

UNIVERSIDADE FEDERAL DOS VALES DO JEQUITINHONHA E MUCURI

Departamento de Computação

Raphael Sousa Silva

**SUMARIZAÇÃO AUTOMÁTICA DE TEXTOS UTILIZANDO O PRAGMASUM EM
UM *CORPUS* DE ARTIGOS EDUCACIONAIS**

Diamantina

2019

Raphael Sousa Silva

**SUMARIZAÇÃO AUTOMÁTICA DE TEXTOS UTILIZANDO O PRAGMASUM EM
UM *CORPUS* DE ARTIGOS EDUCACIONAIS**

Trabalho de Conclusão de Curso apresentado ao
Curso de Sistemas de Informação da Universidade
Federal dos Vales do Jequitinhonha e Mucuri, como
requisito parcial para conclusão do curso.

Orientador: Prof. Dr. Marcus Vinícius Carvalho
Guelpe

Diamantina

2019


RAPHAEL SOUSA SILVA

**SUMARIZAÇÃO AUTOMÁTICA DE TEXTOS UTILIZANDO O
PRAGMASUM EM UM *CORPUS* DE ARTIGOS EDUCACIONAIS.**

Trabalho de conclusão de curso de
Graduação apresentado à Universidade
Federal dos Vales do Jequitinhonha e
Mucuri, como requisito parcial para
obtenção do grau de Bacharel em
Sistemas de Informação.

Orientador: Prof. Dr. Marcus Vinícius
Carvalho Guelpeli

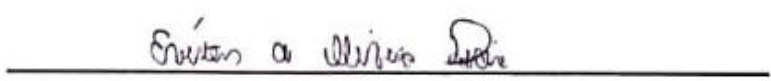
Data de aprovação 13/07/2019.



Prof. Dr. Marcus Vinícius Carvalho Guelpeli - UFVJM



Prof.º Dr. Alexandre Ramos Fonseca - UFVJM



Msc. Éverton de Oliveira Paiva – UFVJM

Dedico este trabalho à minha
família e aos meus amigos.

AGRADECIMENTOS

Quero agradecer primeiramente à Deus, por ter me proporcionado saúde e sabedoria para chegar até aqui. Aos meus pais, João Dias da Silva e Maria Helena Martins Silva, pelo apoio, aprendizado, amor e por sempre me auxiliarem em meio a todas as dificuldades. A minha irmã Renata Sousa Silva, pelo apoio e conselhos.

Agradeço aos meus amigos da República Beira do Fanado pelo companheirismo em todos os momentos. A todos aqueles que contribuíram para minha formação e estiveram, de alguma forma, envolvidos durante a minha trajetória.

Ao meu orientador Prof. Dr. Marcus Vinícius Carvalho Guelpe, pela disponibilidade, ajuda, acompanhamento e ensinamentos, fundamentais durante esta etapa. A todos aqueles que contribuíram para minha formação e estiveram, de alguma forma, envolvidos durante a minha trajetória.

Mas pra quem tem pensamento forte
O impossível é só questão de opinião.
(Charlie Brown Jr.).

RESUMO

A sumarização de textos, como um todo, é uma atividade comum em nosso cotidiano. Está presente em conversas informais à trabalhos acadêmicos. A sumarização automática busca retirar as informações mais importantes de um texto fonte e apresentá-las de uma forma mais condensada em um sumário. Pode ser aplicada em acontecimento, filme, texto, entre outras. Sumários são versões reduzidas dos textos a que se referem e contêm as ideias principais dos mesmos. Este trabalho tem como objetivo utilizar o sumarizador de textos PragmaSUM em artigos educacionais com novas técnicas de sumarização utilizando os métodos de palavras chaves do autor e palavras com maior frequência. A utilização de métodos de personificação do sumário tem como objetivo melhorar o desempenho e a precisão do PragmaSUM e seus sumários. Diante disso, foi utilizado um *corpus* formado por 50 artigos científicos da área educacional, divididos em 10 domínios diferentes. O desempenho do sumarizador foi medido pela métrica *F-Measure* presente na ferramenta ROUGE e validados com o teste estatístico *T-Student*. Os resultados obtidos apontaram que os dois métodos utilizados na sumarização obtiveram resultados equivalentes, o que indica a importância da escolha de palavras chaves adequadas para fazer a indexação dos artigos, facilitando assim a busca do usuário.

Palavras-chaves: Sumarização Automática de Textos. PragmaSUM. ROUGE. Corpus. Sumarização.

ABSTRACT

Summarizing texts is a common activity in our daily lives. It is present in informal conversations and academic works as well. Automatic summarization seeks to extract the most important information from a source text and show it in a condensed form as a summary. It can be applied to events, films, texts, etc. Summaries are reduced versions of the texts to which they refer, and contain the main ideas of the texts. This work aims to use the PragmaSUM text summarizer in educational articles combined with new techniques of summarization using author's key words and the more frequent words. The use of summary impersonation methods intendeds to improve the performance and accuracy of PragmaSUM and its summaries. Thus, it was used a corpus of 50 scientific articles about education, divided into 10 different domains. The performance of the summarizer was measured by the F-Measure metric present in the ROUGE tool, and validated by the T-Student statistical test. The obtained results indicated that the two methods used in the summarization acquired equivalent results, which indicates the importance of choosing suitable keywords to index the articles, thus facilitating user's searches.

Keywords: Automatic Summarization of Texts. PragmaSUM. ROUGE. Corpus. Summarization.

LISTA DE ILUSTRAÇÕES

Figura 1 - Esquema do ROUGE para avaliação de sumários.....	30
Figura 2 – Pasta Raiz do Corpus.	32
Figura 3 – Diagrama do Domínio Educação Especial.....	32
Figura 4 – Taxa de 80% de Compressão no PragmaSUM	33
Figura 5 – Comparação entre F-Measure no Domínio Educação Especial	37
Figura 6 – Comparação entre F-Measure no Domínio Educação Permanente	38
Figura 7 – Comparação entre F-Measure no Domínio Educação Pré-Escolar	39
Figura 8 – Comparação entre F-Measure no Domínio Ensino Aprendizagem	40
Figura 9 – Comparação entre F-Measure no Domínio Filosofia da Educação	41
Figura 10 – Comparação entre F-Measure no Domínio História da Educação	42
Figura 11 – Comparação entre F-Measure no Domínio Política Educacional	43
Figura 12 – Comparação entre F-Measure no Domínio Psicologia Educacional	44
Figura 13 – Comparação entre F-Measure no Domínio Sociologia da Educação	45
Figura 14 – Comparação entre F-Measure no Domínio Tecnologia Educacional	46
Figura 15 – Teste de variâncias equivalentes com taxa de compressão de 50% do domínio Educação Especial	59
Figura 16 – Teste de variâncias equivalentes com taxa de compressão de 70% do domínio Educação Especial	60
Figura 17 – Teste de variâncias equivalentes com taxa de compressão de 80% do domínio Educação Especial	60
Figura 18 – Teste de variâncias equivalentes com taxa de compressão de 90% do domínio Educação Especial	61
Figura 19 – Teste de variâncias equivalentes com taxa de compressão de 50% do domínio Educação Permanente	61
Figura 20 – Teste de variâncias equivalentes com taxa de compressão de 70% do domínio Educação Permanente	62
Figura 21 – Teste de variâncias equivalentes com taxa de compressão de 80% do domínio Educação Permanente	62
Figura 22 – Teste de variâncias equivalentes com taxa de compressão de 90% do domínio Educação Permanente	63

Figura 23 – Teste de variâncias equivalentes com taxa de compressão de 50% do domínio Educação Pré-Escolar	63
Figura 24 – Teste de variâncias equivalentes com taxa de compressão de 70% do domínio Educação Pré-Escolar	64
Figura 25 – Teste de variâncias equivalentes com taxa de compressão de 80% do domínio Educação Pré-Escolar	64
Figura 26 – Teste de variâncias equivalentes com taxa de compressão de 90% do domínio Educação Pré-Escolar	65
Figura 27 – Teste de variâncias equivalentes com taxa de compressão de 50% do domínio Ensino Aprendizagem	65
Figura 28 – Teste de variâncias equivalentes com taxa de compressão de 70% do domínio Ensino Aprendizagem	66
Figura 29 – Teste de variâncias equivalentes com taxa de compressão de 80% do domínio Ensino Aprendizagem	66
Figura 30 – Teste de variâncias equivalentes com taxa de compressão de 90% do domínio Ensino Aprendizagem	67
Figura 31 – Teste de variâncias equivalentes com taxa de compressão de 50% do domínio Filosofia da Educação	67
Figura 32 – Teste de variâncias equivalentes com taxa de compressão de 70% do domínio Filosofia da Educação	68
Figura 33 – Teste de variâncias equivalentes com taxa de compressão de 80% do domínio Filosofia da Educação	68
Figura 34 – Teste de variâncias equivalentes com taxa de compressão de 90% do domínio Filosofia da Educação	69
Figura 35 – Teste de variâncias equivalentes com taxa de compressão de 50% do domínio História da Educação	69
Figura 36 – Teste de variâncias equivalentes com taxa de compressão de 70% do domínio História da Educação	70
Figura 37 – Teste de variâncias equivalentes com taxa de compressão de 80% do domínio História da Educação	70
Figura 38 – Teste de variâncias equivalentes com taxa de compressão de 90% do domínio História da Educação	71
Figura 39 – Teste de variâncias equivalentes com taxa de compressão de 50% do domínio Política Educacional	71

Figura 40 – Teste de variâncias equivalentes com taxa de compressão de 70% do domínio Política Educacional	72
Figura 41 – Teste de variâncias equivalentes com taxa de compressão de 80% do domínio Política Educacional	72
Figura 42 – Teste de variâncias equivalentes com taxa de compressão de 90% do domínio Política Educacional	73
Figura 43 – Teste de variâncias equivalentes com taxa de compressão de 50% do domínio Psicologia Educacional	73
Figura 44 – Teste de variâncias equivalentes com taxa de compressão de 70% do domínio Psicologia Educacional	74
Figura 45 – Teste de variâncias equivalentes com taxa de compressão de 80% do domínio Psicologia Educacional	74
Figura 46 – Teste de variâncias equivalentes com taxa de compressão de 90% do domínio Psicologia Educacional	75
Figura 47 – Teste de variâncias equivalentes com taxa de compressão de 50% do domínio Sociologia da Educação	75
Figura 48 – Teste de variâncias equivalentes com taxa de compressão de 70% do domínio Sociologia da Educação	76
Figura 49 – Teste de variâncias equivalentes com taxa de compressão de 80% do domínio Sociologia da Educação	76
Figura 50 – Teste de variâncias equivalentes com taxa de compressão de 90% do domínio Sociologia da Educação	77
Figura 51 – Teste de variâncias equivalentes com taxa de compressão de 50% do domínio Tecnologia Educacional	77
Figura 52 – Teste de variâncias equivalentes com taxa de compressão de 70% do domínio Tecnologia Educacional	78
Figura 53 – Teste de variâncias equivalentes com taxa de compressão de 80% do domínio Tecnologia Educacional	78
Figura 54 – Teste de variâncias equivalentes com taxa de compressão de 90% do domínio Tecnologia Educacional	78

LISTA DE TABELAS

Tabela 1 – Estudos relevantes na Área de Sumarização Automática.....	29
Tabela 2 – Variância do T-Student no Domínio Educação Especial.	47
Tabela 3 – Variância do T-Student no Domínio Educação Permanente	48
Tabela 4 – Variância do T-Student no Domínio Pré-Escola	48
Tabela 5 – Variância do T-Student no Domínio Ensino Aprendizagem	49
Tabela 6 – Variância do T-Student no Domínio Filosofia da Educação	49
Tabela 7 – Variância do T-Student no Domínio História da Educação	50
Tabela 8 – Variância do T-Student no Domínio Política Educacional	50
Tabela 9 – Variância do T-Student no Domínio Psicologia Educacional	51
Tabela 10 – Variância do T-Student no Domínio Sociologia da Educação	51
Tabela 11 – Variância do T-Student no Domínio Tecnologia Educacional	52
Tabela 12 – Resultados da F-Measure com o aumento da taxa de compressão do História da Educação	53

SUMÁRIO

1 INTRODUÇÃO.....	23
1.1 Motivação.....	24
1.2 Problema	24
1.3 Hipótese.....	24
1.4 Contribuições.....	24
1.5 Estrutura da Proposta	25
1.5.1 Capítulo 2 – <i>Fundamentação Teórica</i>	25
1.5.2 Capítulo 3 – <i>Metodologia</i>	25
1.5.3 Capítulo 4 – <i>Resultados</i>	25
1.5.4 Capítulo 5 – <i>Conclusão</i>	26
2 FUNDAMENTAÇÃO TEÓRICA	27
2.1 Corpus.....	27
2.2 Sumarização	27
2.2.1 <i>Sumarização Automática</i>	28
2.3 Rouge.....	29
2.3.1 <i>Métrica</i>	30
3 METODOLOGIA.....	31
3.1 Corpora	31
3.2 PragmaSUM	33
3.2.1 <i>Exemplo de sumarização no PragmaSUM</i>	33
3.2.2 <i>Métodos Utilizados na Sumarização no PragmaSUM</i>	34
3.3 Teste de Hipótese.....	34
3.3.1 <i>Teste T-Student</i>	35
4 RESULTADOS	37
4.1 Resultados Por Domínio	37

<i>4.1.1 Educação Especial</i>	37
<i>4.1.2 Domínio Educação Permanente</i>	38
<i>4.1.3 Domínio Educação Pré-Escolar</i>	39
<i>4.1.4 Domínio Ensino Aprendizagem</i>	40
<i>4.1.5 Domínio Filosofia da Educação</i>	41
<i>4.1.6 Domínio História da Educação</i>	42
<i>4.1.7 Domínio Política Educacional</i>	43
<i>4.1.8 Domínio Psicologia Educacional</i>	44
<i>4.1.9 Domínio Sociologia da Educação</i>	45
<i>4.1.10 Domínio Tecnologia Educacional</i>	46
4.2 Verificação da Hipótese	47
<i>4.2.1 Domínio Educação Especial</i>	47
<i>4.2.2 Domínio Educação Permanente</i>	48
<i>4.2.3 Domínio Educação Pré-Escolar</i>	48
<i>4.2.4 Domínio Ensino Aprendizagem</i>	49
<i>4.2.5 Domínio Filosofia da Educação</i>	49
<i>4.2.6 Domínio História da Educação</i>	50
<i>4.2.7 Domínio Política Educacional</i>	50
<i>4.2.8 Domínio Psicologia Educacional</i>	51
<i>4.2.9 Domínio Sociologia da Educação</i>	51
<i>4.2.10 Domínio Tecnologia Educacional</i>	52
5 DISCUSSÃO	53
6 CONCLUSÃO	55
6.1 Limitações	55
6.2 Trabalhos Futuros	56
REFERÊNCIAS	57
APÊNDICES	59

1 INTRODUÇÃO

Com a ampliação do acesso à internet e a criação de ferramentas que possibilitam às pessoas criarem conteúdo, a informação disponível cresce de forma acelerada. Textos sobre os mais diversos assuntos e autores são criados todos os dias. É impossível absorver a quantidade de informação disponível, o que dificulta a escolha da mais adequada para determinado interesse ou público.

