UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
GRADUAÇÃO EM ENGENHARIA DA COMPUTAÇÃO

SISTEMA DE SUMARIZAÇÃO AUTOMÁTICA DE TEXTOS BASEADO EM CLASSES DE DOCUMENTOS

IHAGO HENRIQUE LUCENA E SILVA

RECIFE - PE
JULHO, 2017

IHAGO HENRIQUE LUCENA E SILVA

SISTEMA DE SUMARIZAÇÃO AUTOMÁTICA DE TEXTOS BASEADO EM CLASSES DE DOCUMENTOS

Monografia apresentada ao Centro de Informática da Universidade Federal de Pernambuco como pré-requisito para obtenção do Título de Bacharel em Engenharia da Computação.

Orientador: Prof. Dr. Rafael Dueire Lins

RECIFE - PE
JULHO, 2017

IHAGO HENRIQUE LUCENA E SILVA

SISTEMA DE SUMARIZAÇÃO AUTOMÁTICA DE TEXTOS BASEADO EM CLASSES DE DOCUMENTOS

Monografia apresentada ao Centro de Informática da Universidade Federal de Pernambuco como pré-requisito para obtenção do Título de Bacharel em Engenharia da Computação.

Recife, 10 de Julho de 2017

BANCA EXAMINADORA

_____

Prof. Dr. Rafael Dueire Lins
Orientador

_____

Prof. Dr. Frederico Luiz Gonçalves de Freitas
Avaliador

AGRADECIMENTOS

Agradeço, primeiramente, à **Deus** que iluminou o meu caminho e me deu muitas forças para chegar até aqui.

Agradeço, aos meus pais **Normando José Silva** e **Luciene Ferreira de Lucena e Silva**, pois graças à eles, foi possível transformar todos os meus sonhos em realidade. Por isto e tudo mais, eu os agradeço e os dedico este trabalho. Agradeço, também, ao meu irmão **Augusto César de Lucena e Silva**, que mesmo morando longe, sempre me apoiou e contribuiu indiretamente para a minha formação acadêmica.

Agradeço a **Kerolayne Gomes Tato Cota**, uma das minhas melhores amigas e das pessoas mais incríveis e extraordinárias que eu já conheci ao longo de toda minha vida. Sem ela, com toda certeza, eu não teria chegado aqui. A **Heitor Fonseca de Araujo** que sempre se mostrou muito solícito em me ajudar em todas as atividades acadêmicas, inclusive na realização deste trabalho.

Agradeço ao Prof. **Rafael Dueire Lins**, orientador deste trabalho, por ter aceitado me orientar, por me auxiliar na produção desta monografia e, principalmente por prestar sua paciência, atenção e boa vontade. Ao Prof. **Rafael Ferreira Leite de Mello** que, juntamente com meu orientador, me instruiu da melhor maneira possível e me concedeu os meios necessários para a concepção e implantação deste trabalho. Ao Prof. **Frederico Luiz Gonçalves de Freitas** pela disponibilidade de participar da banca de avaliação, por seus comentários e correções feitos ao trabalho.

Ao Prof. **Eduardo Antônio Guimarães Tavares** e à **Eric Rodrigues Borba** por todo apoio no meu crescimento acadêmico durante a minha iniciação científica no Convênio CIn-Motorola e durante a minha graduação em geral.

Aos demais professores do Centro de Informática da Universidade Federal de Pernambuco por toda sabedoria que me transmitiram durante esses mais de cinco anos. Especialmente aos professores **Adriano Augusto de Moraes Sarmento**, **Carlos Alexandre Barros de Mello**, **Kiev Santos da Gama** e **Marcelo Bezerra d'Amorim** pela preocupação com o meu futuro pessoal e profissional.

Por fim, agradeço a todos os meus sinceros, leais e grandes amigos que são muito difíceis de encontrar. Por isso, gratulo os poucos, mas valiosos que ainda fazem parte da minha vida. Juntamente a todas as outras pessoas que de alguma forma participaram dessa longa e árdua jornada.

RESUMO

A sumarização automática de textos é uma das principais aplicações da área de Processamento de Linguagem Natural (PLN) e consiste, sobretudo, em criar automaticamente, como o próprio nome já diz, uma versão mais curta de um ou mais documentos de texto, mantendo suas informações fundamentais. Um dos seus usos mais relevantes é auxiliando usuários a encontrar informações desejadas em grandes bibliotecas de texto, principalmente na *Internet*, de forma rápida e precisa. Devido, então, ao grande aumento no volume de dados da Internet, o interesse por essa área cresceu.

Assim, o trabalho aqui proposto visa descobrir se a classificação de um documento está correlacionada com a técnica de sumarização extrativa mais eficiente e, em tal caso, determinar também as técnicas mais adequadas, dentre as mais amplamente difundidas para a sumarização extrativa, para cada classe de documentos. Logo, a fim de ratificar tal asserção, um sistema de sumarização automática de textos baseada nas classes dos documentos foi implementado e teve sua eficácia mensurada através da avaliação da qualidade dos sumários gerados.

Palavras-chave: Sumarização Automática de Textos. Sumarização Extrativa. Classificação de Documentos. Processamento de Linguagem Natural.

ABSTRACT

Automatic text summarization is one of the main applications in the area of Natural Language Processing (PLN). It consists, above all, of creating a shorter version of one or more text documents, keeping their fundamental key information. One of its most relevant uses is supporting users to find desired information in large text databases, especially on the Internet, in a quick and accurate path. The demand for new automatic summarization approaches has risen considerably in the last years.

This research aims to figure out whether the classification of a document is correlated to extractive summarization techniques and, in such a case, to determine the most appropriate approach for each specific class of documents. An automatic text summarization system based on the classes of documents was implemented and had its effectiveness measured through the evaluation of the quality of the summaries generated.

Keywords: Automatic Text Summarization. Extractive Summarization. Document Classification. Natural Language Processing.

SUMÁRIO

LISTA DE QUADROS

LISTA DE TABELAS

CAPÍTULO 1 - INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO E MOTIVAÇÃO

Atualmente, bilhões de usuários navegam na Internet regularmente e, quase sempre, estão interagindo com algum engenho de busca como o *Google* ou o *Bing*. Cada um desses usuários tem suas próprias necessidades e procuram por informações específicas. Por mais que os mecanismos de recuperação de informação tenham evoluído ao longo dos anos, trazendo sempre os melhores resultados de uma determinada consulta de forma ordenada, ainda é muito comum o usuário perder um tempo significante reformulando suas consultas e buscando o que, de fato, procura na *web* [1]. Excluindo-se os resultados com anúncios e o fato de que alguns usuários não estão formulando corretamente suas *queries* de busca, essa perda de tempo também pode ser proveniente de informações ruidosas presentes nos documentos recuperados, geralmente informações que não dizem respeito ao conteúdo essencial do arquivo.

Devido ao crescimento exponencial da quantidade de informações disponíveis *online*, tornou-se inconcebível para os usuários analisar documento por documento [2] dos encontrados pelos engenhos de busca afim de selecionar as informações mais relevantes. Por isso, há um crescente interesse entre a comunidade de pesquisa no desenvolvimento de novas abordagens para resumir automaticamente o texto [3] tanto para que o próprio usuário não precise peneirar documentos enormes, quanto na própria atividade de recuperação de informação.

Também existem grandes benefícios da geração de sumários para a área de classificação automática de textos [4**,** 5, 6], uma vez que a sumarização diminuí a quantidade de informações que geram uma sobreposição entre as classes não disjuntas, aumentando, dessa forma a precisão do classificador [7]. Entende-se aqui por sumários todo tipo de versão menor, mais resumida, de um ou mais documentos de texto que preserve suas informações essenciais.

Há diversas técnicas e formas de se extrair as informações mais essenciais de um ou mais documentos, bem como diversas variáveis a serem consideradas nessa tarefa, como o idioma do texto, o seu domínio linguístico, a quantidade de informações redundantes, dentre outras. Ultimamente, várias abordagens para melhorar a qualidade dos sumarizadores automáticos, e consequentemente, dos seus resumos estão sendo propostas. Principalmente, as que levam em conta as características do texto a ser sumarizado [8], visto que, obviamente, é quase impossível criar um método genérico que consiga resumir diferentes tipos de texto de forma eficiente.

Vislumbrando esse contexto, este trabalho expõe a proposta de um sistema que utiliza uma nova abordagem para sumarizar os documentos: a sua própria classificação. Por exemplo, as técnicas de sumarização utilizadas para resumir documentos da classe “Economia” provavelmente são diferentes das utilizadas para resumir os da classe “Esportes”. Exposto isso, tornou-se então necessário verificar sistematicamente e experimentalmente tal abordagem proposta.

1.2 OBJETIVOS

Como já evidenciado anteriormente, o objetivo geral proposto aqui neste trabalho foi, através da implementação de um sistema de sumarização de textos, verificar se classificação de um documento está correlacionada com a técnica mais eficiente de sumarização e, nessa situação, identificar quais as melhores técnicas de sumarização extrativa produzem os melhores resumos para cada uma das classificações do documento.

1.3 ESTRUTURA

Para atingir os objetivos apresentados na subseção anterior, este trabalho está estruturado em 5 capítulos, sendo o primeiro esta introdução.

O capítulo 2 apresenta toda a fundamentação teórica necessária para o entendimento deste trabalho realizando uma breve, porém importante, abordagem histórica da área, caracterizando o processo de sumarização automática de textos em si, bem como os tipos de sumarização utilizados hoje, igualmente o procedimento de sumarização extrativa e, finalmente, apresentando, ainda, as dificuldades e metodologias de avaliação da qualidade de sumários mais empregadas atualmente.

