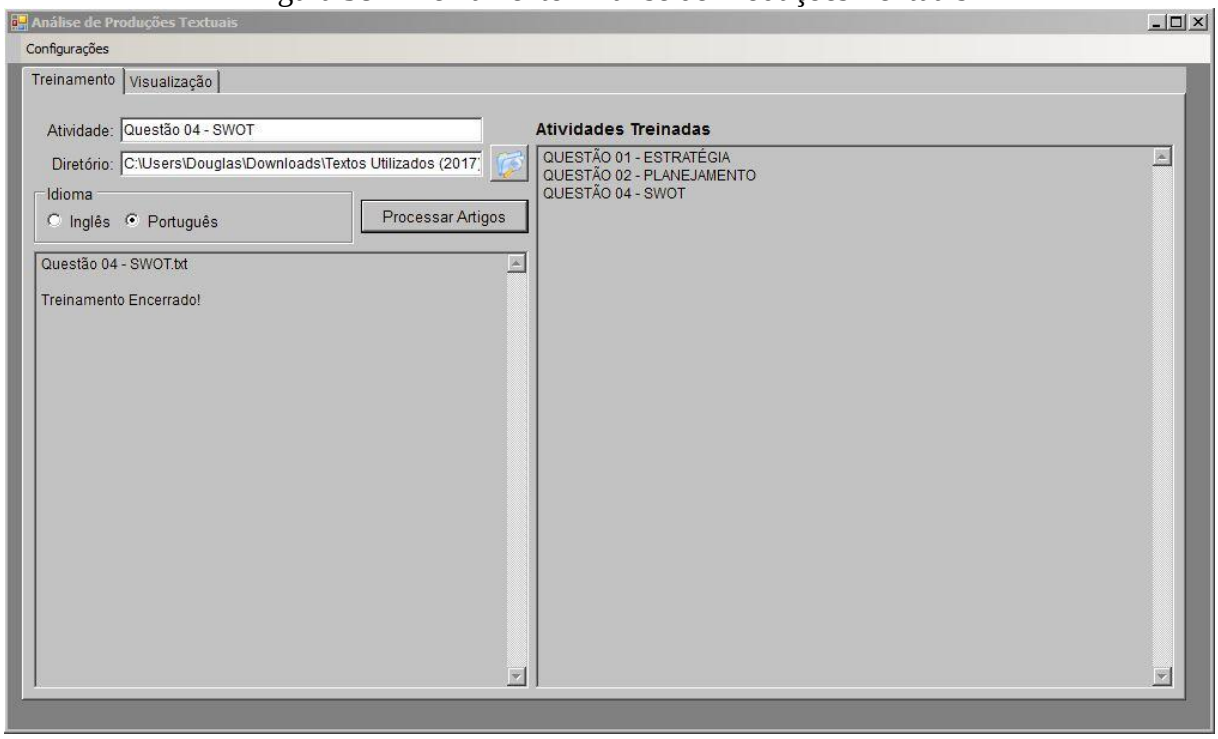
2.2 TREINAMENTO

Descrição das funcionalidades da tela:

a) Atividade: é a descrição da atividade que será treinada na ferramenta. Esta descrição é importante pois os textos analisados posteriormente deverão conter o respectivo assunto ao serem importados;
b) Diretório: deverá ser informado o diretório que contém os corpus de texto em formato .txt que serão importados no treinamento;
c) Idioma: o sistema está preparado para aceitar textos nos idiomas inglês e português. É importante informar o idioma correto, pois os textos analisados deverão estar no respectivo idioma;

d) Processar Artigos: irá importar os artigos conforme o diretório informado para o treinamento. Esta ação irá criar um diretório “C:\listasTermos” caso seja a primeira vez que a ferramenta faça algum treinamento. Este diretório armazenará as atividades treinadas;
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 na ferramenta.