A Sumarização Automática (SA) é uma subárea de estudo de Processamento de Linguagem Natural (PLN) e propõe a geração de resumos de maneira automatizada com o objetivo de reduzir o volume de informações sem perder a qualidade original do texto fonte. O sistema que realiza esta tarefa é chamado na literatura de sumarizador automático ou sumarizador.

Oliveira e Guelpeli (2014a) ressalta que atualmente é possível encontrar diversos sumarizadores automáticos, a maioria deles é voltada para o idioma inglês.

Para a realização deste trabalho foi utilizado um *corpus* com artigos no Domínio Educacional em português e divididos em dez domínios (ROCHA, 2014). A escolha por artigos da área educacional se deu por conterem resumos e palavras chaves dos seus respectivos autores. Os artigos escolhidos tinham que ter necessariamente cinco palavras chaves, para que os testes fossem realizados no PragmaSUM. Na fase captura e manipulação foram retirados dos textos gráficos, tabelas, bibliografias, notas de rodapé, figuras e números, restando somente o corpo dos textos. Para possibilitar o processamento computacional os arquivos foram convertidos dos formatos originais para o formato *.txt* (GUELPELI, 2012).

Este trabalho utiliza o PragmaSUM (GUELPELI, 2007; ROCHA & GUELPELI, 2014) que é um sumarizador automático de textos que utiliza palavras na construção do sumário. A escolha destas palavras é feita pelo usuário. Foram utilizados dois métodos para a seleção dessas palavras: no primeiro o usuário utiliza as palavras chaves que estão nos artigos originais, o segundo método o usuário utiliza as palavras com maior frequência ranqueadas pelo PragmaSUM.

Foram realizados testes no PragmaSUM que geraram sumários com quatro taxas de compressão, 50%, 70%, 80% e 90%. A partir destes testes, seus resultados foram avaliados pela métrica *F-Measure*, utilizando a ferramenta ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) (LIN E HOVY, 2003). Foram realizados testes estatísticos utilizando o teste *T-Student*, presente no Microsoft Excel (2013),

onde-se observou que os dois métodos utilizados na sumarização no PragmaSUM tiveram desempenho equivalentes.

Com este trabalho espera-se poder contribuir, principalmente, apresentando uma nova forma de SA, com um modelo que melhore a precisão dos sumários gerados. Espera-se ainda, considerando as circunstâncias propostas por este trabalho, viabilizar estudos sobre a importância da escolha das palavras-chave em artigos científicos.

1.1 Motivação

O principal objetivo deste trabalho é a análise da sumarização automática em textos na área educacional. Adotando técnicas de personificação na sumarização automática, para alcançar sumários mais exatos e de acordo com as escolhas do usuário. Contribuindo com métodos de sumarização que possam auxiliar na simplificação de textos.

1.2 Problema

Os sumarizadores automáticos disponíveis na literatura contemporânea não dispõem de métodos de personificação do sumário.

1.3 Hipótese

A hipótese deste trabalho, consiste na equivalência entre dois métodos na sumarização automáticos de textos utilizando o sumarizador PragmaSUM. Utilizando os métodos de palavras chaves dos autores e palavras com maior frequência nos textos, os sumários gerados, não têm diferenças significativas, logo eles se equivalem.

1.4 Contribuições

Com as análises dos resultados dos sumários, advindos dos experimentos realizados, gerados pelo PragmaSUM, destacam-se alguns pontos que contribuem para a área de sumarização automática de textos:

- A importância do uso de palavras chaves no processo de indexação de artigos científicos;
- Criação de um modelo que melhore a precisão do sumário;
- Criação de sumários mais personificados, de acordo com a escolha do usuário;
- Possibilitar uma estrutura algorítmica no uso do aprendizado de máquina.

1.5 Estrutura da Proposta

O trabalho desenvolvido e a estrutura estão organizados nos seguintes capítulos que serão brevemente descritos:

1.5.1 Capítulo 2 – Fundamentação Teórica

Este capítulo traz consigo a conceituação básica dos principais tópicos da pesquisa realizada, tornando-se assim um trabalho didático e de fácil entendimento. Os conceitos de *Corpus*, Sumarização e posteriormente, as métricas e sumário utilizado para a avaliação dos resultados, serão descritos do decorrer do capítulo.

1.5.2 Capítulo 3 – Metodologia

Será posto em pauta, neste capítulo, a descrição da metodologia aplicada nos testes realizados com os sumarizadores, a construção do *corpus* de domínio educacional, os domínios escolhidos, sua organização e as fontes de onde foram extraídos. Serão expostos os critérios utilizados na coleta e organização dos dados textuais e as estatísticas que constituem o *corpus*.

1.5.3 Capítulo 4 – Resultados

Em síntese, o capítulo apresenta os resultados das simulações e as comparações realizadas entre cada método. Explana tabelas e gráficos nos quais se podem conferir as modificações causadas pelas mudanças de domínio, taxas de compressão, e utilização de palavras chave do autor e palavras com maior frequência na sumarização com o PragmaSUM.

1.5.4 Capítulo 5 – Conclusão

Este capítulo mostra as contribuições e as limitações deste trabalho e apresenta sugestões para serem aplicadas em trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

A fundamentação teórica será apresentada para o entendimento dos conceitos básicos da pesquisa. São descritos conceitos de *Corpus*, Sumarização e posteriormente, a métrica e sumarizadores utilizadas para a avaliação dos resultados.

2.1 Corpus

A Linguística de *Corpus* (LC) é uma área de pesquisa contida no Processamento de Linguagem Natural (PLN). A PLN estuda os problemas da compreensão e geração automática de línguas humanas naturais, e também é uma subárea da inteligência artificial.

Segundo Oliveira e Guelpeli (2014b), a palavra *corpus* (plural, *corpora*) vem do latim *corpo*, que a LP conceitua como um conjunto de textos organizados e selecionados que seguem critérios para servir à pesquisa científica.

Segundo Fernandes e Guelpeli (2014), a criação de *corpora* baseia-se em gerar conjuntos de dados linguísticos textuais que foram coletados criteriosamente com a finalidade de abastecer a pesquisa de uma língua ou variedade linguística. Tais dados podem ser alcançados de diversas maneiras, dependendo do que se quer trabalhar, sendo as mais comuns: escanear livros, obter textos já digitalizados, e textos provenientes da internet.

2.2 Sumarização

De acordo com Pardo *et al.* (2008), sumários são textos reduzidos, que transmitem as ideias mais relevantes e principais de um texto original, de uma forma coesa e objetiva, sem que haja a perda da informatividade.

Com o aumento dos meios de comunicação e distribuição, provenientes da globalização e crescentes invenções tecnológicas, pode ser observado, nos dias de hoje, um acúmulo expressivo de textos de diversas vertentes, junto com a nossa falta de tempo, nos leva à incapacidade de consumi-los em sua totalidade. Muitas vezes, acabamos recorrendo às principais manchetes e notícias dos jornais de nossas preferências.

A quantidade de informação disponível no atual momento é de grande escala, impossível de ser absorvida em sua totalidade. Para amenizar o problema, costuma-se procurar versões menores, mais enxutas: resumos. Também chamados de sumários, são versões reduzidas dos textos a que se referem e contêm as ideias principais dos mesmos (MANI, 2001). Há diversos tipos de documentos que são exemplos de sumários: previsões meteorológicas, sinopses de novelas, chamadas de notícias jornalísticas, resenhas e resumos de livros e teses.

Hutchins (1987) classifica sumários científicos em três tipos:

- Indicativos contêm apenas os tópicos essenciais de um texto, não necessariamente contendo detalhes de resultados, argumentos e conclusões;
- Informativos, por sua vez, são considerados substitutos do texto, devendo conter todos os seus aspectos principais;
- Críticas assumem a função de avaliadores que, onde apresentam uma análise comparativa do conteúdo do texto-fonte com o contexto de outros trabalhos relacionados a esse conteúdo.

Segundo Hutchins (1987), seria mais simples produzir automaticamente sumários indicativos, devido à alta complexidade de se modelar adequadamente a sumarização humana para os demais tipos de sumários.

2.2.1 Sumarização Automática

A sumarização automática busca simular as principais características da tarefa humana: identificar informações relevantes de um texto e compô-los, para produzir os sumários correspondentes.

A sumarização automática vem sendo explorada desde a década de 50, quando começaram a surgir os primeiros métodos para a produção de extratos, sendo o método das palavras-chave (LUHN, 1958) o mais significativo então.

A Tabela 1 traz de forma cronológica as informações mais relevantes da área de sumarização automática apresentados por Guelpeli (2012).

Tabela 1 - Estudos relevantes na Área de Sumarização Automática.

Ano	Autor /Tipo de estudo
1958	Luhn apresenta o método das palavras-chaves
1958	Baxendale versa sobre a importância da primeira e da última frase do texto original.
1969	Edmundson aborda a escolha computacional das frases com o maior potencial de transmitir significância do texto original.
1975	Pollock e Zamora reforçam relevância da restrição do domínio.
1987	Hutchins classifica os sumários em indicativos, informativos e de críticas
1993	Paice e Jones hierarquizam os papéis semânticos em cada frase.
1993	Maybury propõe o uso da abordagem híbrida.
1997	Hovy e Lin fazem uso do conhecimento simbólico e de técnicas estatísticas para sumarização.
1999	Sparck afirma que a taxonomia criada por Hutchins (1987) é um fator determinante para estabelecer sua aplicabilidade e criar uma avaliação consistente deste processo.

Fonte: ROCHA; GUELPELI, 2014.

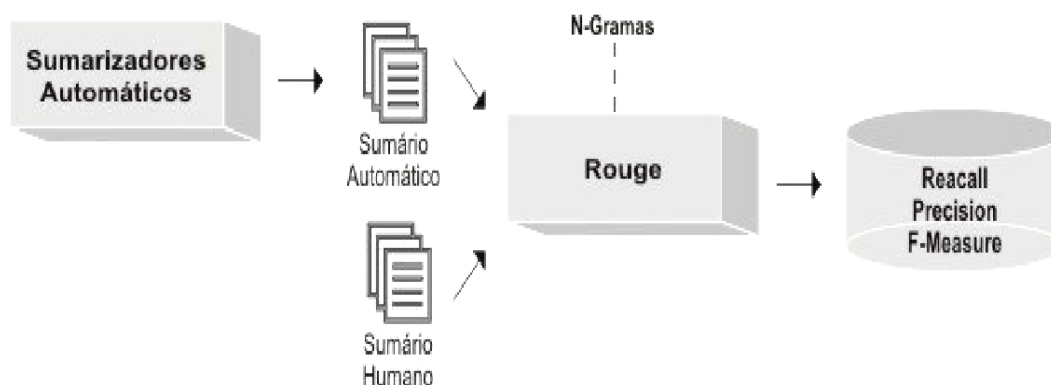
Segundo Oliveira e Guelpeli (2014a) gerar um sumário que não perca a sua informatividade é o maior desafio da área de sumarização automática.

2.3 Rouge

Segundo Oliveira e Guelpeli (2014b), a ferramenta ROUGE trata-se de um pacote de avaliação automática de sumários que compara a qualidade dos resumos gerados pelos sumarizadores automáticos com aqueles feitos por seres humanos. O uso de avaliação automatizada se justifica pela grande quantidade de textos a serem analisados e pelo alto custo que essa teria se fosse feita por especialistas.

A ROUGE, para avaliar, considera um intervalo que está entre 0 e 1 e quanto mais próximo de 1 estiver o resultado do cálculo, mais o sumário automático se assemelha ao sumário humano comparado (LUCCHI e RIBEIRO, 2011). A Figura 1 mostra o processo de avaliação com a ROUGE.

Figura 1 - Esquema do ROUGE para avaliação de sumários



Fonte: DELGADO, 2010.

2.3.1 Métrica

O ROUGE utiliza três métricas estatísticas *Recall* (cobertura), *Precision* (precisão) e *F-Measure* (média harmônica) para a avaliação dos sumários gerados. Na avaliação são comparados os resumos retirados dos artigos científicos com sumários automáticos. Segundo Oliveira e Guelpeli (2014a) essas métricas são definidas a seguir:

- *Recall* (R): indica o quanto do sumário manual permanece no sumário automático;
- *Precision* (P): indica o quanto do sumário automático coincide com o sumário manual;
- *F-Measure* (FM): indica a média harmônica entre a cobertura e a precisão.

3 METODOLOGIA

Este capítulo descreve a metodologia utilizada para os testes realizados com os sumarizadores, a construção do *corpus* de domínio educacional, os domínios escolhidos, sua organização e as fontes de onde foram extraídos. Serão apresentados os critérios utilizados na coleta e organização dos dados textuais e as estatísticas que compõem o *corpus*.