O capítulo 3 descreve toda a metodologia utilizada neste trabalho: desde à seleção dos métodos de sumarização extrativa analisados até a avaliação dos extratos gerados pelo sistema proposto, exibindo a elaboração da arquitetura do sistema e a organização do corpus de texto empregue para avaliação do sistema.

No capítulo 4 os resultados obtidos nas avaliações dos sumários gerados são exibidos e discutidos. Por fim o capítulo 5 apresenta as considerações finais, contribuições e possíveis trabalhos futuros.

Ao final, na secção de apêndice encontram-se as tabelas com os resultados detalhados obtidos na experimentação deste trabalho.

CAPÍTULO 2 - REFERENCIAL TEÓRICO

2.1 ABORDAGEM HISTÓRICA

A sumarização automática vem sendo objeto de estudo desde os primórdios da computação. Já no final da década de 50 começaram a surgir alguns métodos estatísticos para extrair as sentenças principais de um texto [9]. Os resumos automáticos criados, naquela mesma década, por computadores IBM 704 e IBM 705 foram, talvez os primeiros exemplos de sumários gerados por máquina [10].

Na década de 60 a avaliação da qualidade de resumos passou a despertar o interesse de estudiosos dessa área. Nessa década várias avaliações de sumários foram realizadas. Embora, nessa época, não se tenha chegado a um consenso sobre questões de avaliação de qualidade de um sumário, um panorama bastante claro começou a emergir do estudo dos métodos de avaliação explorados até o momento [11].

Durantes as décadas de 70 e 80 alguns novos métodos de sumarização começaram a ser bastante explorados. Métodos como o das palavras-chaves, palavras-chaves do título, localização, palavras sinalizadoras (*Cue phrases*) ou dos marcadores linguísticos, relacional e de frase auto indicativa [9].

Houve o ressurgimento da área de mineração de textos (devido a uma reorientação de pesquisa) na década de 90 e consequentemente o intercâmbio de conhecimentos entre essa área e a de sumarização automática de textos. A primeira é responsável por buscar padrões linguísticos em textos através de técnicas de inteligência artificial. Geralmente, utilizam-se conhecimentos dessa área para extrair determinadas informações em um grande volume de textos [9].

Desde a década de 50 até a década de 90, considera-se que a área de sumarização não teve muitos avanços devido à utilização de técnicas muito simples que não rendiam bons resultados. Outro motivo dessa estagnação devia-se a pequena necessidade até aquele momento, cenário este que mudou drasticamente com a popularização da Internet e do enorme surgimento de documentos on-line [9].

Somente em maio de 1998, o governo dos Estados Unidos completou o SUMMAC (*TIPSTER Text Summarization Evaluation*) que foi o primeiro avaliador automático de textos em larga escala e, obviamente, independente de desenvolvedores de sistemas automáticos de sumarização de texto. Os objetivos do SUMMAC era julgar os sistemas individuais de sumarização de textos em termos de sua utilidade em tarefas específicas de sumarização e obter uma melhor compreensão das questões envolvidas na construção e avaliação de tais sistemas [12].

Desde os anos 2000, várias novas abordagens foram elaboradas para tentar melhorar ou até sanar muitos dos problemas da sumarização, principalmente os de coesão e coerência [9]. Métodos baseados em modelos de grafos, vetores de características, clusters, contexto/gênero, dentre outros.

A sumarização automática de texto tornou-se, então, uma maneira muito importante de encontrar informações relevantes em grandes bibliotecas de texto ou na Internet [8]. Na verdade, ela tem sido vista como uma abordagem eficaz para lidar com o aumento exponencial da quantidade de informações na Internet hoje em dia [7].

2.2 SUMARIZAÇÃO AUTOMÁTICA DE TEXTOS

Conforme já exposto, sumarizar é o processo de filtrar as informações mais significativas de um ou mais documentos [13]. Por ser um processo muito trabalhoso e dispendioso para ser executado manualmente, em alguns casos, passou-se cada vez mais a buscar formas automáticas e computacionais de realizar essa atividade através da aplicação de técnicas de sumarização.

Existem diversas técnicas de sumarização automática e geralmente elas ou envolvem explorações estatísticas ou conhecimentos linguísticos ou ambos [1]. As técnicas também tem seu domínio de aplicabilidade. Muitas delas podem ser aplicadas a qualquer domínio, obviamente, entretanto é provável que os resumos produzidos por elas sejam de baixa qualidade, enquanto que outras podem produzir resumos de alta qualidade, mas apenas em domínios muito restritos [14].

A utilização de técnicas de PLN trouxe muitos benefícios à área de sumarização, visto que a maioria das informações presentes nas grandes bibliotecas virtuais são não-estruturadas ou semi-estruturadas. As técnicas extrativas, por serem mais simples e exigirem menos de uma base de conhecimento linguístico, são as mais estudadas e empregadas nas diversas aplicações de sumarização automática. Predominantemente, são técnicas para o idioma inglês [15].

Um dos fatores mais importantes no processo de sumarização automática é o nível de compressão, isto é, a proporção do comprimento do resumo para o comprimento do texto dos documentos originais [13]. Como as técnicas de sumarização podem ser empregadas em vários tipos de documentos como biografias, enredos de filmes, sequências de e-mails, pautas de reuniões, artigos científicos, postagens de blogs, etc. o número de sentenças e tamanho do texto final, consequentemente, também apresentará variações. Inclusive há métodos de sumarização como Sentence Length que tende a evitar selecionar sentenças muito longas ou muito curtas [1].

2.3 TIPOS DE SUMARIZAÇÃO

Basicamente, os sistemas de sumarização automática de textos podem ser classificados de acordo com 3 tipos de critérios: abstrativo ou extrativo, genérico ou baseado em perguntas / consultas e mono documentos ou multi documentos.

Essencialmente, as técnicas desse tipo de processo são classificadas como extrativas ou abstrativas [16]. Sumarização extrativa produz um conjunto com as sentenças mais significativas de um documento, exatamente da mesma forma como elas aparecem (extratos). Já a sumarização abstrativa pode gerar novas sentenças a partir de palavras diferentes, mas que tenham o mesmo significado. A sumarização abstrativa tenta melhorar a coerência entre as sentenças produzidas [8, 17].

O processo de sumarização também pode ser genérico ou focado em uma pergunta / consulta. O processo genérico resume o documento retornando uma ou mais sentenças, varia de acordo com o tamanho do sumário desejado, com as maiores pontuações para um conjunto de características [18]. Para o processo de sumarização com ênfase em uma pergunta, procura-se pela sentença que contenha a informação requerida na pergunta em vez da sentença de maior pontuação do documento.

A sumarização também pode ser classificada de acordo com a quantidade de documentos a serem analisados simultaneamente: em mono documento, no qual só existe um único documento, e em multi-documento, onde uma coleção de textos sobre o mesmo assunto ou um único tópico é analisada para gerar um único resumo [18]. No geral, sumarização de múltiplos documentos tendem a ser mais genéricas do que focadas em perguntas dos usuários [18].

2.4 ETAPAS DA SUMARIZAÇÃO EXTRATIVA

Os métodos de sumarização extrativa, geralmente, são realizados em três etapas: (i) seleção das sentenças através da criação de uma representação intermediária do texto original; (ii) ordenação das sentenças a partir de suas pontuações; e (iii) elaboração do sumário a partir das sentenças de maior pontuação [19].

O primeiro dos passos acima cria uma representação de um documento. Normalmente, o texto é dividido em parágrafos, sentenças e *tokens*. Nesse passo também é provável que seja realizado algum pré-processamento como, por exemplo, remoção de *stopwords*.

O segundo dos passos tenta determinar quais as sentenças são mais importantes para o documento, através das suas pontuações, ou em que medida são combinadas informações sobre diferentes tópicos por pontuação das sentenças (o quão relevante é uma sentença para a compreensão do texto como um todo) [20, 8]. Geralmente os sistemas determinam quais as

sentenças mais representativas do conteúdo de um determinado documento podendo utilizar as três seguintes abordagens: pontuação baseada em palavras - atribuindo pontuações às palavras mais importantes do documento; pontuação baseada em sentenças - verificando características de sentenças, como sua posição no documento, semelhança com o título, etc.; pontuação baseada em grafos - analisando a relação entre as sentenças [21, 20].

O último passo apenas combina as partituras fornecidas pelo passo anterior e gera um extrato [8].

2.5 AVALIAÇÃO DE SUMÁRIOS

Avaliar a qualidade dos sumários produzidos é algo muito importante e ao mesmo tempo muito complexo. A importância da avaliação da qualidade serve para selecionar os melhores sumarizadores para um determinado contexto e, também, para observar a evolução das técnicas de sumarização [22].

A complexidade inerente a atividade de sumarização reside no fato de que não há como se chegar em um resumo ideal, uma vez que há um significativo grau de discordância na elaboração, e, consequentemente, na avaliação humana de sumários. Nesse sentido, essa avaliação pode envolver uma diversidade de fatores como legibilidade, coerência, concisão, conteúdo, etc. Além disso, sumarizar determinados corpus de textos demandaria muito esforço e tempo. Portanto, avaliadores automáticos de resumos têm chamado muito a atenção da comunidade de pesquisa de sumarização de textos [23].