Figura 38 - Treinamento Análise de Produções Textuais



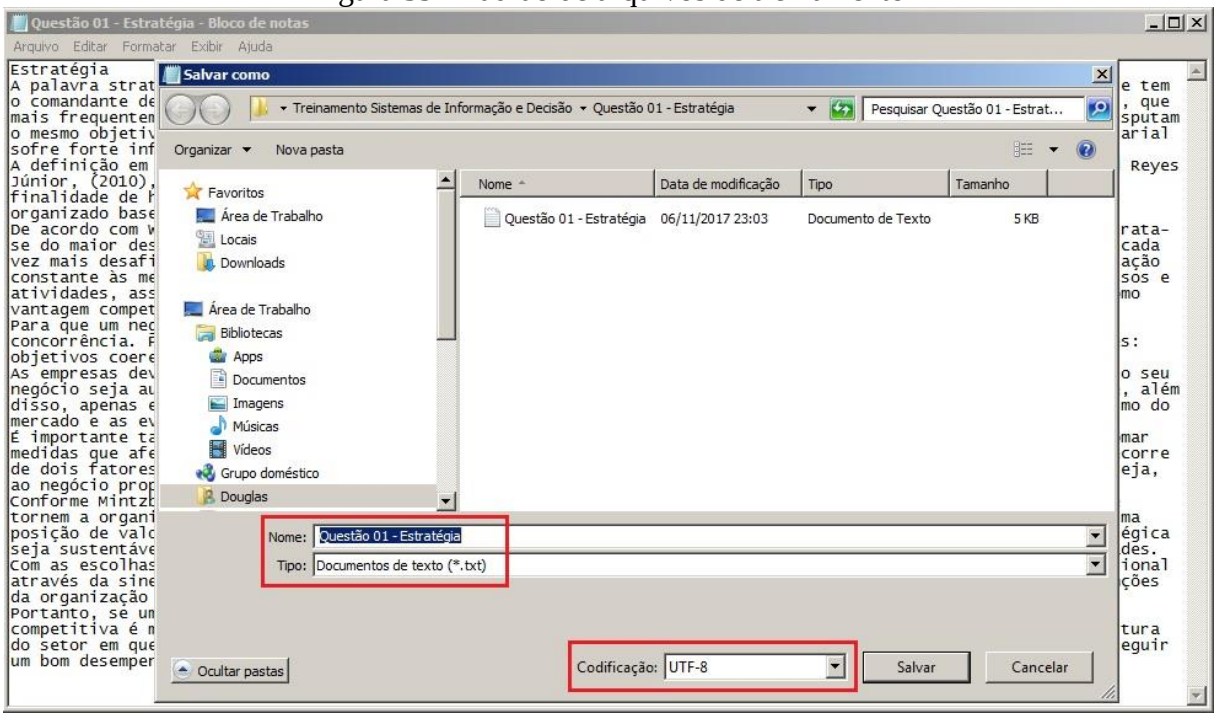
Fonte: elaborado pelo autor.

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 da ferramenta sobre determinada atividade, deve-se treinar novamente com todos os corpus de texto desta atividade.

2.2.1 Arquivos de entrada

Tanto os arquivos de entrada para o treinamento quanto para a análise devem estar no formato .txt e devem estar salvos utilizando o padrão de codificação UTF-8. A Figura 39 representa um arquivo de treinamento sendo salvo conforme os padrões necessários.

Figura 39 - Padrão de arquivos de treinamento



Fonte: elaborado pelo autor.