3.1 Corpora

O *corpus* utilizado (ROCHA; GUELPELI, 2014) neste trabalho tem um total de 50 artigos científicos em português, separados em 10 áreas de conhecimento do domínio Educacional, cada um contendo 5 artigos. Todos estes artigos foram retirados da tabela de áreas de conhecimento da grande área da educação, disponível no site da CAPES. Foram escolhidas as seguintes categorias: Educação Especial, Educação Permanente, Educação Pré-Escolar, Ensino Aprendizagem, Filosofia da Educação, História da Educação, Política Educacional, Psicologia Educacional, Sociologia da Educação e Tecnologia Educacional.

Foram escolhidos artigos científicos por eles conterem resumos dos seus autores, que foram utilizados na avaliação do ROUGE, e também foi utilizado o PragmaSUM para fazer a sumarização de textos usando as palavras chaves de cada artigo.

Tabelas, gráficos, figuras, notas de rodapé, bibliografias e números de páginas foram retiradas dos textos na fase de captura e manipulação, restando somente o corpo do texto. Os arquivos foram convertidos dos formatos originais para o formato *txt* a fim de possibilitar o processamento computacional (GUELPELI, 2012).

A Figura 2 mostra a organização dos arquivos, na pasta raiz *Corpus* existem 10 pastas de cada domínio de conhecimento.

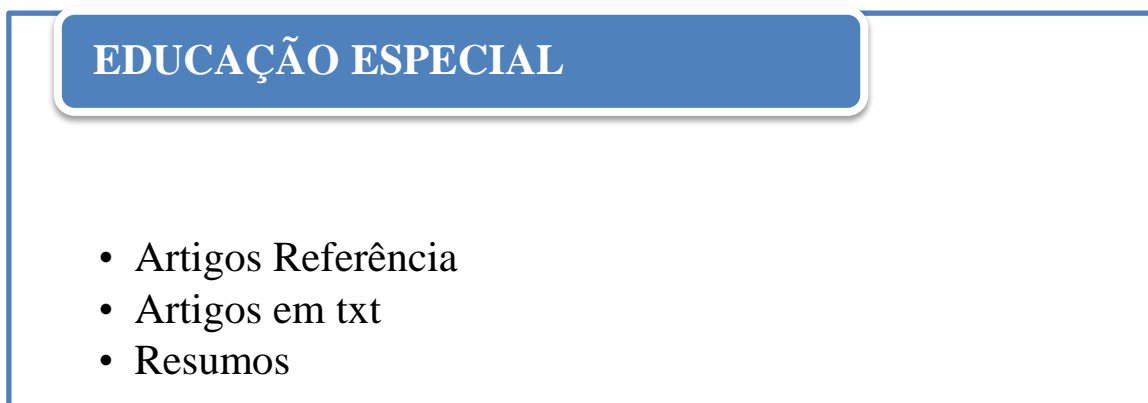
Figura 2 - Pasta Raiz do Corpus.

Fonte: O AUTOR, 2019.

Cada domínio contém três diretórios organizados da seguinte maneira:

- Artigos Referências: que contém os artigos originais em formato pdf;
- Artigos em txt: que trazem os arquivos em .txt para serem processados;
- Resumos: que contém arquivos com os resumos produzidos pelos autores de cada artigo.

A Figura 3 apresenta o diagrama simplificado do domínio Educação Especial.

Figura 3 - Diagrama do Domínio Educação Especial

Fonte: O AUTOR, 2019.

O *corpus* possui 150 arquivos de texto, sendo 15 de cada domínio de conhecimento. Cada domínio possui 5 artigos referências, 5 artigos em txt e 5 resumos.

3.2 PragmaSUM

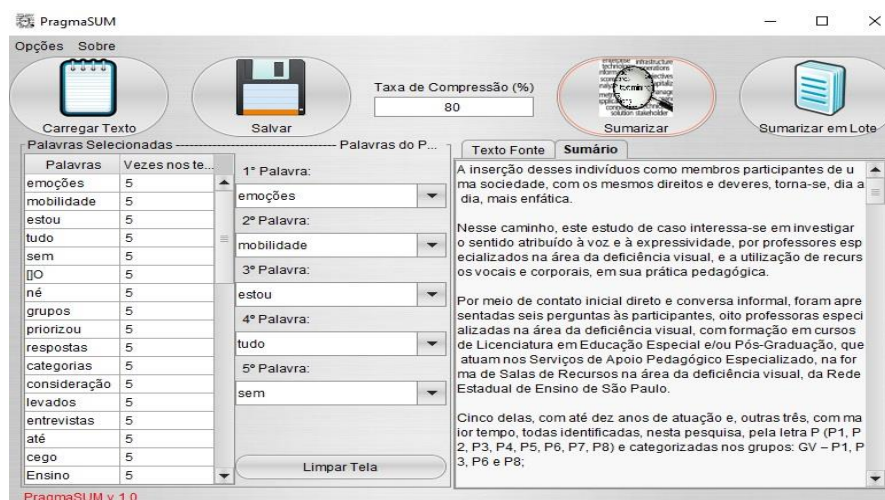
Desenvolvido por Rocha & Guelpli (2014), o PragmaSUM é um sumarizador automático de textos independentemente do idioma e domínio. O PragmaSUM permite personificar a sumarização individualmente ou sumarizar vários textos de uma só vez com os mesmos parâmetros.

Na etapa de pré-processamento ocorre a limpeza dos textos, a preparação para o processo computacional, a redução do número de palavras é a maior preocupação, não só para a viabilização computacional, mas também para proporcionar um ganho qualitativo e quantitativo para o processamento (GUELPELI, 2012).

3.2.1 Exemplo de sumarização no PragmaSUM

Nessa seção será apresentado um exemplo da interação do usuário com o PragmaSUM na sumarização de textos. A Figura 7 apresenta um sumário gerado pelo PragmaSUM com a taxa de compressão de 80%. A figura mostra as palavras selecionadas, há a opção de escolher cinco delas ou inserir manualmente as palavras para sumarização personificada. E por fim apresenta o sumário gerado.

Figura 4 - Taxa de 80% de Compressão no PragmaSUM.



Fonte: O AUTOR, 2019.

3.2.2 Métodos Utilizados na Sumarização no PragmaSUM

Foram utilizados dois métodos para fazer a sumarização automática utilizando o PragmaSUM (GUELPELI, 2007; ROCHA & GUELPELI, 2014) que é um sumarizador automático de textos que utiliza palavras na construção do sumário. A escolha destas palavras é feita pelo usuário.

São dois os principais métodos utilizados pelo PragmaSUM. No primeiro método o usuário utiliza as palavras chaves que estão nos artigos originais. Já no segundo método, o usuário utiliza as palavras com maior frequência ranqueadas no PragmaSUM.

3.3 Teste de Hipótese

Teste de hipótese é uma metodologia estatística que nos auxilia a tomar decisões sobre uma ou mais populações baseado na informação obtida da amostra. Nos permite verificar se os dados amostrais trazem evidência que apoiem ou não uma hipótese estatística formulada.

Um teste de hipótese é uma regra que especifica se deve aceitar ou rejeitar uma alegação sobre uma população de acordo com as provas fornecidas por uma amostra de dados.

Segundo Guelpeleli (2012) um teste de hipótese examina duas hipóteses opostas sobre uma população: a hipótese nula e a hipótese alternativa. A hipótese nula é a declaração que está sendo testada. Normalmente, a hipótese nula é uma declaração de "nenhum efeito" ou "nenhuma diferença". A hipótese alternativa é a declaração que você quer ser capaz de concluir que é verdadeira com base em evidências fornecidas pelos dados da amostra.

Com base nos dados amostrais, o teste determina se devemos rejeitar a hipótese nula. Você usa um valor-p para fazer a determinação. Se o valor de p for menor que o nível de significância (denotado como α ou alfa), então você pode rejeitar a hipótese nula.

3.3.1 *Teste T-Student*

O teste *T-Student* pode ser utilizado para avaliação das médias de duas amostras, com o objetivo de verificar se há uma diferença significativa entre suas médias. Por ser simples e sua implementação em software, é bastante utilizado nos testes de hipótese.

As hipóteses para o teste com duas amostras podem ser:

- Nula: quando a diferença entre as médias da população for igual a diferença hipotética;
- Alternativas: quando a diferença entre as médias da população não for igual a diferença hipotética.

Para Raju (2005), o T-Student pode ser utilizado, quando:

- Tamanhos iguais e variâncias iguais: as amostras dos dois grupos possuem tamanho iguais;
- Tamanhos diferentes e variâncias iguais: quando duas amostras possuem a mesma variância;

Tamanhos diferentes e variâncias diferentes: neste caso o teste é utilizado para amostras que possuem variâncias diferentes.

4 RESULTADOS

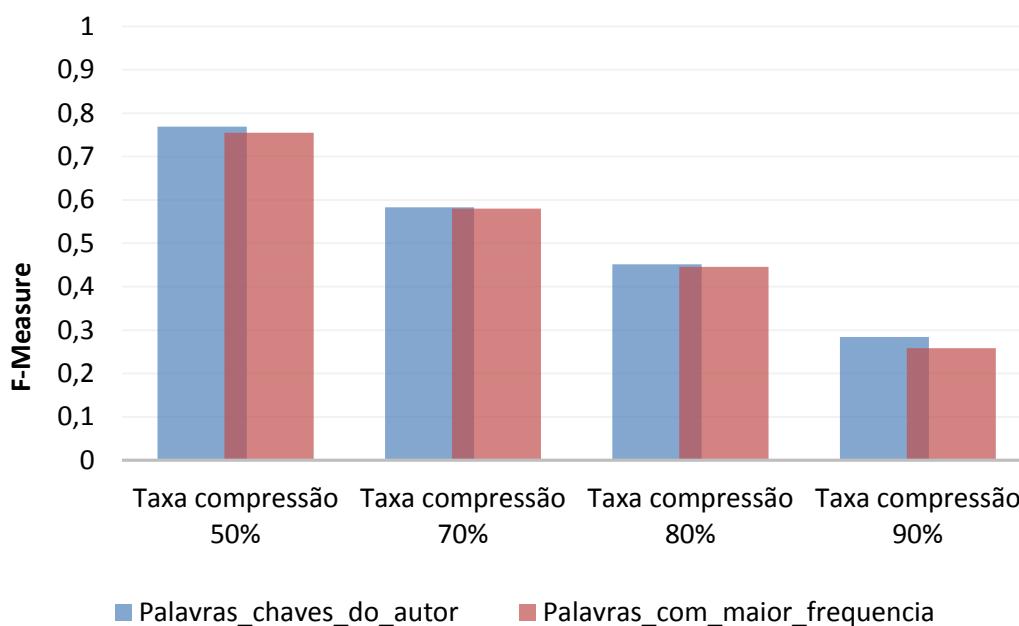
Resultados das simulações feitas por cada domínio educacional e subdivido em quatro taxas de compressão (50, 70, 80 e 90), utilizando os métodos de palavras chaves do autor e palavras com maior frequência. Foram geradas tabelas e gráficos utilizando a métrica *F-Measure* (indica a média harmônica entre a cobertura e a precisão) para a avaliação dos sumários gerados. Os resultados variam em uma escala de 0 a 1, em que: valores mais próximo de 1 indicam melhores resultados e valores mais próximo de 0 indicam piores resultados.

4.1 Resultados Por Domínio

4.1.1 Educação Especial

A Figura 5 apresenta os resultados da comparação feita entre as taxas de compressão no domínio Educação Especial utilizando a métrica *F-Measure*. Nota-se que com a taxa de compressão de 90%, o método de palavras chaves do autor teve um desempenho ligeiramente melhor que o método de palavras com maior frequência, já nas outras taxas os desempenhos entre os métodos foram equivalentes.

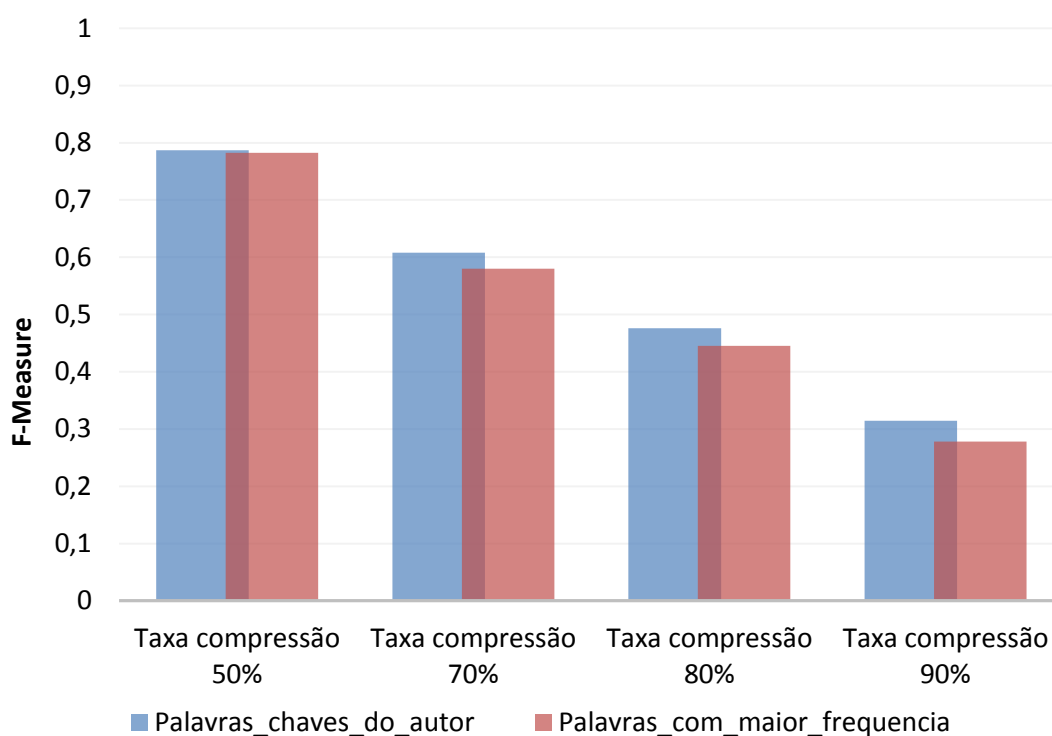
Figura 5 - Comparação entre *F-Measure* no Domínio Educação Especial.