Uma ferramenta muito utilizada para o propósito da avaliação automática de resumos é o ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*). O ROUGE avalia automaticamente a qualidade de um resumo comparando-o com outros resumos (ideais) criados por seres humanos. Dessa forma ele consegue eliminar a subjetividade humana do processo de avaliação [23, 22].

CAPÍTULO 3 - METODOLOGIA

3.1 DESCRIÇÃO GERAL

A metodologia utilizada neste trabalho consiste em sumarizar todos os documentos de um *corpus* textual, previamente classificado em um conjunto de classes, com cada um dos métodos de sumarização extrativa selecionados para este trabalho e avaliar a qualidade de todos os extratos gerados, identificando os métodos mais eficientes para cada classe. Essa metodologia será executada através da implementação de um sistema de sumarização automática de textos baseado nas classes dos documentos que realizará todo esse processo de forma automatizada.

O sistema dispõe de um conjunto de técnicas de sumarização extrativa e visa correlacionar cada classe de documentos com a melhor técnica de sumarização dentre as implementadas. Para avaliar da melhor forma os extratos gerados pelo sistema proposto, foi utilizado um grande *corpus* de texto, um *software* de avaliação automática de extratos e o número de equivalências das sentenças dos extratos gerados com as sentenças dos respectivos documentos modelo, estes últimos, presentes, também, no *corpus* textual.

Os documentos presentes no *corpus* resumem-se a textos de notícias (textos de cunho informativo escritos em linguagem formal e bem estruturada) e seus referentes modelos, aqui também chamados de padrões-ouro, devidamente classificados em um conjunto de classes. O *software* de avaliação automática indica quantitativamente o quão bom é um extrato, ou melhor, o quanto o extrato gerado automaticamente se aproxima do seu padrão-ouro. O número de correspondências entre as sentenças do extrato gerado e do seu respectivo padrão-ouro também é contabilizado como uma medida de qualidade para a seleção do método de sumarização mais eficiente, isto é, o número de sentenças presentes simultaneamente em ambos os documentos consiste em outra importante medida de avaliação do sistema.

3.2 MÉTODOS DE SUMARIZAÇÃO

Os métodos de sumarização utilizados neste trabalho, e aqui exibidos, seguem as três seguintes abordagens: pontuação baseada em palavras, pontuação baseada em sentenças e pontuação baseada em grafos [8, 21, 20]. Os primeiros métodos a serem exibidos consistem em pontuação baseada em palavras, onde cada palavra recebe uma pontuação e a pontuação total de cada sentença é a soma de todas as pontuações de suas palavras constituintes [8, 4].

- **Frequência de Palavra (*Word Frequency*)**: como o próprio nome do método sugere, quanto mais frequentemente uma palavra ocorrer no texto, maior será o fator de importância da mesma [8, 4, 24];
- **Frequência do Termo - Inverso da Frequência nos Documentos (*TF/IDF*)**: quanto mais frequente é um termo em um documento e quanto mais raro ele é dentro de um corpus de documentos, simultaneamente, maior será a pontuação desse termo [8, 4, 24];
- **Letras maiúsculas (*Upper Case*)**: atribui pontuações mais altas a palavras que contêm uma ou mais letras maiúsculas [8, 4, 24];
- **Nomes Próprios (*Proper Noun*)**: pressupõe que as sentenças que contêm um maior número de substantivos próprios são, possivelmente, mais importantes do que outras [8, 4, 24];
- **Co-ocorrência de palavras (*Word Co-occurrence*)**: mede a probabilidade de duas palavras ou mais palavras ocorrerem em sequência ordenada (N-grams) em um texto [8, 4, 24];
- **Similaridade léxica (*Lexical Similarity*)**: baseia-se no pressuposto de que as sentenças importantes são identificadas por cadeias de palavras fortemente similares (tanto na sua forma quanto no seu significado), de forma que quanto maior a similaridade léxica entre duas sentenças, mais próximos serão os vocábulos de ambas [8, 4, 24].

Na abordagem de pontuação baseada sentenças são analisadas as características da própria sentença, como a presença de palavras sinalizadoras. Os métodos mais importantes que seguem esta ideia são descritos abaixo:

- **Semelhança de frases com o título (*Sentence Resemblance to the Title*)**: sobreposição de vocabulário entre uma frase e o título do documento de forma que quanto mais similar seja a frase do título, mais pontos a sentença terá [8, 4, 24];
- **Centralidade da Sentença (*Sentence Centrality*)**: a centralidade da sentença é a sobreposição de vocabulário entre uma sentença e outras sentenças no documento de forma que quanto mais a sentença fornece uma quantidade necessária e suficiente de informações relacionadas ao tema principal, mais central ela será [8, 4, 24];
- **Posição da Sentença (*Sentence Position*)**: a posição da sentença, em geral, tem influência na sua importância. Por exemplo, as frases mais importantes tendem a aparecer no início de um documento [8, 4, 24];
- **Palavras Sinalizadores (*Cue-Phrases*)**: atribui maior peso às sentenças iniciadas por bons indicadores de conteúdo como “resumindo”, “concluindo", “o mais importante”, etc. [8, 4, 24];

- **Inclusão de Dados Numéricos em Frases (*Sentence Inclusion of Numerical Data*)**: geralmente a frase que contém dados numéricos é considerada importante e é muito provável que seja incluída no resumo do documento [8, 4, 24];
- **Comprimento da Sentença (*Sentence Length*)**: este método é muitas vezes utilizado para penalizar frases que são muito curtas ou longas, geralmente em sumarizadores que o usuário especifica o tamanho máximo do resumo [8, 4, 24].

Nos métodos cujas pontuações são baseadas em grafos, a pontuação é gerada pela relação entre sentenças do texto. Quando uma sentença se refere a outra, ela gera uma ligação com um peso associado entre elas. Os pesos são usados para gerar as pontuações de uma frase. Abaixo, uma lista dos três principais métodos desse tipo de abordagem:

- **Classificação de Texto (*Text Rank*)**: extrai as palavras-chave importantes de um documento de texto e também determina o peso da "importância" dessas palavras dentro de todo o documento usando um modelo baseado em grafos para identificar as dependências da mesma [8, 4, 17, 24];
- **Caminho espesso do nó (*Bushy Path of the Node*)**: o caminho espesso de um nó (sentença) em um mapa é definido como o número de ligações que o ligam a outros nós no mapa [8, 4, 24];
- **Similaridade agregada (*Aggregate Similarity*)**: ao invés de contar o número de ligações que ligam um nó (sentença) a outros nós (caminho espesso do nó), a similaridade agregada soma os pesos (similaridades) das ligações [8, 4, 24].

3.3 CORPUS DE TEXTO UTILIZADO NOS EXPERIMENTOS

O *corpus* de texto aqui utilizado foi o da CNN (*Cable News Network*) [25], desenvolvido pelo professor orientador deste trabalho e seus colaboradores em no projeto de Pesquisa e Desenvolvimento Tecnológico chamado FLIP, em parceria com a empresa *Hewlett-Packard* a partir de textos do site da CNN (*www.cnn.com*). O corpus CNN na sua versão atual consiste de 3.000 textos em inglês atribuídos.

Os textos estavam originalmente classificados em 11 classes distintas: "África", "América Latina", "Ásia", "Esportes", "Estados Unidos da América", "Europa", "Negócios", "Oriente Médio", "Tecnologia", "Viagens" e "Mundo". Tal classificação foi considerada inadequada uma vez que a classe "Mundo", por exemplo, era demasiadamente genérica.

Uma nova ontologia foi desenvolvida pelo grupo do projeto FLIP baseada nas classificações da *Google* e *Yahoo*. Cada um dos 3.000 textos do corpus CNN foi automaticamente reclassificado segundo a nova ontologia desenvolvida em até três classes, em ordem decrescente de importância, sendo tal classificação verificada por três especialistas.

As classes da nova ontologia que foram utilizadas neste trabalho são: “Alimentos e Bebidas” (“*Food and Beverage*”), “Artes” (“*Arts*”), “Automóveis” (“*Automotive*”), “Ciência” (“*Science*”), “Computação e Internet” (“*Computer and Internet*”), “Economia e Finanças” (“*Economy and Finance*”), “Educação” (“*Education*”), “Emprego e Trabalho” (“*Employment and Work*”), “Entretenimento” (“*Entertainment*”), “Esportes” (“*Sports*”), “Jogos” (“Games”), “Governo e Política” (“*Government and Politics*”), “Lei” (“*Law*”), “Negócios” (“*Business*”), “Saúde” (“*Health*”), “Sociedade e Cultura” (“*Society and Culture*”), “Viagens” (“*Travel*”), “Tempo” (“*Weather*”).

Tabela 1 - Relação entre o número de documentos por classe

Nome da Classe	Nº de Documentos
Alimentos e Bebidas	22
Artes	274
Automóveis	19
Ciência	128
Computação e Internet	57
Economia e Finanças	33
Educação	101
Emprego e Trabalho	69
Entretenimento	420
Esportes	61
Governo e Política	446
Jogos	27
Lei	206
Negócios	143
Saúde	119
Sociedade e Cultura	833
Tempo	16
Viagens	26
TOTAL	**3.000**

Fonte: Próprio autor

Devido ao conteúdo das notícias, seria muito improvável que todas fossem enquadradas em apenas uma única classe, dentre as 18 existentes, de forma unânime por todos os especialistas. Justamente, por esse motivo, cada texto do corpus foi classificado em,

no mínimo, uma classe, e, no máximo, em três classes. A primeira classe de classificação é mais importante que a segunda e a terceira classe, assim como, a segunda é mais importante que a terceira.

