Apêndice A

Considerações prévias à sumarização automática

Apêndice do Capítulo 5

Jackson Wilke da Cruz Souza
Paula Christina Figueira Cardoso
Crysttian Arantes Paixão

Publicado em: 13/03/2024

Ao analisarmos os métodos e sistemas de sumarização, torna-se necessário estabelecer e detalhar alguns procedimentos importantes para trabalhar com sumarização de textos. Muitas vezes, o usuário de sistemas de SA tem em mente o que deseja realizar com o seu texto, porém nem sempre se atenta aos detalhes que podem influenciar no processamento, análise e resultados.

A) Formalidades e variações da língua

O primeiro detalhe é conhecer a qualidade do texto que se deseja fazer a sumarização. A origem do texto determina a qualidade e o padrão da grafia textual. Um texto pode vir de diferentes fontes para compor um *corpus* ou corpora; Por exemplo, em textos de redes sociais, nos quais o processo de escrita não segue, necessariamente a variedade, é formal da língua e possui limitação de caracteres, dependendo do público que gerou as mensagens, demandará uma atenção especial. Uma base de texto que siga uma padronização, de preferência normalizado, reduzirá em muito o tratamento das inconsistências, otimizando a análise (cf. Capítulo Recuperação de Informação).

Problemas envolvendo acentuação são os mais comuns, desconsiderando o problema de grafia. Vamos ao exemplo do uso da palavra "não". Em mensagens do Twitter, um usuário pode escrevê-la de várias formas para transmitir a mensagem, como "n", "não" e "nao", sendo que todas essas variações correspondem ao "não". Para resolver esse problema, o uso de dicionário seria uma opção para corrigir as palavras (corretor ortográfico), como existem nos editores de textos tradicionais.

B) Obtenção dos textos-fonte

Dependendo da origem dos textos, alguns problemas podem ser encontrados. No caso das redes sociais, quando se possui uma interface de programação de aplicações - API, o texto é coletado de forma automatizada. Porém, os problemas com a falta de padronização ainda podem persistir.

Alguns pesquisadores realizam estudos envolvendo apenas textos que possuem uma normalização, exemplo os que analisam trabalhos de conclusão de curso - TCC, dissertações e teses. Existem ainda os que utilizam artigos, de revistas e jornais, para realizarem suas pesquisas.

Nesses textos, por conta da formalização textual, muitos problemas são minimizados. Entretanto, eles ainda podem ocorrer. Um exemplo é quando a base é composta por

textos que seguem a formação dos programas Word da Microsoft, Write do OpenOffice, documentos do Google e em PDF (*Portable Document Format*). Esses formatos utilizam uma linguagem para a formatação dos textos que não é visível ao usuário, como *Extensible Markup Language* - XML do docx do Word, que tem que ser tratada no momento da importação para análise, pois apenas o texto deve compor o *corpus*. Para visualizar a fonte de um arquivo docx, sugerimos que o arquivo docx analisado seja aberto em um editor de texto simples, como o bloco de notas do Windows. No caso de arquivos no formato PDF, a conversão também é necessária. Existem ainda textos que são oriundos de páginas da internet que são capturadas, nos quais a linguagem que estrutura o site na maioria das vezes, é a Linguagem de Marcação de HiperTexto - HTML.

Esse exemplo é fácil de se constatar. Abra uma página da internet, selecione o texto e cole no Word. Você notará que a formatação das palavras também foi mantida. A pergunta é: como? A resposta é simples: junto com o texto desejado, você copiou um texto de marcação que irá compor o seu documento; logo, este deve ser tratado antes de realizar a análise.

Uma forma de contornar esse problema é copiar o texto da internet, colar em um editor de textos simples e, em seguida, copiar novamente o texto do editor para o Word. Dessa forma,a linguagem invisível ao olhar humano é removida. Porém, trata-se de um serviço manual e dependente da quantidade de textos, o que o torna inviável. Assim, em algum momento nesse processo, será necessário utilizar recursos computacionais para o processamento textual. Em suma, independente da fonte, deve-se realizar uma limpeza dos arquivos, o que usualmente recebe o nome de pré-processamento, deixando apenas o texto a ser analisado.

Uma sugestão é que os textos a serem analisados, na medida do possível, estejam em formato de texto puro, como os criados em blocos de notas. Recomenda-se sempre que, após a importação ou carregamento de uma base textual, uma inspeção visual seja realizada em uma amostra do texto para a verificação de possíveis problemas.