4.1.2 Domínio Educação Permanente

A Figura 6 mostra a comparação entre o desempenho dos métodos de palavras chaves do autor e palavras com maior frequência com as taxas de compressão 50, 70, 80 e 90 no domínio Educação Permanente, utilizando a métrica *F-Measure*. Nota-se que os desempenhos entre os métodos foram equivalentes, com uma ligeira vantagem para método de palavras chaves do autor nas taxas de 70, 80 e 90.

Figura 6 - Comparação entre *F-Measure* no Domínio Educação Permanente.

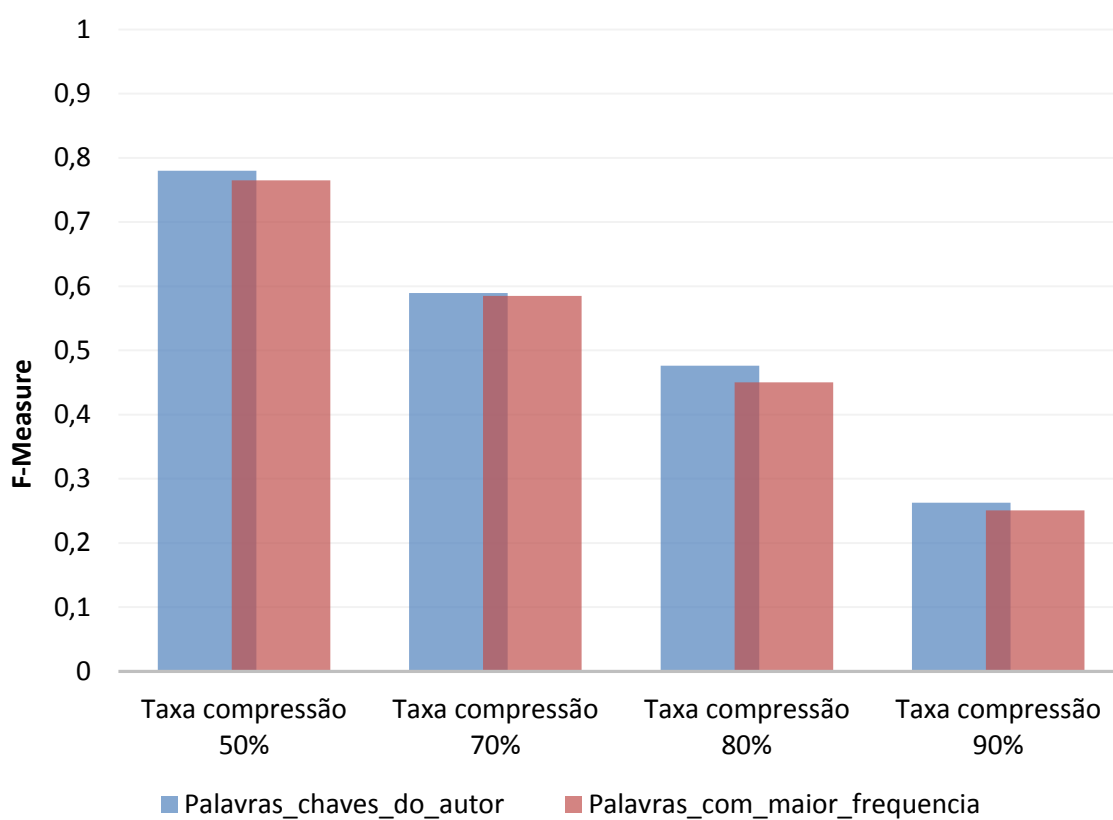


Fonte: O AUTOR, 2019.

4.1.3 Domínio Educação Pré-Escolar

A Figura 7 mostra a comparação feita entre as taxas de compressão 50, 70, 80 e 90 no domínio Educação Pré-Escolar, utilizando a métrica *F-Measure*. Nota-se que os desempenhos entre os métodos de palavras chaves do autor e palavras com maior frequência tiveram resultados equivalentes, mas com uma ligeira vantagem na taxa de compressão de 80 para o método de palavras chaves do autor.

Figura 7 - Comparação entre *F-Measure* no Domínio Educação Pré-Escolar.

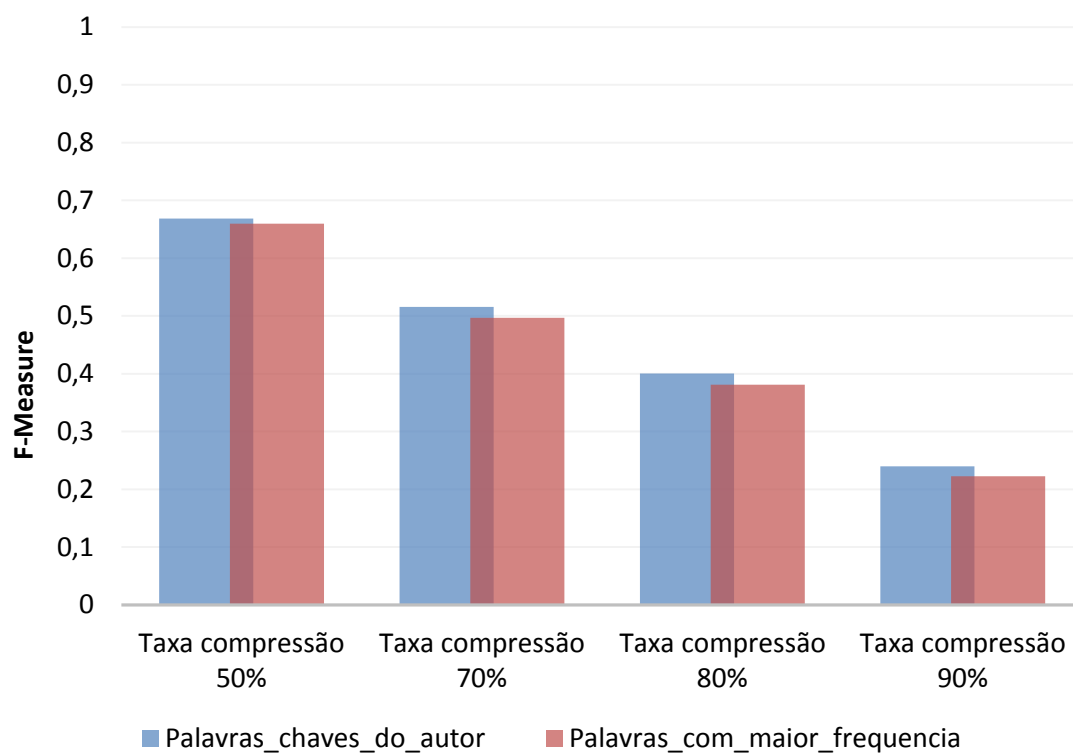


Fonte: O AUTOR, 2019.

4.1.4 Domínio Ensino Aprendizagem

A Figura 8 apresenta os resultados da comparação feita entre os métodos de palavras chaves e palavras com maior frequência no domínio Ensino Aprendizagem com as taxas de compressão 50, 70, 80 e 90, utilizando a métrica *F-Measure*. Nota-se que os desempenhos entre os métodos foram equivalentes, mas com o método de palavras chaves do autor com uma ligeira vantagem nas taxas de 70, 80 e 90.

Figura 8 - Comparação entre *F-Measure* no Domínio Ensino Aprendizagem.

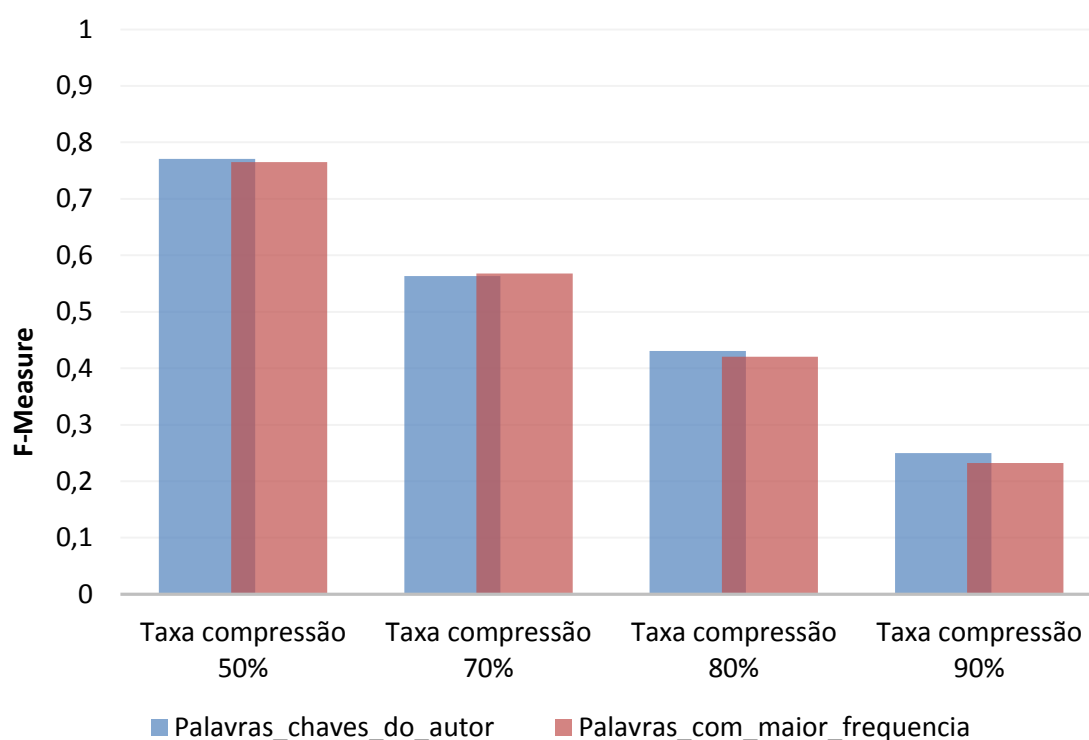


Fonte: O AUTOR, 2019.

4.1.5 Domínio Filosofia da Educação

A Figura 9 mostra uma comparação feita entre as taxas de compressão 50, 70, 80 e 90 no domínio Filosofia da Educação, utilizando a métrica *F-Measure*. Nota-se que o desempenho entre os métodos de palavras chaves do autor e palavras com maior frequência tiveram resultados equivalentes, com uma pequena vantagem na taxa de 90 para o método de palavras chaves do autor.

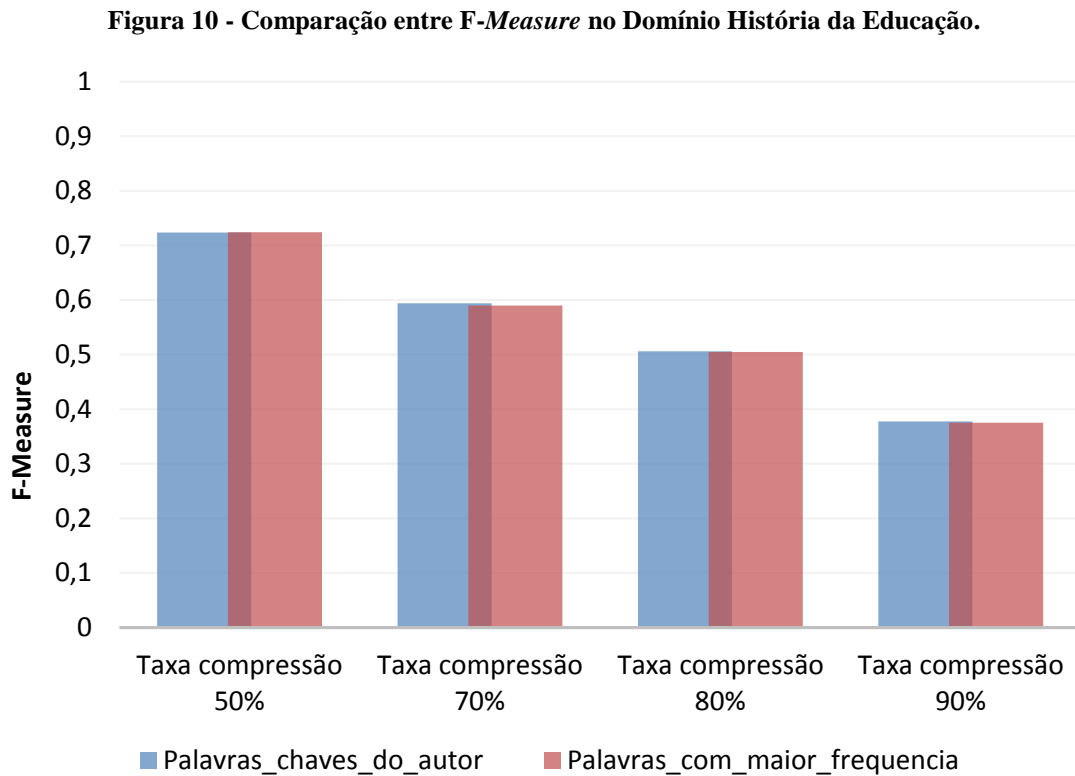
Figura 9 - Comparação entre *F-Measure* no Domínio Filosofia da Educação.



Fonte: O AUTOR, 2019.

4.1.6 Domínio História da Educação

A Figura 10 mostra que os métodos de palavras chaves do autor e palavras com maior frequência tiveram desempenhos equivalentes na comparação feita entre as taxas de compressão 50, 70, 80 e 90 no domínio História da Educação, utilizando a métrica *F-Measure*.