A notícia "*E-book lending: Your public library's best kept secret?*" [26], por exemplo, tem como sua primeira classe de classificação "Computação e Internet", como sua segunda classe "Entretenimento" e não apresenta uma terceira classe, visto que tal notícia não se enquadrou em mais nenhuma outra categoria existente. Essa classificação em três classes, também indica a classe predominante de um determinado documento. No exemplo da notícia citada, então, a sua classe principal é "Computação e Internet". Para efeitos práticos neste trabalho, só foi considerada a primeira classe, a classe mais importante, dentre as três existentes para cada documento do *corpus*.

A quantidade de documentos por classe da nova ontologia não são uniformes, há algumas classes que possuem um número de documentos bem maiores do que outras. A relação do número de documentos por classe é ilustrada na Tabela 1.

Quadro 1 - Sentenças do padrão-ouro referente ao texto "*E-book lending: Your public library's best kept secret?*".

Nº da Sentença no Texto Original	Sentença
2	Still, well over half of U.S. library card holders don't know whether their local public library lends e-books, according to a new Pew report.
7	Earlier this year, Pew research found that about one in five U.S. adults have read an e-book in the past year, and that e-book users tend to read over 30% more books per year than people who only read printed material.
8	If e-books encourage people to read more, that's good for public libraries -- which have a core mission to foster literacy and community engagement with information, culture and civic life.
25	But evidence indicates that publishers may not have much to fear from library e-book lending -- just like they haven't been hurt by library lending of print editions of current bestsellers.

Fonte: Próprio autor

Quadro 2 - Sentenças do documento de destaques referente ao texto "*E-book lending: Your public library's best kept secret?*".

Nº da Sentença no Destaque	Sentença
1	Many Americans don't know whether their public library has e-books, a new Pew report says.
2	Those who use e-books read 30% more books per year than those who don't.
3	If e-books encourage people to read more, that's good for public libraries.
4	However, library e-book acquisition has been complicated by higher publisher prices.

Fonte: Próprio autor

Para ilustrar melhor a estrutura dos padrões-ouro e dos destaques, os Quadros 1 e 2 exibem as sentenças presentes no padrão-ouro e nos destaques, respectivamente, referentes ao texto "*E-book lending: Your public library's best kept secret?*" [26].

A referência [27] apresenta uma análise preliminar da relação entre técnicas de sumarização automática e classificação de textos, porém como foi aí concluído, devido à extrema generalidade da classificação original do corpus CNN, não foi possível realizar uma correlação clara entre o conteúdo do documento (sua classificação) e as técnicas mais eficientes para sua sumarização. Apenas nas classes “Negócios” e “Esportes” ficou evidenciada uma correlação da maior adequabilidade de algumas técnicas de sumarização com o tipo de documento.

3.4 AVALIAÇÃO DOS EXTRATOS GERADOS PELO SISTEMA

Uma primeira avaliação quantitativa foi realizada utilizando o ROUGE, mais especificamente, a medida de avaliação ROUGE-2, uma vez que o mesmo consegue avaliar automaticamente a qualidade de um resumo comparando-o com outros resumos (ideais) criados por seres humanos, geralmente. Dessa forma foi possível eliminar a subjetividade humana do processo de avaliação [23] e comparar numericamente a qualidade de um resumo. Os resultados da medida do ROUGE utilizado neste trabalho são expressos através dos valores de Precisão, Cobertura e Medida F (*F-measure*).

Devido à existência de padrões-ouro elaborados com base nos resumos dos próprios autores foi possível realizar, também, uma avaliação qualitativa dos resumos produzidos pelo sistema. Essa avaliação consiste no número de casamentos entre as frases dos extratos gerados com as frases dos padrões-ouro e constitui uma medida da qualidade dos resumos considerada neste trabalho. Essa medida será aqui designada por *MATCHES* e calculada através da Equação 3.4.1.

$$MATCHES = \frac{NCSS}{TSGS}$$

Equação 3.4.1

Onde:
NCSS é o número de sentenças presentes simultaneamente no sumário e no padrão-ouro.
TSGS é o número total de sentenças do padrão-ouro.

Por regra, foi adotado neste trabalho que o número de sentenças de cada sumário seria cinco, uma vez que todos os padrões-ouro apresentavam cinco sentenças no máximo. Entretanto, para as situações que os padrões-ouro apresentavam uma quantidade inferior de sentenças, os sumários, referentes aos mesmos textos originais, que apresentassem um número maior de sentenças não foram penalizados, isto é, caso um sumário contivesse todas as sentenças presentes no seu referente padrão-ouro e outras mais, a pontuação da medida *MATCHES*, nesse caso, seria máxima.

Para ilustrar melhor ambas as avaliações empregues neste trabalho, ROUGE-2 e *MATCHES*, os Quadros 3 e 4 exibem as sentenças presentes em dois sumários gerados pelos métodos *Word Frequency* e *Resemblance-Title*, respectivamente, referentes ao texto "*E-book lending: Your public library's best kept secret?*" [26].

Quadro 3 - Sentenças do sumário referente ao texto "*E-book lending: Your public library's best kept secret?*" gerado pelo método *Word Frequency*.

Nº da Sentença no Texto Original	Sentença
2	Still, well over half of U.S. library card holders don't know whether their local public library lends e-books, according to a new Pew report.
7	Earlier this year, Pew research found that about one in five U.S. adults have read an e-book in the past year, and that e-book users tend to read over 30% more books per year than people who only read printed material.
30	But, interestingly, slightly more library e-books borrowers told Pew that when they want to find an e-book, the first place they look is an online bookstore -- not the library.
36	Pew asked people who don't already borrow library e-books whether they might avail themselves of library programs that would prepare them for using e-books.
45	You'll need a library card, of course -- so if you don't already have one (more than 40% of Americans don't), borrowing free e-books might be a good reason to finally visit your public library and get a library card.

Fonte: Próprio autor

Quadro 4 - Sentenças do sumário referente ao texto "*E-book lending: Your public library's best kept secret?*" gerado pelo método *Resemblance-Title*.

Nº da Sentença	Sentença
1	About three quarters of American public libraries currently lend out e-books, and in the past year libraries have seen a sharp growth in e-book borrowing.
3	Pew also found that 12% of all Americans age 16 and older who read e-books have borrowed an e-book from a library in the past year -- and they're generally pretty happy with the experience.
16	Even though libraries have options to get many e-books for free, lending out many of the most popular titles definitely costs money.
32	Last week, Penguin announced a pilot program to allow public libraries in New York City and Brooklyn to lend out its e-books via 3M's Cloud Drive service.
45	You'll need a library card, of course -- so if you don't already have one (more than 40% of Americans don't), borrowing free e-books might be a good reason to finally visit your public library and get a library card.

Fonte: Próprio autor

Comparando as sentenças presentes no sumário exibido no Quadro 3 com as sentenças presentes no padrão-ouro exibido no Quadro 1, é possível observar que as sentenças 2 e 7 estão presentes em ambos os documentos. Prontamente, comparando agora as sentenças presentes no sumário exibido no Quadro 4 novamente com as sentenças

expostas no Quadro 1, é possível observar que nenhuma sentença está presente em ambos os documentos, nesse caso.

Os valores comparativos entre os resultados das avaliações dos sumários exibidos nos Quadros 3 e 4 são indicados na Tabela 2. Pode-se perceber que o sumário produzido pelo método *Word Frequency* obteve um desempenho melhor que o sumário produzido pelo método *Resemblance-Title* na medida MATCHES e em todos os valores da medida ROUGE-2.

Tabela 2 - Tabela comparativa dos resultados das avaliações dos sumários gerados pelos métodos *Word Frequency* e *Resemblance-Title* referente ao texto "*E-book lending: Your public library's best kept secret?*" gerado pelo método *Resemblance-Title*.

Método de Sumarização	MATCHES	Cobertura ROUGE-2	Precisão ROUGE-2	Medida F ROUGE-2
Word Frequency	0,5000	0,5489	0,4269	0,4803
Sentence Resemblance to the Title	0,0000	0,1053	0,0886	0,0962

Fonte: Próprio autor

Como neste trabalho foram processados um conjunto de documentos por classe, para cada grupamento de valores de medidas, foram calculados os valores da média (*AVG*), do desvio padrão amostral (*SD*) e do coeficiente de variação de Pearson (*CV*) através das Equações 3.4.2, 3.4.3 e 3.4.4, respectivamente.

$$AVG = \frac{\sum_{i=1}^{n} x_i}{n}$$

Equação 3.4.2

$$SD = \sqrt{\frac{\sum_{i=1}^{n} (x_i - AVG)^2}{n-1}}$$

Equação 3.4.3

$$CV = \frac{SD}{AVG}$$

Equação 3.4.4

Onde:
x_i é o valor de cada amostra.
n é o número de amostras.

Os métodos de sumarização mais eficientes para cada classe de documentos foram obtidos considerando os maiores valores médios da Medida F do ROUGE-2 e os maiores valores médios da medida MATCHES. Para ambos os casos, foram considerados apenas os métodos com os menores coeficientes de variação.

CAPÍTULO 4 - RESULTADOS E DISCUSSÃO

Aplicando a metodologia descrita no capítulo anterior e condensando os dados obtidos com as experimentações, foram concebidas as Tabelas 3 e 4 que expõem os métodos cujos extratos gerados demonstraram os melhores resultados perante as medidas de avaliação ROUGE-2 e MATCHES, nesta ordem. No caso da Tabela 3, os melhores métodos foram selecionados pelo maior valor médio da Medida F. Todos os valores médios das medidas MATCHES e ROUGE-2 obtidos por classe para cada método de sumarização extrativa encontram-se do Apêndice A ao Apêndice R.