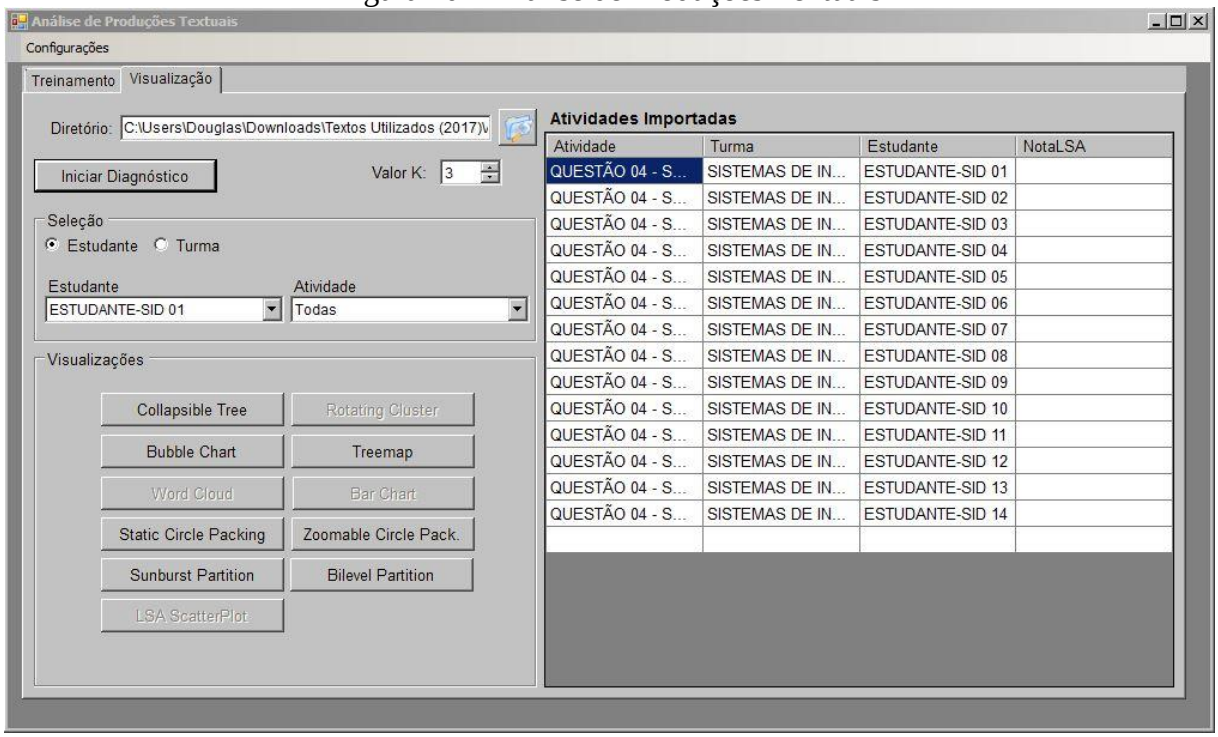
2.3 ANÁLISE

Descrição das funcionalidades da tela:

a) Diretório: deverá ser informado o diretório que contém as produções textuais em formato .txt que serão carregados para análise;
b) Iniciar Diagnóstico: irá realizar a importação das produções textuais conforme o diretório informado para carregamento. Após esta ação a ferramenta esta pronta para ser usada com o objetivo de analisar os textos importados;
c) Atividades Importadas: apresenta uma listagem dos textos que foram importados, contendo informações sobre a atividade, a turma e os estudantes de cada produção textual;
d) Seleção: a combinação dos textos para análise é representada na Figura 41. 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 serão abertas no navegador padrão da máquina. As visualizações estão disponíveis conforme a seleção abaixo:
    - Collapsible Tree – todas as seleções;