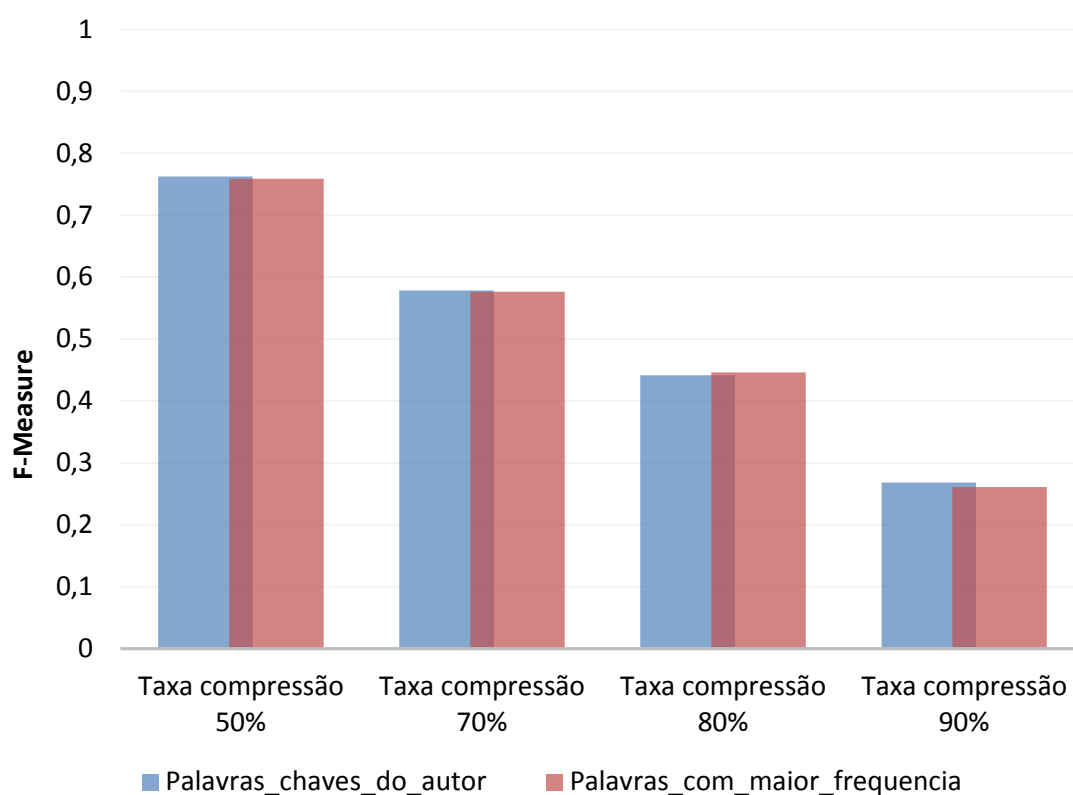


Fonte: O AUTOR, 2019.

4.1.7 Domínio Política Educacional

A Figura 11 mostra uma comparação feita entre as taxas de compressão 50, 70, 80 e 90 no domínio Política Educacional, utilizando a métrica *F-Measure*, nota que o desempenho entre os métodos de palavras chaves do autor e palavras com maior frequência tiveram resultados equivalentes.

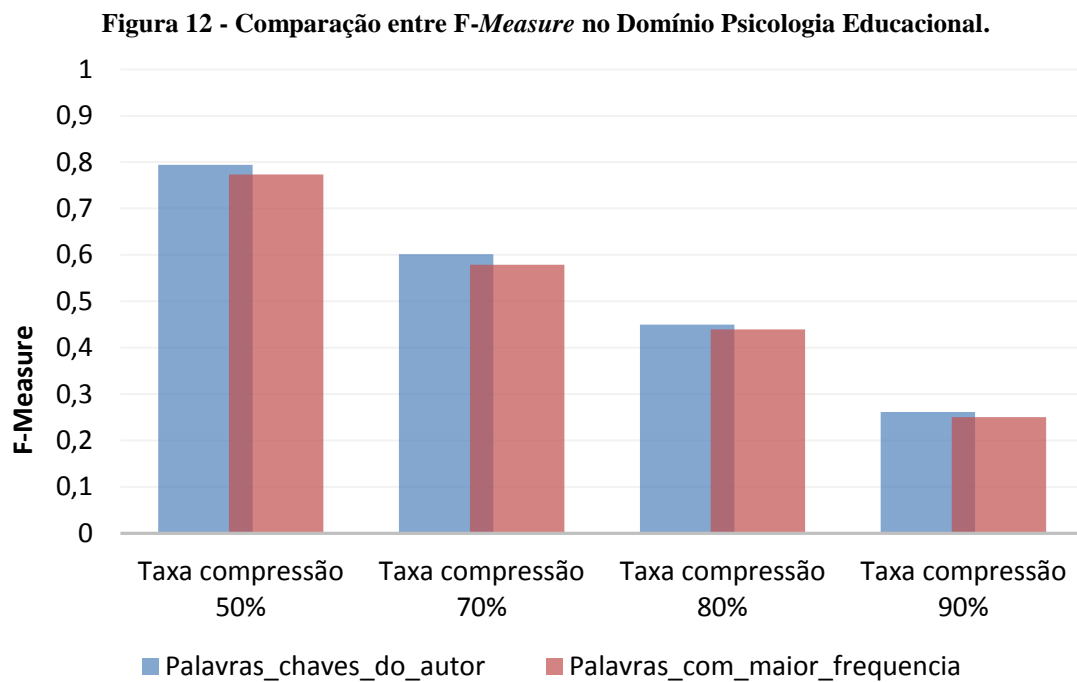
Figura 11 - Comparação entre *F-Measure* no Domínio Política Educacional.



Fonte: O AUTOR, 2019.

4.1.8 Domínio Psicologia Educacional

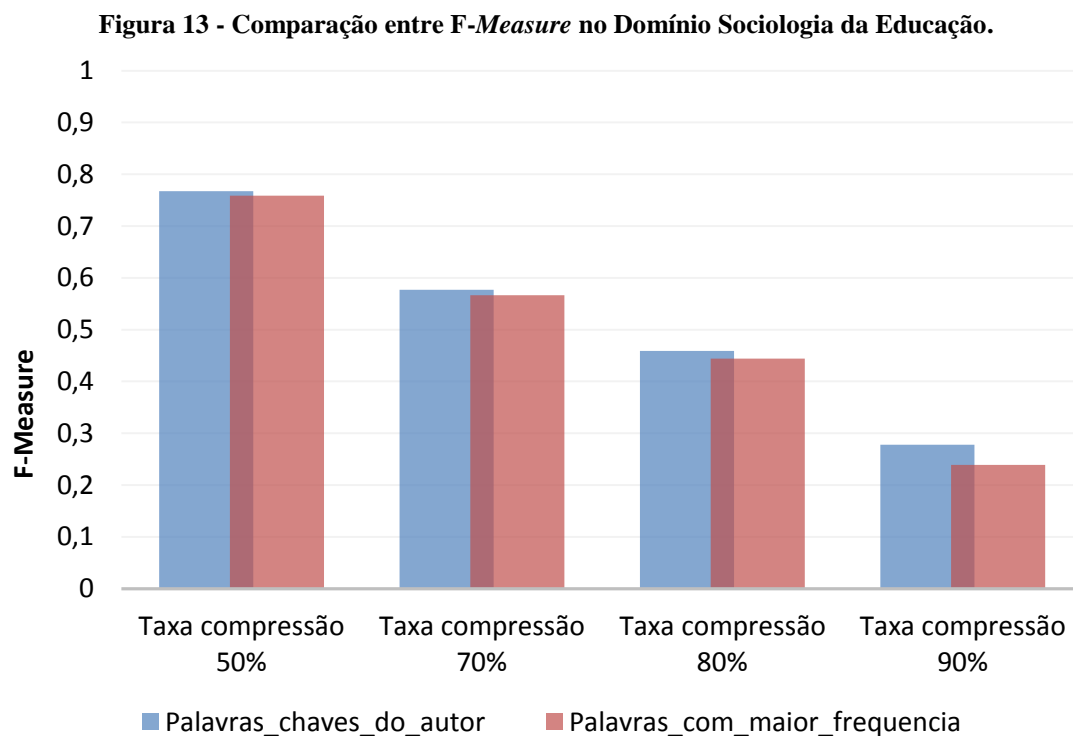
A Figura 12 apresenta a comparação entre os métodos de palavras chaves do autor e palavras com maior frequência com as taxas de compressão 50, 70, 80 e 90 no domínio Psicologia Educacional, utilizando a métrica *F-Measure*. Nota-se a equivalência no desempenho entre os métodos, com uma ligeira vantagem nas taxas de 50 e 70 para o método de palavras chaves do autor.



Fonte: O AUTOR, 2019.

4.1.9 Domínio Sociologia da Educação

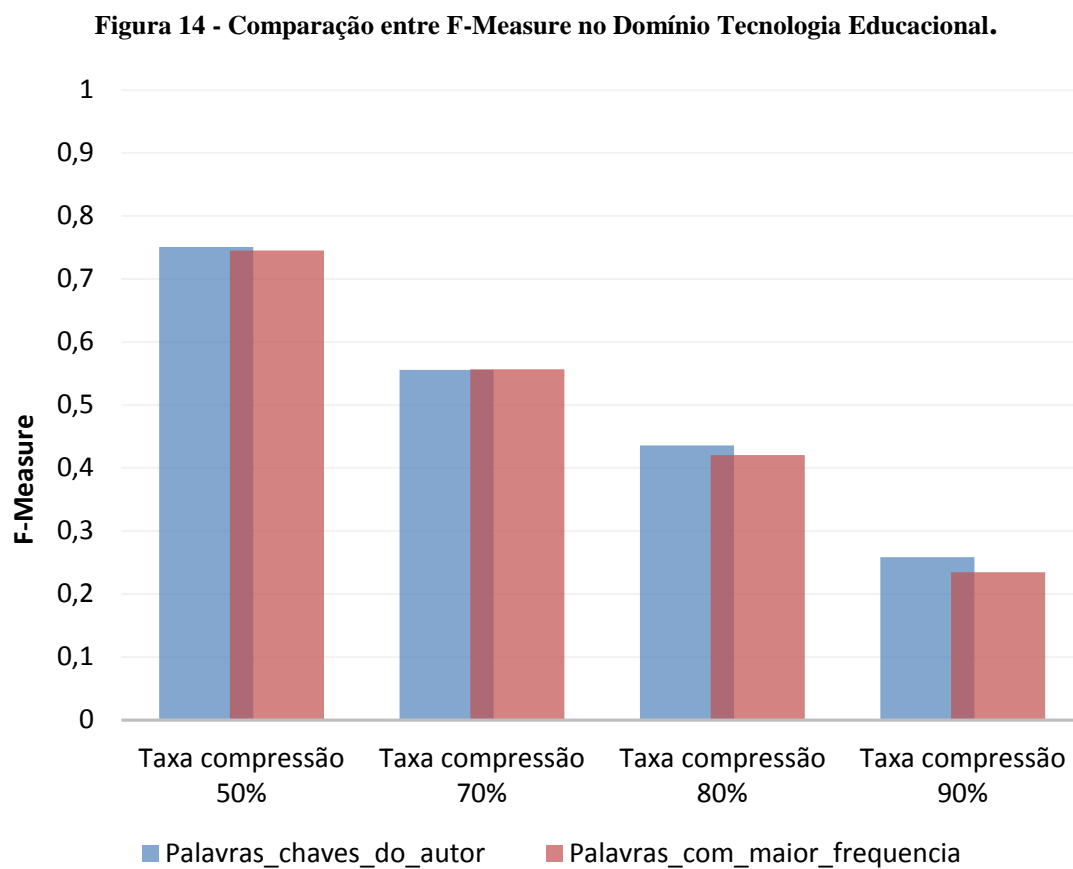
A Figura 13 mostra uma comparação feita entre as taxas de compressão 50, 70, 80 e 90 no domínio Sociologia da Educação utilizando a métrica *F-Measure*. Nota-se que o desempenho entre os métodos de palavras chaves do autor e palavras com maior frequência nos textos tiveram resultados equivalentes.



Fonte: O AUTOR, 2019.

4.1.10 Domínio Tecnologia Educacional

A Figura 14 apresenta a comparação no domínio Tecnologia Educacional entre os métodos de palavras chaves do autor e palavras com maior frequência com as taxas de compressão 50, 70, 80 e 90, utilizando a métrica *F-Measure*. Nota-se que os desempenhos entre os métodos tiveram resultados equivalentes, com uma pequena vantagem para o método de palavras chaves do autor nas taxas de 80 e 90.



Fonte: O AUTOR, 2019.

4.2 Verificação da Hipótese

Esta seção tem como objetivo discutir a hipótese deste trabalho. Com os resultados obtidos pelo PragmaSUM, a utilização dos métodos de sumarização com palavras chaves do autor e palavras com maior frequência tiveram resultados equivalentes, o que comprava a hipótese deste trabalho.

Para a comprovação da hipótese foi utilizado o teste estatístico *T-Student*, presente na ferramenta Microsoft Excel (2013), que é um teste de hipótese utilizado para avaliar se há diferenças significativas entre as médias de duas amostras. Para a realização dos testes, foram analisadas duas amostras: palavras chaves do autor e palavras com maior frequência. Notou-se que não houve diferenças significativas como apresentado nos resultados dos gráficos da seção 4.1, logo utilizou-se o teste *T-Student* para verificar se a hipótese do trabalho é aceita.

A seguir, serão apresentados os resultados dos testes estatísticos para cada domínio.

4.2.1 Domínio Educação Especial

A Tabela 2 mostra os resultados dos testes estatísticos *T-Student* no Domínio Educação Especial para cada taxa de compressão. Para realizar os testes estatísticos, foram utilizadas as amostras obtidas através da sumarização automática no PragmaSUM. Notou-se que na taxa de compressão de 90%, a variância teve o maior índice entre os métodos, mas a diferença não foi significativa, logo os métodos de sumarização se equivalem e comprovam a hipótese deste trabalho.

Tabela 2 - Variância do *T-Student* no Domínio Educação Especial.

Taxa de compressão	Palavras Chaves do Autor	Palavras com maior frequência
50	0,006130	0,004677
70	0,004300	0,003780
80	0,002492	0,001467
90	0,002965	0,000975

4.2.2 Domínio Educação Permanente

A Tabela 3 apresenta os resultados obtidos pelos métodos de sumarização no PragmaSUM do Domínio Educação Permanente, através do teste estatístico *T-Student* com as taxas de compressão. Pode-se notar que os métodos tiveram variância equivalentes, logo comprovam a hipótese deste trabalho.