Analisando cuidadosamente os resultados obtidos, pode-se verificar que eles realmente indicam que a classificação não está correlacionada com a técnica mais eficiente de sumarização, posto que os valores obtidos pela medida do ROUGE-2 ficaram muito abaixo do estado da arte.

Tabela 3 - Tabela comparativa com os melhores métodos de sumarização extrativa por classe segundo os valores da medida ROUGE-2.

Classe	Método de Sumarização mais Eficiente	Cobertura ROUGE-2	Precisão ROUGE-2	Medida F ROUGE-2
Alimentos e Bebidas	*Sentence Resemblance to the Title*	0,5095	0,3905	0,4395
Artes	*TF/IDF*	0,4776	0,2693	0,3382
Automóveis	*Sentence Resemblance to the Title*	0,4650	0,2676	0,3325
Ciência	*Word Frequency*	0,4534	0,2522	0,3196
Computação e Internet	*TF/IDF*	0,4486	0,2377	0,3057
Economia e Finanças	*Word Frequency*	0,5425	0,2915	0,3739
Educação	*TF/IDF*	0,4552	0,2451	0,3138
Emprego e Trabalho	*Word Frequency*	0,5578	0,3095	0,3926
Entretenimento	*Word Frequency*	0,4491	0,2573	0,3219
Esportes	*TF/IDF*	0,5954	0,3663	0,4446
Governo e Política	*Word Frequency*	0,4826	0,2754	0,3457
Jogos	*Proper Noun*	0,5690	0,3695	0,4392
Lei	*TF/IDF*	0,5324	0,2949	0,3738
Negócios	*Word Frequency*	0,4549	0,2563	0,3234
Saúde	*Word Frequency*	0,4261	0,2275	0,2917
Sociedade e Cultura	*TF/IDF*	0,4553	0,2576	0,3244
Tempo	*TF/IDF*	0,4918	0,2832	0,3533
Viagens	*TF/IDF*	0,3257	0,1825	0,2313

Fonte: Próprio autor

Na maioria das classes os métodos *Word Frequency* e *TF/IDF* produziram, em média, os melhores sumários. No caso da classe “Economia”, por exemplo, talvez intuitivamente era de se esperar que o método *Sentence Inclusion of Numerical Data* mostrasse uma um desempenho melhor, entretanto o método *Word Frequency* para ambas as medidas de avaliação.

Tabela 4 - Tabela comparativa com os melhores métodos de sumarização extrativa por classe segundo os valores da medida MATCHES.

Classe	Método de Sumarização mais Eficiente	MATCHES
Alimentos e Bebidas	*Sentence Resemblance to the Title*	0,4061
Artes	*TF/IDF*	0,3714
Automóveis	*Sentence Resemblance to the Title*	0,4263
Ciência	*Word Frequency*	0,3531
Computação e Internet	*Upper Case*	0,3424
Economia e Finanças	*Word Frequency*	0,4242
Educação	*Word Frequency*	0,3657
Emprego e Trabalho	*Word Frequency*	0,4300
Entretenimento	*Word Frequency*	0,3556
Esportes	*Word Frequency*	0,5005
Governo e Política	*Word Frequency*	0,3820
Jogos	*Proper Noun*	0,5007
Lei	*Word Frequency*	0,4238
Negócios	*Word Frequency*	0,3562
Saúde	*Word Frequency*	0,3172
Sociedade e Cultura	*TF/IDF*	0,3403
Tempo	*TF/IDF*	0,3729
Viagens	*TF/IDF*	0,2410

Fonte: Próprio autor

Salvo poucas exceções, no geral, quase todos os métodos que apresentaram a melhor eficiência para uma determinada classe na medida ROUGE-2, apresentavam, também a melhor eficiência para a medida MATCHES. O Quadro 5 exibe um comparativo dos melhores métodos para ambas as medidas de avaliação dos resumos gerados.

Quadro 5 - Quadro comparativo com os melhores métodos de sumarização extrativa por classe segundo os valores da medida ROUGE-2 e os valores da medida MATCHES.

Classe	Método de Sumarização mais Eficiente segundo o valor da Medida F do ROUGE-2	Método de Sumarização mais Eficiente segundo a medida MATCHES
Alimentos e Bebidas	*Sentence Resemblance to the Title*	*Sentence Resemblance to the Title*
Artes	*TF/IDF*	*TF/IDF*
Automóveis	*Sentence Resemblance to the Title*	*Sentence Resemblance to the Title*
Ciência	*Word Frequency*	*Word Frequency*
Computação e Internet	*TF/IDF*	*Upper Case*
Economia e Finanças	*Word Frequency*	*Word Frequency*
Educação	*TF/IDF*	*Word Frequency*
Emprego e Trabalho	*Word Frequency*	*Word Frequency*
Entretenimento	*Word Frequency*	*Word Frequency*
Esportes	*TF/IDF*	*Word Frequency*
Governo e Política	*Word Frequency*	*Word Frequency*
Jogos	*Proper Noun*	*Proper Noun*
Lei	*TF/IDF*	*Word Frequency*
Negócios	*Word Frequency*	*Word Frequency*
Saúde	*Word Frequency*	*Word Frequency*
Sociedade e Cultura	*TF/IDF*	*TF/IDF*
Tempo	*TF/IDF*	*TF/IDF*
Viagens	*TF/IDF*	*TF/IDF*

Fonte: Próprio autor

Examinando o Quadro 5 é possível observar que apenas para as classes “Computação e Internet”, “Educação”, “Esportes” e “Lei” os métodos de sumarização extrativa mais eficientes para os documentos da classe, apontados pela medida ROUGE-2 e pela medida MATCHES são diferentes. O Quadro 6 é bem semelhante ao Quadro 5, exibindo a mesma informação, entretanto a agrupando pelos métodos de sumarização extrativa ao invés das classes.

Quadro 6 - Quadro comparativo com as classes mais adequadas para cada método de sumarização extrativa segundo os valores da medida ROUGE-2 e os valores da medida MATCHES.

Método de Sumarização	Classes mais adequadas segundo o valor da Medida F do ROUGE-2	Classes mais adequadas segundo a medida MATCHES
Proper Noun	“Jogos”	“Jogos”
Sentence Resemblance to the Title	“Alimentos e Bebidas”, “Automóveis”	“Alimentos e Bebidas”, “Automóveis”
TF/IDF	“Artes”, “Computação e Internet”, “Educação”, “Esportes”, “Lei”, “Sociedade e Cultura”, “Tempo”, “Viagens”	“Artes, ”Sociedade e Cultura”, “Tempo”, “Viagens”
Upper Case	-	“Computação e Internet”
Word Frequency	“Ciência”, “Economia e Finanças”, “Emprego e Trabalho”, “Entretenimento”, “Governo e Política”, “Negócios”, “Saúde”	“Ciência”, “Economia e Finanças”, “Educação”, “Emprego e Trabalho” “Entretenimento”, “Esportes”, “Governo e Política”, “Lei”, “Negócios”, “Saúde”

Fonte: Próprio autor

CAPÍTULO 5 - CONSIDERAÇÕES FINAIS

5.1 CONTRIBUIÇÕES

O projeto aqui descrito se propôs a verificar o quanto a classificação de um documento consistia em um bom critério para a escolha das técnicas de sumarização extrativa mais eficientes, uma vez que é muito difícil elaborar um método genérico para resumir diferentes tipos de textos [8]. Além disso, no caso de um possível cenário confirmativo, também seria um propósito deste trabalho, identificar quais as técnicas produzem resumos de melhor qualidade para cada uma das classes de documentos empregadas.

Contudo, pela análise dos resultados obtidos, utilizar apenas a classificação dos textos não é uma boa estratégia para a escolha das técnicas mais eficientes de sumarização automática. Talvez para a escolha das combinações mais eficientes desses algoritmos, a utilização das classes de classificação dos textos produza resultados melhores, ou seja, embora alguns métodos de sumarização isoladamente não ofereceram bons resultados, tem-se que analisar a possibilidade da junção de técnicas oferecer uma maior cobertura. Infelizmente, gerar as combinações entre todos os métodos não foi o escopo deste trabalho, mas, certamente isso é um bom direcionamento para trabalhos futuros.

A principal contribuição deste trabalho foi mostrar experimentalmente que a apenas a classificação de um conjunto de documentos de textos não configura um bom parâmetro para a escolha do método de sumarização a ser empregue para todos os documentos da classe. Alguns fatores corroboram para esse desfecho como o desbalanceamento da quantidade de documentos das classes do corpus CNN, isto é, enquanto a classe “Sociedade e Cultura” possui mais de 800 documentos, a classe “Alimentos e Bebidas” possui pouco mais que 20. A seleção de documentos utilizando apenas uma das três classificações da nova ontologia proposta [25] também agravou esse desbalanceamento para algumas das classes.

5.2 TRABALHOS FUTUROS

Como já declarado anteriormente, analisar as mais diversas combinações de técnicas de sumarização para cada classe de documento é um possível direcionamento futuro, visto que duas técnicas que não se mostraram como uma das melhores, possivelmente combinadas podem superar os resultados obtidos por outras técnicas isoladamente. Nesse específico caso, a escolha de tais técnicas pode estar correlacionada com a classe dos documentos.