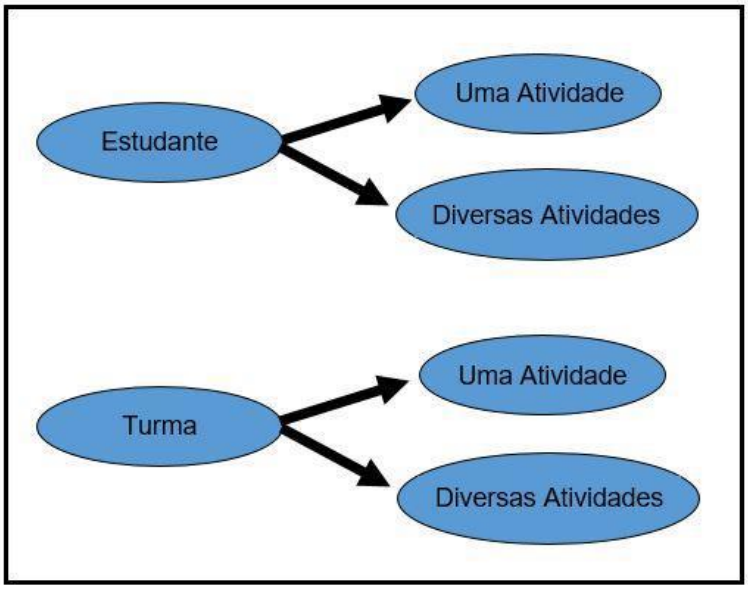
- Bubble Chart – todas as seleções;
- Word Cloud – apenas um estudante em uma atividade;
- Rotating Cluster – todas as seleções;
- Treemap – todas as seleções;
- Bar Chart – apenas uma turma em uma atividade;
- Static Circle Packing – todas as seleções;
- Zoomable Circle Packing – todas as seleções;
- Sunburst Partition – todas as seleções;
- Bilevel Partition – todas as seleções;
- LSA ScatterPlot – apenas uma turma em uma atividade.

Figura 40 - Análise de Produções Textuais



Fonte: elaborado pelo autor.

Figura 41 - Combinação para seleção de textos para análise



Fonte: Gil (2016).

2.3.1 Arquivos de entrada

As produções textuais para análise, além de estarem salvos em formato .txt e codificação UTF-8 conforme a imagem da Figura 39, devem possuir um cabeçalho que os identifique. O cabeçalho deve ser escrito conforme o padrão a seguir:

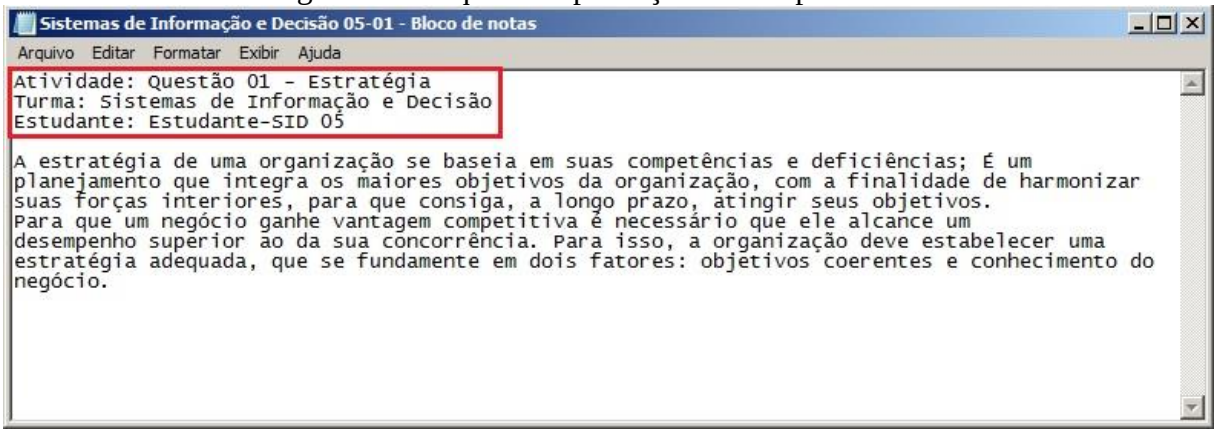
Atividade: “Nome da Atividade”;

Turma: “Nome da Turma”;

Estudante: “Nome do Estudante”.

Importante ressaltar que a atividade precisa ter exatamente o mesmo nome da atividade que foi colocada no treinamento conforme a Figura 38, inclusive com o mesmo espaçamento e acentuação, caso utilizado. Um exemplo de cabeçalho e arquivo de entrada para análise é mostrado na Figura 42.

Figura 42 - Arquivo de produção textual para análise



Fonte: elaborado pelo autor.