Tabela 3 - Variância do *T-Student* no Domínio Educação Permanente.

Taxa de compressão	Palavras Chaves do Autor	Palavras com maior frequência
50	0,002001	0,001505
70	0,003929	0,002235
80	0,006634	0,003572
90	0,008269	0,006065

Fonte: O AUTOR, 2019.

4.2.3 Domínio Educação Pré-Escolar

A Tabela 4 mostra os resultados da variância entre os métodos de sumarização utilizados no PragmaSUM do Domínio Educação Pré-Escolar. Para a realização dos testes foi utilizado o teste *T-Student*. Notou-se na taxa de compressão de 70%, a variância teve o maior índice, mas a diferença não foi significativa, logo os métodos de sumarização se equivalem.

Tabela 4 - Variância do *T-Student* no Domínio Pré-Escola.

Taxa de compressão	Palavras Chaves do Autor	Palavras com maior frequência
50	0,000363	0,000461
70	0,000205	0,000537
80	0,000762	0,000542
90	0,000366	0,000421

Fonte: O AUTOR, 2019.

4.2.4 Domínio Ensino Aprendizagem

A Tabela 5 mostra os resultados dos testes estatísticos do Domínio Ensino Aprendizagem para cada taxa de compressão na sumarização automática utilizando o PragmaSUM. Pode-se notar que a variância entre os métodos não houve diferenças significativas, logo os métodos de sumarização se equivalem.

Tabela 5 - Variância do *T-Student* no Domínio Ensino Aprendizagem.

Taxa de compressão	Palavras Chaves do Autor	Palavras com maior frequência
50	0,073067	0,068168
70	0,054731	0,046059
80	0,031513	0,027833
90	0,012992	0,010968

Fonte: O AUTOR, 2019.

4.2.5 Domínio Filosofia da Educação

A Tabela 6 mostra os resultados dos testes estatísticos do Domínio Filosofia da Educação para cada taxa de compressão. Para cálculo da variância foi utilizado o teste estatísticos *T-Student*, pode-se notar que a variância entre os métodos de palavras chaves do autor e palavras com maior frequência não foi significativa, por tanto os métodos de sumarização se equivalem.

Tabela 6 - Variância do *T-Student* no Domínio Filosofia da Educação.

Taxa de compressão	Palavras Chaves do Autor	Palavras com maior frequência
50	0,000181	0,000355
70	0,000654	0,000187
80	0,000750	0,000547
90	0,000367	0,000279

Fonte: O AUTOR, 2019.

4.2.6 Domínio História da Educação

A Tabela 7 apresenta os resultados dos testes estatísticos do Domínio História da Educação para cada taxa de compressão na sumarização utilizando o PragmaSUM. Pode-se notar que a variância entre os métodos não foi significativa, logo os métodos de sumarização se equivalem.

Tabela 7 - Variância do *T-Student* no Domínio História da Educação.

Taxa de compressão	Palavras Chaves do Autor	Palavras com maior frequência
50	0,012954	0,013144
70	0,012977	0,012130
80	0,012532	0,013844
90	0,020465	0,022903

Fonte: O AUTOR, 2019.

4.2.7 Domínio Política Educacional

A Tabela 8 mostra os resultados dos testes estatísticos *T-Student* no Domínio Política Educacional para cada taxa de compressão. Para realizar os testes estatísticos, foram utilizadas as amostras obtidas através da sumarização automática no PragmaSUM. Notou-se que na taxa de compressão de 90%, a variância teve o maior índice entre os métodos, mas a diferença não foi significativa, logo os métodos de sumarização se equivalem e comprovam a hipótese deste trabalho.

Tabela 8 - Variância do *T-Student* no Domínio Política Educacional

Taxa de compressão	Palavras Chaves do Autor	Palavras com maior frequência
50	0,000710	0,000912
70	0,001596	0,001272
80	0,002495	0,001070
90	0,001209	0,000483

Fonte: O AUTOR, 2019.

4.2.8 Domínio Psicologia Educacional

A Tabela 9 mostra os resultados dos testes estatísticos do Domínio Psicologia Educacional para cada taxa de compressão. Para cálculo da variância foi utilizado o teste estatísticos *T-Student*, pode-se notar que a variância entre os métodos de palavras chaves do autor e palavras com maior frequência não foi significativa, por tanto os métodos de sumarização se equivalem.

Tabela 9 - Variância do *T-Student* no Domínio Psicologia Educacional.

Taxa de compressão	Palavras Chaves do Autor	Palavras com maior frequência
50	0,001037	0,002175
70	0,001507	0,004468
80	0,001842	0,003708
90	0,000872	0,003191

Fonte: O AUTOR, 2019.

4.2.9 Domínio Sociologia da Educação

A Tabela 10 apresenta os resultados obtidos pelos métodos de sumarização no PragmaSUM do Domínio Sociologia da Educação, através do teste estatístico *T-Student* para cada taxa de compressão. Notou-se que na taxa de compressão de 90% a variância teve o maior índice entre os métodos, mas a diferença não foi significativa, logo comprovam a hipótese deste trabalho.

Tabela 10 - Variância do *T-Student* no Domínio Sociologia da Educação.

Taxa de compressão	Palavras Chaves do Autor	Palavras com maior frequência
50	0,000668	0,001751
70	0,001229	0,001836
80	0,002115	0,001785
90	0,000777	0,001771

Fonte: O AUTOR, 2019.

4.2.10 Domínio Tecnologia Educacional

A Tabela 11 apresenta os resultados dos testes estatísticos do Domínio Tecnologia Educacional para cada taxa de compressão na sumarização utilizando o PragmaSUM. Pode-se notar que na taxa de compressão de 80 % variância entre os métodos teve o maior índice, mas a diferença não foi significativa, logo os métodos de sumarização se equivalem.

Tabela 11 - Variância do *T-Student* no Domínio Tecnologia Educacional.

Taxa de compressão	Palavras Chaves do Autor	Palavras com maior frequência
50	0,000659	0,001105
70	0,001781	0,001304
80	0,001338	0,000510
90	0,000755	0,000909

Fonte: O AUTOR, 2019.

5 DISCUSSÃO

Nessa seção serão discutidos os resultados com um todo, da análise da métrica *F-Measure*.

Notou-se que os resultados obtidos pelos dois métodos de sumarização com o PragmaSUM foram equivalentes, como apresentado nos gráficos do capítulo 4. Observou-se também uma queda no desempenho dos métodos de palavras chaves do autor e palavras com maior frequência, com o aumento da taxa de compressão. A tabela a seguir apresenta os resultados do Domínio História da Educação.

Tabela 12 - Resultados da *F-Measure* com o aumento da taxa de compressão do História da Educação.

Taxa de compressão	Palavras Chaves do Autor	Palavras com maior frequência
50	0,723334	0,724140
70	0,593850	0,589784
80	0,506014	0,505028
90	0,377726	0,375006

Fonte: O AUTOR, 2019.

Analisando os resultados gerados pelos gráficos, o teste estatístico obteve valores equivalentes, como apresentado na seção 4.1. Pode-se concluir então que a hipótese de equivalência entre os métodos de sumarização foi aceita.

6 CONCLUSÃO

Este trabalho teve como objetivo principal avaliar o desempenho de dois métodos na sumarização automática utilizando o PragmaSUM. Foi utilizado um *corpus* na área da educação, formado por 50 artigos científicos da língua portuguesa e divididos em dez domínio educacionais.

Foram realizados testes que geraram sumários com quatro taxas de compressão, 50%, 70%, 80% e 90%. A partir destes testes, seus resultados foram avaliados pela métrica *F-Measure*, utilizando a ferramenta ROUGE.

Nos resultados dos testes estatísticos neste trabalho, pode-se observar que os dois métodos utilizados na sumarização no PragmaSUM, palavras chaves do autor e palavras com maior frequência, tiveram desempenho equivalentes, o que comprova a hipótese deste trabalho.

Conclui-se então que, pelo fato dos autores dos artigos deste *corpus* não conhecerem o processo de indexação não souberam individualizar os artigos, ou seja, não souberam identificar as palavras chaves que melhor representam seus textos. Percebe-se então a importância da escolha das palavras chaves pelos autores, fazendo com que os resultados das buscas feitas pelos usuários sejam intuitivas e apresentem os documentos que realmente precisam. A indexação tem como benefícios as buscas intuitivas, rápidas e otimizadas.

A partir dessas considerações, serão apresentadas as limitações e possíveis trabalhos futuros.

6.1 Limitações

Neste trabalho foi utilizado um *corpus* com apenas textos em português, que contém palavras chaves e resumos retirados de artigos científicos. Um fator limitante é a quantidade de domínios, seria interessante o seu aumento. Outro fator seria a necessidade de testar em outros idiomas e em outras áreas.

6.2 Trabalhos Futuros

Utilizar um *corpus* em outras áreas, computacional, saúde, exatas, humanas e entre outras. Utilização do *corpus* em mais domínios e em diferentes idiomas. Utilizar outros métodos de personificação do sumário, criar novas formas de selecionar palavras-chaves.

REFERÊNCIAS

DÁVILA, VÍCTOR HUGO LANCHOS. **Teste de Hipótese**. Universidade Estadual de Campinas (UNICAMP).

DELGADO, Carlos Henrique; VIANNA, Caroline Evangelista; GUELPELI, Marcus Vinicius C. Comparando sumários de referência humano com extratos ideais no processo de avaliação de sumários extrativos. **IADIS Ibero-Americana WWW/Internet**, v. 1, n. 1, p. 293-300, 2010.

EXCEL, M.O. **Programa de planilha eletrônica**. 2013.

ESPINA, A.P.; RINO, L.H.M.; **Utilização de Métodos Extrativos na Sumarização Automática de Textos**; Série de Relatórios do NILC. NILC-TR-02-06, São Carlos, Março, 2002;

FERNANDES H. M.; GUELPELI, M. V. C. **Creación de corpus en lengua española para su utilización en testes acerca de Sumarización Automática**. In: 6th International Conference on Corpus Linguistics-CILC 2014, 2014, Las Palmas de Gran Canaria. 6th International Conference on Corpus Linguistics-CILC 2014, 2014.

GUELPELI, M.V.C; BERNARDINI, F.C; GARCIA, A.C.B; **Todas as Palavras da Sentença como Métrica para um sumarizador Automático**. In: Tecnologia da Informação e da Linguagem Humana-TIL, WebMedia, 2008. P. 287-291, Vila Velha, Brasil, 2008.

GUELPELI, M.V.C; **Cassiopeia: Um modelo de agrupamento de textos baseado em sumarização**. 2012. 220f. Tese (Doutorado em Computação) – Universidade Federal Fluminense, Rio de Janeiro, 2012.

LEITE, D.S; RINO, L.H.M; **Uma comparação entre sistemas de sumarização automática extrativa**; Departamento de Computação, UFSCar CP 676, 13565-905, São Carlos, Outubro, 2006;

LIN, C-Y. and HOVY, E.H. Automatic Evaluation of Summaries Using N-gram Co occurrence Statistics. **In the Proceedings of Language Technology Conference – HLT**. Edmonton, Canadá, 2003.

LUCHI, D.; RIBEIRO, E. **Sumarização Automática de Textos via Ranqueamento de Sentenças**. 2011. Universidade Federal do Espírito Santo – UFES.

LUHN, H. P. The Automatic Creation of Literature Abstracts. **IBM Journal of Research and Development**, vol. 2, 157-165, 1958.

MARTINS, C.B.; PARDO, T.A.S.; ESPINA, A.P.; RINO, L.H.M. (2001). **Introdução à Sumarização Automática**. Relatório Técnico RT-DC 002/2001. Departamento de Computação, Universidade Federal de São Carlos. São Carlos-SP, Fevereiro, 38p.

NOGUEIRA, B.M; **Avaliação de métodos não-supervisionados de seleção de atributos para Mineração de Textos**. 2009. 104f. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) - Universidade de São Paulo, São Carlos, 2009.

OLIVEIRA, R.R.; GUELPELI, M.V.C. Building a Corpus in Italian Written Language. In: **6th International Conference on Corpus Linguistics (CILC2014)**. Las Palmas de Gran Canaria, Espanha, 2014a. No prelo.

OLIVEIRA, R.R.; GUELPELI, M.V.C. Corpus in Italian of the Journalism and Medical Fields. In: **Second Asia Pacific Corpus Linguistics Conference (APCLC 2014)**, Abstract. Hong Kong, China, 2014b. No prelo.

PARDO, T. A. S. *et al.* Sumarização automática: principais conceitos e sistemas para o português brasileiro. **Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional**. NILC-TR-06-01, São Carlos, SP, 2008.

RAJU, Tonse NK; William Sealy Gosset and William A. Silverman: two “students” of science. **Pediatrics**, v. 116, n. 3, p. 732-735, 2005.

ROCHA, V. J. C.; GUELPELI, M. V. C. **Pragmasum**: um sumarizador automático de textos baseado em perfil de usuário. Trabalho de Conclusão de Curso de Sistemas de Informação da Universidade Federal dos Vales do Jequitinhonha e Mucuri. Diamantina-MG, Brasil, 2014.

ROCHA, V. J. C. **Pragmasum**: novos métodos na utilização de palavras-chave na sumarização automática. Dissertação (Mestrado Profissional - Programa de PósGraduação em Educação) –Universidade Federal dos Vales do Jequitinhonha e Mucuri, 2014.

APÊNDICES

APÊNDICE A – GRÁFICOS COM OS RESULTADOS DO TESTE ESTATÍSTICO *T-STUDENT*

O apêndice A traz os resultados dos testes estatísticos realizados num corpus de textos em português com textos de 10 domínios de conhecimento.

1 DOMÍNIO EDUCAÇÃO ESPECIAL

Esta seção descreve os resultados do teste *T-Student* no domínio Educação Especial. As Figuras 15, 16, 17 e 18 trazem dos resultados dos testes estatísticos para os dois métodos da sumarização automática utilizada no PragmaSUM em quatro taxas de compressão.