REFERÊNCIAS

[1] GOLDSTEIN, Jade et al. **Summarizing text documents: sentence selection and evaluation metrics**. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999. p. 121-128.

[2] WANG, Sicui et al. **A survey on automatic summarization**. In: Information Technology and Applications (IFITA), 2010 International Forum on. IEEE, 2010. p. 193-196.

[3] GAMBHIR, Mahak; GUPTA, Vishal. **Recent automatic text summarization techniques: a survey.** Artificial Intelligence Review, v. 47, n. 1, p. 1-66, 2017.

[4] FERREIRA, Rafael et al. **Automatic Document Classification using Summarization Strategies**. In: Proceedings of the 2015 ACM Symposium on Document Engineering. ACM, 2015. p. 69-72.

[5] SHEN, Dou et al. **Web-page classification through summarization**. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2004. p. 242-249.

[6] DALAL, Mita K.; ZAVERI, Mukesh A. **Semisupervised learning based opinion summarization and classification for online product reviews**. Applied Computational Intelligence and Soft Computing, v. 2013, p. 10, 2013.

[7] MIHALCEA, Rada; HASSAN, Samer. **Using the essence of texts to improve document classification**. In: Proceedings of RANLP. 2005.

[8] FERREIRA, Rafael et al. **A context based text summarization system**. In: Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on. IEEE, 2014. p. 66-70.

[9] MARTINS, Camilla Brandel et al. **Introdução à sumarização automática**. Relatório Técnico RT-DC, v. 2, p. 2001, 2001.

[10] LUHN, Hans Peter. **The automatic creation of literature abstracts**. IBM Journal of research and development, v. 2, n. 2, p. 159-165, 1958.

[11] MANI, Inderjeet. **Summarization evaluation: An overview**. 2001.

[12] MANI, Inderjeet et al. **SUMMAC: a text summarization evaluation**. Natural Language Engineering, v. 8, n. 1, p. 43-68, 2002.

[13] BHATIA, Neelima; JAISWAL, Arunima. **Automatic text summarization and it's methods-a review**. In: Cloud System and Big Data Engineering (Confluence), 2016 6th International Conference. IEEE, 2016. p. 65-72.

[14] ORASAN, Constantin; HASLER, Laura. **Computer-aided summarisation–what the user really wants**. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006). 2006. p. 1548-1551.

[15] CABRAL, Luciano de Souza et al. **A platform for language independent summarization**. In: Proceedings of the 2014 ACM symposium on Document engineering. ACM, 2014. p. 203-206.

[16] PAICE, Chris D. **The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases**. In: Proceedings of the 3rd annual ACM conference on Research and development in information retrieval. Butterworth & Co., 1980. p. 172-191.

[17] FERREIRA, Rafael et al. **A four dimension graph model for automatic text summarization**. In: Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 01. IEEE Computer Society, 2013. p. 389-396.

[18] FERREIRA, Rafael et al. **A multi-document summarization system based on statistics and linguistic treatment**. Expert Systems with Applications, v. 41, n. 13, p. 5780-5787, 2014.

[19] NENKOVA, Ani; MCKEOWN, Kathleen. **A survey of text summarization techniques**. Mining text data, p. 43-76, 2012.

[20] FERREIRA, Rafael et al. **Assessing sentence scoring techniques for extractive text summarization**. Expert systems with applications, v. 40, n. 14, p. 5755-5764, 2013.

[21] FERREIRA, Rafael et al. **A new sentence similarity assessment measure based on a three-layer sentence representation**. In: Proceedings of the 2014 ACM symposium on Document engineering. ACM, 2014. p. 25-34.

[22] SANTOS, Ângelo Filipe da Silva dos. **Sumarização automática de texto**. 2012. Tese de Doutorado.

[23] LIN, Chin-Yew. **ROUGE: A package for automatic evaluation of summaries**. In: Text summarization branches out: Proceedings of the ACL-04 workshop. 2004.

[24] MEENA, Yogesh Kumar; GOPALANI, Dinesh. **Analysis of sentence scoring methods for extractive automatic text summarization**. In: Proceedings of the 2014 International Conference on Information and Communication Technology for Competitive Strategies. ACM, 2014. p. 53.

[25] LINS, Rafael Dueire et al. **A multi-tool scheme for summarizing textual documents**. In: Proc. of 11st IADIS International Conference WWW/INTERNET 2012. 2012. p. 1-8.

[26] **E-book lending: Your public library's best kept secret?**. Disponível em: <http://edition.cnn.com/2012/06/26/tech/web/ebook-lending-pew/index.html>. Acesso em: 01 de Julho de 2017.

[27] FERREIRA, Rafael et al. **Automatic Document Classification using Summarization Strategies**. In: Proceedings of the 2015 ACM Symposium on Document Engineering. ACM, 2015. p. 69-72.

APÊNDICE A - Tabela com os resultados dos métodos de sumarização extrativa para a classe “Alimentos e Bebidas”.

Método de Sumarização	MATCHES	Medida F ROUGE-2	Precisão ROUGE-2	Cobertura ROUGE-2
Aggregate Similarity	0,2841	0,2753	0,2457	0,3204
Bushy Path of the Node	0,2386	0,2441	0,2213	0,2769
Cue-Phrases	0,2083	0,2294	0,2310	0,2383
Lexical Similarity	0,1545	0,1644	0,1528	0,1832
Proper Noun	0,2636	0,2849	0,2594	0,3267
Sentence Centrality	0,1492	0,1380	0,1466	0,1453
Sentence Inclusion of Numerical Data	0,3023	0,3161	0,2919	0,3615
Sentence Length	0,3424	0,3241	0,2613	0,4350
Sentence Position	0,2773	0,2937	0,2771	0,3235
Sentence Resemblance to the Title	0,4061	0,4395	0,3905	0,5095
Text Rank	0,2083	0,2294	0,2310	0,2383
TF/IDF	0,3553	0,3617	0,2945	0,4785
Upper Case	0,2773	0,3039	0,2684	0,3613
Word Co-occurrence	0,2720	0,3266	0,2888	0,3864
Word Frequency	0,3780	0,3738	0,3102	0,4820

APÊNDICE B - Tabela com os resultados dos métodos de sumarização extrativa para a classe “Artes”.

Método de Sumarização	MATCHES	Medida F ROUGE-2	Precisão ROUGE-2	Cobertura ROUGE-2
Aggregate Similarity	0,1885	0,1829	0,1597	0,2289
Bushy Path of the Node	0,1956	0,1885	0,1664	0,2321
Cue-Phrases	0,1754	0,1727	0,1598	0,2006
Lexical Similarity	0,1305	0,1335	0,1277	0,1498
Proper Noun	0,2969	0,2790	0,2348	0,3629
Sentence Centrality	0,0904	0,0826	0,0881	0,0855
Sentence Inclusion of Numerical Data	0,2537	0,2399	0,2128	0,2937
Sentence Length	0,2706	0,2592	0,2036	0,3722
Sentence Position	0,2650	0,2627	0,2423	0,3079
Sentence Resemblance to the Title	0,2908	0,2719	0,2246	0,3622
Text Rank	0,1754	0,1727	0,1598	0,2006
TF/IDF	0,3714	0,3382	0,2693	0,4776
Upper Case	0,2818	0,2690	0,2245	0,3532
Word Co-occurrence	0,2016	0,2151	0,1889	0,2740
Word Frequency	0,3663	0,3299	0,2639	0,4630

APÊNDICE C - Tabela com os resultados dos métodos de sumarização extrativa para a classe “Automóveis”.

Método de Sumarização	MATCHES	Medida F ROUGE-2	Precisão ROUGE-2	Cobertura ROUGE-2
Aggregate Similarity	0,3561	0,2921	0,2400	0,3958
Bushy Path of the Node	0,3254	0,2966	0,2492	0,3855
Cue-Phrases	0,3404	0,2964	0,2559	0,3686
Lexical Similarity	0,2649	0,2278	0,1949	0,2916
Proper Noun	0,3474	0,2981	0,2414	0,4089
Sentence Centrality	0,1614	0,1441	0,1450	0,1520
Sentence Inclusion of Numerical Data	0,3781	0,3171	0,2657	0,4154
Sentence Length	0,2860	0,2564	0,2028	0,3677
Sentence Position	0,2649	0,2438	0,2113	0,3043
Sentence Resemblance to the Title	0,4263	0,3325	0,2676	0,4650
Text Rank	0,3404	0,2964	0,2559	0,3686
TF/IDF	0,3912	0,3217	0,2501	0,4802
Upper Case	0,3386	0,2962	0,2394	0,4053
Word Co-occurrence	0,1377	0,1498	0,1273	0,1955
Word Frequency	0,4000	0,3220	0,2528	0,4724

APÊNDICE D - Tabela com os resultados dos métodos de sumarização extrativa para a classe “Ciência”.

Método de Sumarização	MATCHES	Medida F ROUGE-2	Precisão ROUGE-2	Cobertura ROUGE-2
Aggregate Similarity	0,1890	0,1832	0,1607	0,2276
Bushy Path of the Node	0,1918	0,1856	0,1622	0,2310
Cue-Phrases	0,1553	0,1647	0,1537	0,1883
Lexical Similarity	0,1237	0,1305	0,1149	0,1599
Proper Noun	0,2507	0,2450	0,2038	0,3247
Sentence Centrality	0,0837	0,0868	0,0891	0,0915
Sentence Inclusion of Numerical Data	0,2099	0,2102	0,1851	0,2549
Sentence Length	0,2679	0,2523	0,1965	0,3647
Sentence Position	0,2277	0,2271	0,2053	0,2680
Sentence Resemblance to the Title	0,2903	0,2602	0,2129	0,3465
Text Rank	0,1553	0,1647	0,1537	0,1883
TF/IDF	0,3413	0,3129	0,2447	0,4501
Upper Case	0,2552	0,2466	0,2028	0,3326
Word Co-occurrence	0,1984	0,2160	0,1811	0,2819
Word Frequency	0,3531	0,3196	0,2522	0,4534

APÊNDICE E - Tabela com os resultados dos métodos de sumarização extrativa para a classe “Computação e Internet”.

Método de Sumarização	MATCHES	Medida F ROUGE-2	Precisão ROUGE-2	Cobertura ROUGE-2
Aggregate Similarity	0,2056	0,1973	0,1697	0,2500
Bushy Path of the Node	0,2260	0,2261	0,1958	0,2806
Cue-Phrases	0,1863	0,1962	0,1826	0,2215
Lexical Similarity	0,1228	0,1261	0,1120	0,1529
Proper Noun	0,3082	0,2749	0,2282	0,3650
Sentence Centrality	0,1088	0,1039	0,1068	0,1084
Sentence Inclusion of Numerical Data	0,2684	0,2542	0,2275	0,3045
Sentence Length	0,2871	0,2657	0,2064	0,3871
Sentence Position	0,2652	0,2566	0,2388	0,2968
Sentence Resemblance to the Title	0,3117	0,2840	0,2363	0,3720
Text Rank	0,1863	0,1962	0,1826	0,2215
TF/IDF	0,3418	0,3057	0,2377	0,4486
Upper Case	0,3424	0,3005	0,2492	0,4001
Word Co-occurrence	0,1971	0,2104	0,1766	0,2730
Word Frequency	0,3234	0,2989	0,2380	0,4194

APÊNDICE F - Tabela com os resultados dos métodos de sumarização extrativa para a classe “Economia e Finanças”.

Método de Sumarização	MATCHES	Medida F ROUGE-2	Precisão ROUGE-2	Cobertura ROUGE-2
Aggregate Similarity	0,2652	0,2589	0,2133	0,3405
Bushy Path of the Node	0,3005	0,2819	0,2377	0,3615
Cue-Phrases	0,2273	0,2156	0,1976	0,2516
Lexical Similarity	0,2222	0,2051	0,1754	0,2534
Proper Noun	0,3056	0,2828	0,2232	0,4023
Sentence Centrality	0,1515	0,1421	0,1432	0,1497
Sentence Inclusion of Numerical Data	0,2753	0,2597	0,2199	0,3300
Sentence Length	0,3611	0,3166	0,2452	0,4670
Sentence Position	0,2929	0,2799	0,2554	0,3273
Sentence Resemblance to the Title	0,2449	0,2236	0,1778	0,3124
Text Rank	0,2273	0,2156	0,1976	0,2516
TF/IDF	0,4167	0,3705	0,2884	0,5398
Upper Case	0,3232	0,2928	0,2302	0,4202
Word Co-occurrence	0,3485	0,3367	0,2826	0,4322
Word Frequency	0,4242	0,3739	0,2915	0,5425

APÊNDICE G - Tabela com os resultados dos métodos de sumarização extrativa para a classe “Educação”.

Método de Sumarização	MATCHES	Medida F ROUGE-2	Precisão ROUGE-2	Cobertura ROUGE-2
Aggregate Similarity	0,1881	0,1885	0,1599	0,2391
Bushy Path of the Node	0,1889	0,1853	0,1577	0,2350
Cue-Phrases	0,1777	0,1728	0,1562	0,2048
Lexical Similarity	0,1677	0,1570	0,1445	0,1864
Proper Noun	0,2493	0,2408	0,1993	0,3196
Sentence Centrality	0,1092	0,1089	0,1131	0,1126
Sentence Inclusion of Numerical Data	0,2779	0,2565	0,2280	0,3094
Sentence Length	0,2609	0,2520	0,1963	0,3682
Sentence Position	0,2690	0,2558	0,2261	0,3148
Sentence Resemblance to the Title	0,3234	0,2936	0,2384	0,4018
Text Rank	0,1777	0,1728	0,1562	0,2048
TF/IDF	0,3497	0,3138	0,2451	0,4552
Upper Case	0,2757	0,2635	0,2153	0,3560
Word Co-occurrence	0,2066	0,2260	0,1893	0,2942
Word Frequency	0,3657	0,3134	0,2464	0,4525

APÊNDICE H - Tabela com os resultados dos métodos de sumarização extrativa para a classe “Emprego e Trabalho”.

Método de Sumarização	MATCHES	Medida F ROUGE-2	Precisão ROUGE-2	Cobertura ROUGE-2
Aggregate Similarity	0,3031	0,2796	0,2365	0,3550
Bushy Path of the Node	0,3082	0,2966	0,2575	0,3645
Cue-Phrases	0,2899	0,2896	0,2626	0,3359
Lexical Similarity	0,1478	0,1467	0,1314	0,1735
Proper Noun	0,3461	0,3258	0,2686	0,4300
Sentence Centrality	0,1797	0,1496	0,1453	0,1618
Sentence Inclusion of Numerical Data	0,2990	0,2784	0,2421	0,3478
Sentence Length	0,3510	0,3239	0,2506	0,4757
Sentence Position	0,3159	0,2991	0,2636	0,3622
Sentence Resemblance to the Title	0,4077	0,3662	0,2985	0,4979
Text Rank	0,2899	0,2896	0,2626	0,3359
TF/IDF	0,4203	0,3817	0,2985	0,5509
Upper Case	0,3384	0,3245	0,2647	0,4371
Word Co-occurrence	0,2768	0,2737	0,2266	0,3675
Word Frequency	0,4300	0,3926	0,3095	0,5578

APÊNDICE I - Tabela com os resultados dos métodos de sumarização extrativa para a classe “Entretenimento”.

Método de Sumarização	MATCHES	Medida F ROUGE-2	Precisão ROUGE-2	Cobertura ROUGE-2
Aggregate Similarity	0,2025	0,1927	0,1663	0,2414
Bushy Path of the Node	0,2110	0,2030	0,1766	0,2511
Cue-Phrases	0,1818	0,1814	0,1692	0,2075
Lexical Similarity	0,1356	0,1406	0,1335	0,1588
Proper Noun	0,2908	0,2739	0,2287	0,3599
Sentence Centrality	0,1079	0,0996	0,1061	0,1016
Sentence Inclusion of Numerical Data	0,2487	0,2385	0,2122	0,2874
Sentence Length	0,2845	0,2669	0,2107	0,3808
Sentence Position	0,2733	0,2639	0,2430	0,3093
Sentence Resemblance to the Title	0,3191	0,2901	0,2411	0,3844
Text Rank	0,1818	0,1814	0,1692	0,2075
TF/IDF	0,3461	0,3166	0,2517	0,4449
Upper Case	0,2891	0,2697	0,2242	0,3567
Word Co-occurrence	0,2150	0,2245	0,1908	0,2908
Word Frequency	0,3556	0,3219	0,2573	0,4491

APÊNDICE J - Tabela com os resultados dos métodos de sumarização extrativa para a classe “Esportes”.

Método de Sumarização	MATCHES	Medida F ROUGE-2	Precisão ROUGE-2	Cobertura ROUGE-2
Aggregate Similarity	0,1929	0,1873	0,1750	0,2267
Bushy Path of the Node	0,2287	0,2201	0,2048	0,2665
Cue-Phrases	0,2470	0,2525	0,2353	0,2906
Lexical Similarity	0,1044	0,1154	0,1141	0,1250
Proper Noun	0,4063	0,3684	0,3137	0,4771
Sentence Centrality	0,0934	0,1022	0,1028	0,1121
Sentence Inclusion of Numerical Data	0,3383	0,3172	0,2823	0,3844
Sentence Length	0,3855	0,3627	0,2985	0,4811
Sentence Position	0,4030	0,3659	0,3379	0,4258
Sentence Resemblance to the Title	0,4751	0,4048	0,3476	0,5169
Text Rank	0,2470	0,2525	0,2353	0,2906
TF/IDF	0,4973	0,4446	0,3663	0,5954
Upper Case	0,3740	0,3504	0,2991	0,4497
Word Co-occurrence	0,2432	0,2605	0,2308	0,3202
Word Frequency	0,5005	0,4433	0,3684	0,5881

APÊNDICE K - Tabela com os resultados dos métodos de sumarização extrativa para a classe “Jogos”.

Método de Sumarização	MATCHES	Medida F ROUGE-2	Precisão ROUGE-2	Cobertura ROUGE-2
Aggregate Similarity	0,1815	0,1696	0,1530	0,2017
Bushy Path of the Node	0,2343	0,2125	0,1903	0,2561
Cue-Phrases	0,3028	0,2883	0,2676	0,3257
Lexical Similarity	0,1958	0,2050	0,2004	0,2249
Proper Noun	0,5007	0,4392	0,3695	0,5690
Sentence Centrality	0,1195	0,1351	0,1369	0,1380
Sentence Inclusion of Numerical Data	0,3948	0,3712	0,3260	0,4493
Sentence Length	0,3235	0,3063	0,2451	0,4280
Sentence Position	0,3495	0,3495	0,3243	0,4083
Sentence Resemblance to the Title	0,4100	0,3582	0,2918	0,4848
Text Rank	0,3028	0,2883	0,2676	0,3257
TF/IDF	0,4791	0,4246	0,3400	0,5930
Upper Case	0,4681	0,4153	0,3495	0,5373
Word Co-occurrence	0,2377	0,2330	0,2045	0,2856
Word Frequency	0,4637	0,4101	0,3295	0,5685

APÊNDICE L - Tabela com os resultados dos métodos de sumarização extrativa para a classe “Governo e Política”.