A Figura 15 (teste de variâncias equivalentes) apresenta os resultados dos testes estatísticos para os métodos de palavras chaves do autor e palavras com maior frequência com a taxa de compressão de 50%. Observa-se que há bastante equilíbrio nos resultados, mas o método de palavras chaves do autor possui uma ligeira vantagem.

Figura 15 - Teste de variâncias equivalentes com taxa de compressão de 50% do domínio Educação Especial.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Educação Especial 50%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,768816	0,754784
Variância	0,0061308	0,004677635
Observações	5	5
Variância agrupada	0,005404217	
Hipótese da diferença de média	0	
gl	8	
Stat t	0,301802736	
P(T<=t) uni-caudal	0,38524955	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,7704991	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

Na Figura 16 (teste de variâncias equivalentes) têm-se os resultados para cada um dos métodos utilizados na sumarização automática com a taxa de 70%. Nota-se que as variâncias não têm diferenças significativas entre os dois métodos, com uma pequena vantagem para o método de palavras chaves do autor.

Figura 16 - Teste de variâncias equivalentes com taxa de compressão de 70% do domínio Educação Especial.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Educação Especial 70%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,582868	0,58038
Variância	0,004300058	0,003780914
Observações	5	5
Variância agrupada	0,004040486	
Hipótese da diferença de média	0	
gl	8	
Stat t	0,06188759	
P(T<=t) uni-caudal	0,476085302	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,952170605	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

A Figura 17 (teste de variâncias equivalentes) mostram os resultados da variância entre os dois métodos utilizados no PragmaSUM com a taxa de 80%. É possível perceber que existe uma variação muito pequena dentre resultados aferidos e o método de palavras chaves do autor possui ligeira vantagem.

Figura 17 - Teste de variâncias equivalentes com taxa de compressão de 80% do domínio Educação Especial.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Educação Especial 80%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,45182	0,445914
Variância	0,00249295	0,001467758
Observações	5	5
Variância agrupada	0,001980354	
Hipótese da diferença de média	0	
gl	8	
Stat t	0,2098418	
P(T<=t) uni-caudal	0,419518311	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,839036621	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

É possível perceber na Figura 18 (teste variâncias equivalentes) que os métodos possuem diferenças mínimas entre suas variâncias, com uma ligeira vantagem para o método de palavras chaves do autor.

Figura 18 - Teste de variâncias equivalentes com taxa de compressão de 90% do domínio Educação Especial.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Educação Especial 90%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,283978	0,258594
Variância	0,002965478	0,000975576
Observações	5	5
Variância agrupada	0,001970527	
Hipótese da diferença de média	0	
gl	8	
Stat t	0,904146648	
P(T<=t) uni-caudal	0,196165158	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,392330317	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

2 DOMÍNIO EDUCAÇÃO PERMANENTE

As Figuras desta seção trazem dos resultados dos testes estatísticos entre os métodos de palavras chaves do autor e palavras com maior frequência na sumarização automática utilizada no PragmaSUM no domínio Educação Permanente, utilizando quatro taxas de compressão.

A Figura 19 (teste de variâncias equivalentes) apresenta os resultados dos testes estatísticos para os métodos de palavras chaves do autor e palavras com maior frequência com a taxa de compressão de 50%. Observa-se um equilíbrio nos resultados, mas o método de palavras chaves possui uma ligeira vantagem.

Figura 19 - Teste de variâncias equivalentes com taxa de compressão de 50% do domínio Educação Permanente.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Educação Permanente 50%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,78669	0,782228
Variância	0,002001471	0,001505783
Observações	5	5
Variância agrupada	0,001753627	
Hipótese da diferença de média	0	
gl	8	
Stat t	0,168473245	
P(T<=t) uni-caudal	0,435196253	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,870392506	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

Na Figura 20 (teste de variâncias equivalentes) têm-se os resultados dos métodos de palavras chaves do autor e palavras com maior frequência, com a taxa de compressão de 70%. Nota-se que a variância entre os métodos foi mínima, com uma ligeira vantagem para o método de palavras chaves do autor.

Figura 20 - Teste de variâncias equivalentes com taxa de compressão de 70% do domínio Educação Permanente.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Educação Permanente 70%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,608208	0,580044
Variância	0,003929505	0,002235638
Observações	5	5
Variância agrupada	0,003082572	
Hipótese da diferença de média	0	
gl	8	
Stat t	0,802061661	
P(T<=t) uni-caudal	0,222843276	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,445686552	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

A Figura 21 (teste de variâncias equivalentes) apresenta os resultados dos métodos utilizados na sumarização automática com a taxa de 80%. Nota-se que as variâncias não têm diferenças significativas entre os métodos, com uma ligeira vantagem para o método de palavras chaves do autor.

Figura 21 - Teste de variâncias equivalentes com taxa de compressão de 80% do domínio Educação Permanente.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Educação Permanente 80%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,476186	0,445268
Variância	0,006634786	0,003572382
Observações	5	5
Variância agrupada	0,005103584	
Hipótese da diferença de média	0	
gl	8	
Stat t	0,684295634	
P(T<=t) uni-caudal	0,25655997	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,51311994	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

Na Figura 22 (teste variâncias equivalentes) é possível perceber que os métodos possuem diferenças mínimas entre suas variâncias, com uma pequena vantagem para o método de palavras chaves do autor.

Figura 22 - Teste de variâncias equivalentes com taxa de compressão de 90% do domínio Educação Permanente.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Educação Permanente 90%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,314222	0,278378
Variância	0,008269634	0,006065218
Observações	5	5
Variância agrupada	0,007167426	
Hipótese da diferença de média	0	
gl	8	
Stat t	0,669429531	
P(T<=t) uni-caudal	0,261034381	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,522068761	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

3 DOMÍNIO EDUCAÇÃO PRÉ-ESCOLAR

Esta seção descreve os resultados do teste *T-Student* no domínio Educação Pré-Escolar. As figuras trazem dos resultados dos testes estatísticos para os dois métodos da sumarização automática utilizada no PragmaSUM com quatro taxas de compressão.

A Figura 23 (teste variâncias equivalentes) apresentam os resultados das variâncias entre os métodos de sumarização com a taxa de 50%. Observa-se que há bastante equilíbrio nos resultados, com uma ligeira vantagem para o método de palavras com maior frequência.

Figura 23 - Teste de variâncias equivalentes com taxa de compressão de 50% do domínio Educação Pré-Escolar.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Educação Pré-Escolar 50%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,780026	0,76512
Variância	0,000363379	0,000461193
Observações	5	5
Variância agrupada	0,000412286	
Hipótese da diferença de média	0	
gl	8	
Stat t	1,160731948	
P(T<=t) uni-caudal	0,139606819	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,279213638	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

Na Figura 24 (teste variâncias equivalentes) têm-se os resultados entre dois métodos para de sumarização automática utilizando o PragmaSUM. Nota-se uma pequena vantagem para o método de palavras com maior frequência.

Figura 24 - Teste de variâncias equivalentes com taxa de compressão de 70% do domínio Educação Pré-Escolar

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Educação Pré-Escolar 70%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,589172	0,585124
Variância	0,000205367	0,000537998
Observações	5	5
Variância agrupada	0,000371682	
Hipótese da diferença de média	0	
gl	8	
Stat t	0,331989657	
P(T<=t) uni-caudal	0,374213941	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,748427881	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

A Figura 25 (teste variâncias equivalentes) mostram os resultados entre a variância entre dois métodos na SA. É possível perceber que existe uma variação muita pequena dentre resultados aferidos e o método de palavras chaves do autor possui ligeira vantagem.

Figura 25 - Teste de variâncias equivalentes com taxa de compressão de 80% do domínio Educação Pré-Escolar.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Educação Pré-Escolar 80%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,476186	0,450428
Variância	0,000762346	0,000542678
Observações	5	5
Variância agrupada	0,000407789	
Hipótese da diferença de média	0	
gl	8	
Stat t	0,637974694	
P(T<=t) uni-caudal	0,270658319	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,541316637	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

Percebe-se na Figura 26 (teste variâncias equivalentes) que o método de palavras com maior frequência teve vantagem mínima sobre o método de palavras chaves do autor, utilizando a taxa de compressão de 90%.

Figura 26 - Teste de variâncias equivalentes com taxa de compressão de 90% do domínio Educação Pré-Escolar.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Educação Pré-Escolar 90%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,26258	0,250748
Variância	0,000366186	0,000421341
Observações	5	5
Variância agrupada	0,000393763	
Hipótese da diferença de média	0	
gl	8	
Stat t	0,942780303	
P(T<=t) uni-caudal	0,186694376	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,373388752	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

4 DOMÍNIO ENSINO APRENDIZAGEM

As Figuras desta seção trazem dos resultados dos testes estatísticos entre os métodos de palavras chaves do autor e palavras com maior frequência na SA utilizada no PragmaSUM no domínio Ensino Aprendizagem.

A Figura 27 (teste de variâncias equivalentes) apresenta os resultados dos testes estatísticos para os métodos de palavras chaves do autor e palavras com maior frequência com a taxa de compressão de 50%. Observa-se um equilíbrio nos resultados, mas o método de palavras chaves possui uma ligeira vantagem.

Figura 27 - Teste de variâncias equivalentes com taxa de compressão de 50% do domínio Ensino Aprendizagem.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Ensino-Aprendizagem 50%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,668402	0,659458
Variância	0,07306775	0,068168017
Observações	5	5
Variância agrupada	0,070617884	
Hipótese da diferença de média	0	
gl	8	
Stat t	0,053216272	
P(T<=t) uni-caudal	0,479432241	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,958864481	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

Na Figura 28 (teste de variâncias equivalentes) têm-se os resultados entre dois métodos na SA. Nota-se que a variância foi mínima, com uma pequena vantagem para o método de palavras chaves do autor.

Figura 28 - Teste de variâncias equivalentes com taxa de compressão de 70% do domínio Ensino Aprendizagem.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Ensino-Aprendizagem 70%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,51547	0,49699
Variância	0,054731928	0,046059998
Observações	5	5
Variância agrupada	0,050395963	
Hipótese da diferença de média	0	
gl	8	
Stat t	0,130158968	
P(T<=t) uni-caudal	0,449826979	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,899653957	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

A Figura 29 (teste de variâncias equivalentes) mostram os resultados da SA utilizando o PragmaSUM entre dois métodos. É possível perceber que existe uma variação muito pequena entre os resultados aferidos e o método de palavras chaves do autor possui ligeira vantagem.

Figura 29 - Teste de variâncias equivalentes com taxa de compressão de 80% do domínio Ensino Aprendizagem.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Ensino-Aprendizagem 80%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,400632	0,381052
Variância	0,035138397	0,027833072
Observações	5	5
Variância agrupada	0,031485734	
Hipótese da diferença de média	0	
gl	8	
Stat t	0,174472015	
P(T<=t) uni-caudal	0,43291452	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,865829039	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

Percebe-se na figura a 30 (teste de variâncias equivalentes) que o método de palavras chaves do autor possui uma pequena vantagem sobre o método de palavras com maior frequência.

Figura 30 - Teste de variâncias equivalentes com taxa de compressão de 90% do domínio Ensino Aprendizagem.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Ensino-Aprendizagem 90%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,239996	0,222704
Variância	0,012992381	0,010968678
Observações	5	5
Variância agrupada	0,01198053	
Hipótese da diferença de média	0	
gl	8	
Stat t	0,24979125	
P(T<=t) uni-caudal	0,40452167	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,809043341	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

5 DOMÍNIO FILOSOFIA DA EDUCAÇÃO

As Figuras desta seção trazem os resultados do teste estatístico *T-Student* no domínio Filosofia da Educação com quatro taxas de compressão.

Percebe-se na figura 31 (teste variâncias equivalentes) que o método de palavras com maior frequência teve uma ligeira vantagem sobre o método de palavras chaves do autor, utilizando a taxa de compressão de 50%.

Figura 31 - Teste de variâncias equivalentes com taxa de compressão de 50% do domínio Filosofia da Educação.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Filosofia da Educação 50%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,77077	0,765266
Variância	0,00018148	0,000355038
Observações	5	5
Variância agrupada	0,000268259	
Hipótese da diferença de média	0	
gl	8	
Stat t	0,531338576	
P(T<=t) uni-caudal	0,304807943	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,609615886	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

Na figura 32 (teste variâncias equivalentes) apresenta os resultados entre dois métodos de SA. Percebe-se que o método de palavras com maior frequência teve ligeira vantagem.

Figura 32 - Teste de variâncias equivalentes com taxa de compressão de 70% do domínio Filosofia da Educação.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Filosofia da Educação 70%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,56312	0,56779
Variância	0,000654695	0,000187158
Observações	5	5
Variância agrupada	0,000420927	
Hipótese da diferença de média	0	
gl	8	
Stat t	-0,359901546	
P(T<=t) uni-caudal	0,364118404	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,728236807	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

A Figura 33 (teste variâncias equivalentes) mostram os resultados da variância entre os dois métodos utilizados no PragmaSUM. É possível perceber que existe uma variação muito pequena dentre resultados aferidos e o método de palavras chaves do autor possui ligeira vantagem.

Figura 33 - Teste de variâncias equivalentes com taxa de compressão de 80% do domínio Filosofia da Educação.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Filosofia da Educação 80%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,430386	0,42052
Variância	0,000750818	0,000547779
Observações	5	5
Variância agrupada	0,000649298	
Hipótese da diferença de média	0	
gl	8	
Stat t	0,612193824	
P(T<=t) uni-caudal	0,278702287	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,557404574	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

Percebe-se na figura 34 (teste variâncias equivalentes) que o método de palavras chaves do autor teve uma ligeira vantagem sobre o método de palavras com maior frequência.

Figura 34 - Teste de variâncias equivalentes com taxa de compressão de 90% do domínio Filosofia da Educação.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Filosofia da Educação 90%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,249616	0,23201
Variância	0,000367456	0,000279566
Observações	5	5
Variância agrupada	0,000323511	
Hipótese da diferença de média	0	
gl	8	
Stat t	1,54769828	
P(T<=t) uni-caudal	0,08014196	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,160283921	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

6 DOMÍNIO HISTÓRIA DA EDUCAÇÃO

Esta seção descreve os resultados do teste *T-Student* no domínio Educação Especial. As Figuras 35, 36, 37 e 38 trazem dos resultados dos testes estatísticos para os dois métodos da SA utilizada no PragmaSUM no domínio História da Educação.

A Figura 35 (teste de variâncias equivalentes) apresenta os resultados dos testes estatísticos para dois métodos de SA utilizando a taxa de compressão de 50%. Observa-se que há bastante equilíbrio nos resultados, mas o método de palavras com maior frequência possui uma ligeira vantagem.

Figura 35 - Teste de variâncias equivalentes com taxa de compressão de 50% do domínio História da Educação.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio História da Educação 50%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,723334	0,72414
Variância	0,0129541	0,013144755
Observações	5	5
Variância agrupada	0,013049428	
Hipótese da diferença de média	0	
gl	8	
Stat t	-0,011156021	
P(T<=t) uni-caudal	0,495686078	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,991372157	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

Na Figura 36 (teste de variâncias equivalentes) têm-se os resultados entre os métodos de SA. Nota-se que o método de palavras chaves do autor teve uma pequena vantagem.

Figura 36 - Teste de variâncias equivalentes com taxa de compressão de 70% do domínio História da Educação.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio História da Educação 70%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,59385	0,589784
Variância	0,012977979	0,012130747
Observações	5	5
Variância agrupada	0,012554363	
Hipótese da diferença de média	0	
gl	8	
Stat t	0,057377291	
P(T<=t) uni-caudal	0,477825944	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,955651889	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

A Figura 37 (teste de variâncias equivalentes) mostra que o método de palavras com maior frequência teve uma ligeira vantagem sobre o método de palavras chaves do autor.

Figura 37 - Teste de variâncias equivalentes com taxa de compressão de 80% do domínio História da Educação.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio História da Educação 80%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,506014	0,505028
Variância	0,012532134	0,013844848
Observações	5	5
Variância agrupada	0,013188491	
Hipótese da diferença de média	0	
gl	8	
Stat t	0,013575298	
P(T<=t) uni-caudal	0,494750627	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,989501254	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

Na Figura 38 (teste de variâncias equivalentes) percebe-se que a variância entre os métodos foi mínima, com uma pequena vantagem para o método de palavras com maior frequência.

Figura 38 - Teste de variâncias equivalentes com taxa de compressão de 90% do domínio História da Educação.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio História da Educação 90%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,377726	0,375006
Variância	0,020465496	0,022903126
Observações	5	5
Variância agrupada	0,021684311	
Hipótese da diferença de média	0	
gl	8	
Stat t	0,029205597	
P(T<=t) uni-caudal	0,48870803	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,97741606	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

7 DOMÍNIO POLÍTICA EDUCACIONAL

Esta seção descreve os resultados do teste *T-Student* no domínio Política Educacional. As figuras trazem dos resultados dos testes estatísticos para os dois métodos da sumarização automática utilizada no PragmaSUM.

A Figura 39 (teste variâncias equivalentes) apresentam os resultados das variâncias entre os métodos de sumarização com a taxa de 50%. Observa-se que há bastante equilíbrio nos resultados, com uma ligeira vantagem para o método de palavras com maior frequência.

Figura 39 - Teste de variâncias equivalentes com taxa de compressão de 50% do domínio Política Educacional.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Política Educacional 50%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,762746	0,7587
Variância	0,000710832	0,00091209
Observações	5	5
Variância agrupada	0,000811461	
Hipótese da diferença de média	0	
gl	8	
Stat t	0,224575295	
P(T<=t) uni-caudal	0,413969715	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,82793943	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

Na Figura 40 (teste variâncias equivalentes) têm-se os resultados entre dois métodos na SA. Percebe-se que o método de palavras chaves do autor teve uma ligeira vantagem.

Figura 40 - Teste de variâncias equivalentes com taxa de compressão de 70% do domínio Política Educacional.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Política Educacional 70%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,578346	0,576214
Variância	0,00159669	0,001272372
Observações	5	5
Variância agrupada	0,001434531	
Hipótese da diferença de média	0	
gl	8	
Stat t	0,089002511	
P(T<=t) uni-caudal	0,465633852	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,931267703	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

A Figura 41 (teste variâncias equivalentes) mostram os resultados da variância entre dois métodos na sumarização automática. Nota-se que o método de palavras com maior frequência teve uma pequena vantagem em relação ao método de palavras chaves do autor.

Figura 41 - Teste de variâncias equivalentes com taxa de compressão de 80% do domínio Política Educacional.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Política Educacional 80%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,441484	0,445938
Variância	0,002495474	0,001070114
Observações	5	5
Variância agrupada	0,001782794	
Hipótese da diferença de média	0	
gl	8	
Stat t	-0,166789878	
P(T<=t) uni-caudal	0,435837019	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,871674038	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

Percebe-se na Figura 42 (teste variâncias equivalentes) que o método de palavras chaves do autor obteve uma ligeira sobre o método de palavras com maior frequência.

Figura 42 - Teste de variâncias equivalentes com taxa de compressão de 90% do domínio Política Educacional.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Política Educacional 90%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,268114	0,260946
Variância	0,001209368	0,000483883
Observações	5	5
Variância agrupada	0,000846626	
Hipótese da diferença de média	0	
gl	8	
Stat t	0,389513332	
P(T<=t) uni-caudal	0,353530227	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,707060453	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

8 DOMÍNIO PSICOLOGIA EDUCACIONAL

Esta seção descreve os resultados do teste *T-Student* no domínio Psicologia Educacional. As figuras 43, 44, 45 e 46 trazem dos resultados dos testes estatísticos para os dois métodos da sumarização automática utilizada no PragmaSUM em quatro taxas de compressão.

A Figura 43 (teste de variâncias equivalentes) apresenta os resultados dos testes estatísticos entre dois métodos com a taxa de compressão de 50%. Observa-se que há bastante equilíbrio nos resultados, mas o método de palavras com maior frequência possui uma ligeira vantagem.

Figura 43 - Teste de variâncias equivalentes com taxa de compressão de 50% do domínio Psicologia Educacional.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Psicologia Educacional 50%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,793962	0,77353
Variância	0,001037606	0,002175135
Observações	5	5
Variância agrupada	0,00160637	
Hipótese da diferença de média	0	
gl	8	
Stat t	0,806042689	
P(T<=t) uni-caudal	0,221758121	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,443516242	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

Na Figura 44 (teste de variâncias equivalentes) têm-se os resultados para cada um dos métodos utilizados na sumarização automática com a taxa de 70%. Nota-se que as variâncias não têm diferenças significativas entre os dois métodos, com uma pequena vantagem para o método de palavras com maior frequência.

Figura 44 - Teste de variâncias equivalentes com taxa de compressão de 70% do domínio Psicologia Educacional.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Psicologia Educacional 70%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,601784	0,578648
Variância	0,001507551	0,004468348
Observações	5	5
Variância agrupada	0,00298795	
Hipótese da diferença de média	0	
gl	8	
Stat t	0,669224225	
P(T<=t) uni-caudal	0,261096509	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,522193017	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

É possível perceber na Figura 45 (teste variâncias equivalentes) que os métodos possuem diferenças mínimas entre suas variâncias, com uma ligeira vantagem para o método de palavras com maior frequência.

Figura 45 - Teste de variâncias equivalentes com taxa de compressão de 80% do domínio Psicologia Educacional.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Psicologia Educacional 80%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,449814	0,439112
Variância	0,001842535	0,003708412
Observações	5	5
Variância agrupada	0,002775473	
Hipótese da diferença de média	0	
gl	8	
Stat t	0,321193259	
P(T<=t) uni-caudal	0,378147318	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,756294636	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

Na Figura 46 (teste variâncias equivalentes) têm-se os resultados entre dois métodos para de sumarização automática utilizando o PragmaSUM. Nota-se uma pequena vantagem para o método de palavras com maior frequência.

Figura 46 - Teste de variâncias equivalentes com taxa de compressão de 90% do domínio Psicologia Educacional.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Psicologia Educacional 90%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,261026	0,25037
Variância	0,000872115	0,003191112
Observações	5	5
Variância agrupada	0,002031614	
Hipótese da diferença de média	0	
gl	8	
Stat t	0,373803761	
P(T<=t) uni-caudal	0,359131275	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,71826255	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

9 DOMÍNIO SOCIOLOGIA DA EDUCAÇÃO

As Figuras desta seção trazem os resultados do teste estatístico *T-Student* no domínio Sociologia da Educação com quatro taxas de compressão.

Percebe-se na Figura 47 (teste variâncias equivalentes) que o método de palavras com maior frequência teve uma ligeira vantagem sobre o método de palavras com maior frequência, utilizando a taxa de compressão de 50%.

Figura 47 - Teste de variâncias equivalentes com taxa de compressão de 50% do domínio Sociologia da Educação.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Sociologia da Educação 50%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,767334	0,75866
Variância	0,000668286	0,001751264
Observações	5	5
Variância agrupada	0,001209775	
Hipótese da diferença de média	0	
gl	8	
Stat t	0,394309408	
P(T<=t) uni-caudal	0,351827705	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,70365541	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

Na Figura 48 (teste variâncias equivalentes) apresenta os resultados entre dois métodos de SA. Percebe-se que o método de palavras com maior frequência teve ligeira vantagem.

Figura 48 - Teste de variâncias equivalentes com taxa de compressão de 70% do domínio Sociologia da Educação.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Sociologia da Educação 70%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,576882	0,566524
Variância	0,001229928	0,001836111
Observações	5	5
Variância agrupada	0,001533019	
Hipótese da diferença de média	0	
gl	8	
Stat t	0,418284799	
P(T<=t) uni-caudal	0,343370761	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,686741522	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

A Figura 49 (teste variâncias equivalentes) mostram os resultados da variância entre os dois métodos utilizados no PragmaSUM. É possível perceber que existe uma variação muito pequena dentre resultados aferidos e o método de palavras chaves do autor possui ligeira vantagem.

Figura 49 - Teste de variâncias equivalentes com taxa de compressão de 80% do domínio Sociologia da Educação.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Sociologia da Educação 80%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,45875	0,443902
Variância	0,002115131	0,001785627
Observações	5	5
Variância agrupada	0,001950379	
Hipótese da diferença de média	0	
gl	8	
Stat t	0,531592085	
P(T<=t) uni-caudal	0,304724083	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,609448166	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

Percebe-se na Figura 50 (teste variâncias equivalentes) que o método de palavras chaves com maior frequência teve uma ligeira vantagem sobre o método de palavras chaves do autor.

Figura 50 - Teste de variâncias equivalentes com taxa de compressão de 90% do domínio Sociologia da Educação.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Sociologia da Educação 90%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,277996	0,238842
Variância	0,000777208	0,001771997
Observações	5	5
Variância agrupada	0,001274603	
Hipótese da diferença de média	0	
gl	8	
Stat t	1,734038537	
P(T<=t) uni-caudal	0,060568331	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,121136661	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

10 DOMÍNIO TECNOLOGIA EDUCACIONAL

Esta seção descreve os resultados do teste *T-Student* no domínio Tecnologia Educacional. As figuras trazem dos resultados dos testes estatísticos para os dois métodos da sumarização automática utilizada no PragmaSUM.

A Figura 51 (teste variâncias equivalentes) apresentam os resultados das variâncias entre os métodos de sumarização com a taxa de 50%. Observa-se que há bastante equilíbrio nos resultados, com uma ligeira vantagem para o método de palavras com maior frequência.

Figura 51 - Teste de variâncias equivalentes com taxa de compressão de 50% do domínio Tecnologia Educacional.

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Tecnologia Educacional 50%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,750846	0,745426
Variância	0,000659505	0,001105889
Observações	5	5
Variância agrupada	0,000882697	
Hipótese da diferença de média	0	
gl	8	
Stat t	0,288445364	
P(T<=t) uni-caudal	0,390169137	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,780338274	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

Na Figura 52 (teste variâncias equivalentes) têm-se os resultados entre dois métodos na SA. Percebe-se que o método de palavras chaves do autor teve uma ligeira vantagem.

**Figura 52 - Teste de variâncias equivalentes com taxa de compressão de 70% do domínio
Tecnologia Educacional.**

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Tecnologia Educacional 70%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,555574	0,556724
Variância	0,001781363	0,001304963
Observações	5	5
Variância agrupada	0,001543163	
Hipótese da diferença de média	0	
gl	8	
Stat t	-0,046287316	
P(T<=t) uni-caudal	0,482107928	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,964215855	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

Percebe-se na Figura 53 (teste variâncias equivalentes) que o método de palavras chaves do autor obteve uma ligeira sobre o método de palavras com maior frequência.

**Figura 53 - Teste de variâncias equivalentes com taxa de compressão de 80% do domínio
Tecnologia Educacional.**

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Tecnologia Educacional 80%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,435742	0,420546
Variância	0,001338504	0,000510671
Observações	5	5
Variância agrupada	0,000924587	
Hipótese da diferença de média	0	
gl	8	
Stat t	0,790178662	
P(T<=t) uni-caudal	0,226103842	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,452207683	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.

A Figura 54 (teste de variâncias equivalentes) mostra que o método de palavras com maior frequência teve uma ligeira vantagem sobre o método de palavras chaves do autor.

**Figura 54 - Teste de variâncias equivalentes com taxa de compressão de 90% do domínio
Tecnologia Educacional.**

Teste-t: duas amostras presumindo variâncias equivalentes		
Domínio Tecnologia Educacional 90%		
	<i>Palavras Chaves do Autor</i>	<i>Palavras com Maior Frequência</i>
Média	0,258572	0,234638
Variância	0,000755069	0,000909475
Observações	5	5
Variância agrupada	0,000832272	
Hipótese da diferença de média	0	
gl	8	
Stat t	1,311754528	
P(T<=t) uni-caudal	0,112996606	
t crítico uni-caudal	1,859548038	
P(T<=t) bi-caudal	0,225993213	
t crítico bi-caudal	2,306004135	

Fonte: O AUTOR, 2019.