Método de Sumarização	MATCHES	Medida F ROUGE-2	Precisão ROUGE-2	Cobertura ROUGE-2
Aggregate Similarity	0,2260	0,2172	0,1869	0,2705
Bushy Path of the Node	0,2374	0,2265	0,1959	0,2810
Cue-Phrases	0,2183	0,2132	0,1956	0,2485
Lexical Similarity	0,1627	0,1642	0,1489	0,1944
Proper Noun	0,3333	0,3056	0,2523	0,4026
Sentence Centrality	0,1091	0,1012	0,1060	0,1041
Sentence Inclusion of Numerical Data	0,2668	0,2569	0,2256	0,3113
Sentence Length	0,3035	0,2868	0,2268	0,4051
Sentence Position	0,3120	0,3016	0,2724	0,3561
Sentence Resemblance to the Title	0,3744	0,3389	0,2786	0,4516
Text Rank	0,2183	0,2132	0,1956	0,2485
TF/IDF	0,3770	0,3444	0,2727	0,4848
Upper Case	0,3317	0,3086	0,2547	0,4073
Word Co-occurrence	0,2407	0,2520	0,2167	0,3199
Word Frequency	0,3820	0,3457	0,2754	0,4826

APÊNDICE M - Tabela com os resultados dos métodos de sumarização extrativa para a classe “Lei”.

Método de Sumarização	MATCHES	Medida F ROUGE-2	Precisão ROUGE-2	Cobertura ROUGE-2
Aggregate Similarity	0,2456	0,2321	0,1953	0,3000
Bushy Path of the Node	0,2388	0,2301	0,1946	0,2959
Cue-Phrases	0,2513	0,2421	0,2188	0,2833
Lexical Similarity	0,1826	0,1815	0,1679	0,2108
Proper Noun	0,3484	0,3255	0,2668	0,4363
Sentence Centrality	0,1225	0,1085	0,1137	0,1128
Sentence Inclusion of Numerical Data	0,3153	0,2924	0,2532	0,3645
Sentence Length	0,3479	0,3164	0,2499	0,4498
Sentence Position	0,3365	0,3197	0,2853	0,3859
Sentence Resemblance to the Title	0,3975	0,3557	0,2888	0,4863
Text Rank	0,2513	0,2421	0,2188	0,2833
TF/IDF	0,4198	0,3738	0,2949	0,5324
Upper Case	0,3494	0,3280	0,2686	0,4387
Word Co-occurrence	0,2772	0,2786	0,2326	0,3668
Word Frequency	0,4238	0,3719	0,2939	0,5275

APÊNDICE N - Tabela com os resultados dos métodos de sumarização extrativa para a classe “Negócios”.

Método de Sumarização	MATCHES	Medida F ROUGE-2	Precisão ROUGE-2	Cobertura ROUGE-2
Aggregate Similarity	0,2302	0,2223	0,1883	0,2816
Bushy Path of the Node	0,2382	0,2273	0,1955	0,2828
Cue-Phrases	0,2232	0,2183	0,1974	0,2559
Lexical Similarity	0,1272	0,1343	0,1223	0,1584
Proper Noun	0,2759	0,2630	0,2175	0,3462
Sentence Centrality	0,1117	0,1089	0,1110	0,1126
Sentence Inclusion of Numerical Data	0,2903	0,2645	0,2246	0,3333
Sentence Length	0,2752	0,2634	0,2059	0,3790
Sentence Position	0,2781	0,2732	0,2414	0,3285
Sentence Resemblance to the Title	0,3417	0,3061	0,2511	0,4096
Text Rank	0,2232	0,2183	0,1974	0,2559
TF/IDF	0,3529	0,3156	0,2478	0,4528
Upper Case	0,2491	0,2480	0,2045	0,3276
Word Co-occurrence	0,2032	0,2180	0,1837	0,2859
Word Frequency	0,3562	0,3234	0,2563	0,4549

APÊNDICE O - Tabela com os resultados dos métodos de sumarização extrativa para a classe “Saúde”.

Método de Sumarização	MATCHES	Medida F ROUGE-2	Precisão ROUGE-2	Cobertura ROUGE-2
Aggregate Similarity	0,1992	0,1934	0,1634	0,2471
Bushy Path of the Node	0,1898	0,1884	0,1616	0,2352
Cue-Phrases	0,1514	0,1616	0,1478	0,1867
Lexical Similarity	0,1268	0,1268	0,1080	0,1695
Proper Noun	0,2667	0,2558	0,2091	0,3462
Sentence Centrality	0,1004	0,0909	0,0944	0,0976
Sentence Inclusion of Numerical Data	0,2730	0,2583	0,2242	0,3182
Sentence Length	0,2552	0,2487	0,1927	0,3667
Sentence Position	0,2747	0,2558	0,2260	0,3102
Sentence Resemblance to the Title	0,2881	0,2718	0,2209	0,3732
Text Rank	0,1514	0,1616	0,1478	0,1867
TF/IDF	0,2975	0,2789	0,2166	0,4105
Upper Case	0,2524	0,2447	0,1986	0,3352
Word Co-occurrence	0,1842	0,1994	0,1649	0,2684
Word Frequency	0,3172	0,2917	0,2275	0,4261

APÊNDICE P - Tabela com os resultados dos métodos de sumarização extrativa para a classe “Sociedade e Cultura”.

Método de Sumarização	MATCHES	Medida F ROUGE-2	Precisão ROUGE-2	Cobertura ROUGE-2
Aggregate Similarity	0,1672	0,1745	0,1524	0,2141
Bushy Path of the Node	0,1743	0,1789	0,1567	0,2194
Cue-Phrases	0,1583	0,1674	0,1590	0,1872
Lexical Similarity	0,1225	0,1304	0,1223	0,1489
Proper Noun	0,2639	0,2622	0,2208	0,3362
Sentence Centrality	0,0785	0,0766	0,0855	0,0757
Sentence Inclusion of Numerical Data	0,2290	0,2273	0,2005	0,2756
Sentence Length	0,2546	0,2556	0,2019	0,3619
Sentence Position	0,2570	0,2581	0,2379	0,3002
Sentence Resemblance to the Title	0,2959	0,2865	0,2392	0,3746
Text Rank	0,1583	0,1674	0,1590	0,1872
TF/IDF	0,3403	0,3244	0,2576	0,4553
Upper Case	0,2603	0,2592	0,2171	0,3348
Word Co-occurrence	0,1967	0,2157	0,1834	0,2787
Word Frequency	0,3377	0,3224	0,2577	0,4474

APÊNDICE Q - Tabela com os resultados dos métodos de sumarização extrativa para a classe “Viagens”.

Método de Sumarização	MATCHES	Medida F ROUGE-2	Precisão ROUGE-2	Cobertura ROUGE-2
Aggregate Similarity	0,1769	0,1819	0,1551	0,2308
Bushy Path of the Node	0,2314	0,2211	0,1922	0,2687
Cue-Phrases	0,1372	0,1593	0,1545	0,1796
Lexical Similarity	0,0628	0,0881	0,0870	0,0931
Proper Noun	0,1801	0,1739	0,1520	0,2133
Sentence Centrality	0,0731	0,0872	0,0867	0,1002
Sentence Inclusion of Numerical Data	0,2269	0,2212	0,1906	0,2730
Sentence Length	0,2154	0,2191	0,1736	0,3090
Sentence Position	0,0910	0,1081	0,1070	0,1124
Sentence Resemblance to the Title	0,2237	0,2195	0,1805	0,2865
Text Rank	0,1372	0,1593	0,1545	0,1796
TF/IDF	0,2410	0,2313	0,1825	0,3257
Upper Case	0,1673	0,1621	0,1389	0,2029
Word Co-occurrence	0,0679	0,0940	0,0909	0,1108
Word Frequency	0,1917	0,1970	0,1560	0,2747

APÊNDICE R - Tabela com os resultados dos métodos de sumarização extrativa para a classe “Tempo”.

Método de Sumarização	MATCHES	Medida F ROUGE-2	Precisão ROUGE-2	Cobertura ROUGE-2
Aggregate Similarity	0,2604	0,2650	0,2397	0,3136
Bushy Path of the Node	0,2740	0,2578	0,2249	0,3168
Cue-Phrases	0,2406	0,2283	0,2098	0,2603
Lexical Similarity	0,1958	0,2039	0,1969	0,2363
Proper Noun	0,2583	0,2553	0,2160	0,3290
Sentence Centrality	0,1552	0,1154	0,1268	0,1112
Sentence Inclusion of Numerical Data	0,2802	0,2555	0,2193	0,3231
Sentence Length	0,3396	0,3195	0,2584	0,4400
Sentence Position	0,3260	0,3195	0,2879	0,3840
Sentence Resemblance to the Title	0,3208	0,2973	0,2438	0,3971
Text Rank	0,2406	0,2283	0,2098	0,2603
TF/IDF	0,3729	0,3533	0,2832	0,4918
Upper Case	0,2375	0,2387	0,2032	0,3062
Word Co-occurrence	0,2813	0,2943	0,2667	0,3479
Word Frequency	0,3469	0,3473	0,2822	0,4702