C) Fator cultural

Dependendo da região do país, existem fatores culturais que impactam diretamente a forma de expressão, como o uso de gírias. Portanto, em alguns casos, como desejamos monitorar o que está repercutindo, devemos padronizar essa informação. Para isso, torna-se necessário o uso de léxicos de gírias. O léxico serve como uma orientação na troca das palavras de mesmo significado para se estabelecer um padrão. Basicamente, o léxico funcionaria como um "corretor ortográfico", estabelecendo o padrão textual para análise. Como exemplo, temos o uso dos termos "pão de sal", "pão", "pãozinho", "cassetinho", entre outros, que referem-se a mesma coisa.

O desafio imposto é a obtenção desses léxicos. Na maioria dos casos, o usuário deverá criá-lo, do zero ou a partir de algum outro que contemple os termos que deseja, pois dificilmente encontrará, salvo exceções, um léxico específico para sua aplicação.

D) Emojis

Outro fato relevante em textos de redes sociais é o uso de emojis, ideogramas e smileys. O usuário se vale de tais recursos em suas mensagens para transmitir uma informação relevante, o que em determinada parte do texto pode dificultar o processo de sumarização, demandando configurações especiais. No processo de SA, os emojis serão considerados como palavras e, dependendo da análise, deverá ser utilizado um dicionário de sinônimos para descrevê-los.

Além disso, consideremos outros sistemas, nos quais o usuário de certas redes sociais pode se expressar diretamente com outros elementos multimodais (como vídeos e fotos).

A interação dos usuários também leva em conta o regionalismo, que torna necessário a adaptação de um léxico para tentar minimizar o efeito de idiossincrasias em função da informalidade textual. As temáticas abordadas devem ser consideradas para determinar se o texto pode compor um sumário.

E) Expressões regulares

Aos leitores interessados em análise de textos, recomendamos um estudo, pelo menos introdutório, sobre expressões regulares. Como referência, indicamos o livro de Expressões Regulares (Jargas, 2016). O livro apresenta de forma simples os fundamentos de expressões e suas aplicações. Basicamente, uma expressão regular é utilizada para representar um padrão, o qual é muito utilizado dentro das linguagens de programação que podem ser executadas no processamento textual.

Um exemplo é o caso da palavra "não" com diversas ocorrências que podem ser representadas pela expressão regular "[Nn][aã]?o?", referindo-se às variações "Não", "Nao", "N", "não", "nao" e "n". Grande parte desse pré-processamento depende do uso de funções da linguagem de programação, que recebem como parâmetro uma expressão regular para realizar as operações. O uso dessas funções otimiza o processamento textual, permitindo padronizar o texto, reduzindo problemas e viabilizando as análises.

F) Pré-processamento

Até o momento, relatamos o problema envolvendo a criação da base de dados. Porém, antes de analisar o texto, devem ser feitas, como capitalização, acentuação, símbolos de pontuação e *stopwords*, por exemplo

Algumas perguntas podem surgir ao se pensar que uma palavra escrita com letras maiúsculas e minúsculas é diferente. Como exemplo, temos: o computador diferencia "Carro" e "carro"? ou existe um tipo de caixa para otimizar o processamento? Você já sabe a resposta! A forma com a qual são armazenadas é o que as diferencia. Lembre-se sempre que o computador não possui discernimento entre o significado da palavra, mas sim de como ela é representada. Nesse caso, letras maiúsculas e minúsculas possuem codificações diferentes. O mesmo ocorre com a acentuação.

Já para os símbolos de pontuação, em algum determinado momento, o usuário deve escolher qual o caractere que servirá de delimitador de palavras (*tokens,* em PLN). Em alguns idiomas, esse delimitador é o espaço em branco, como na frase "O dia estava lindo, mas pode chover no período da tarde". Facilmente, você consegue identificar as palavras que compõem essa oração. Porém o computador, dependendo da forma como for programado, pode considerar que a palavra "lindo" não esteja presente na frase, por conta da pontuação. Nesse caso "lindo," com destaque para a presença da vírgula, seria a palavra considerada. A vírgula também é processada como um caractere, logo pode influenciar a análise em decorrência que "lindo" é diferente de "lindo,". Isso ocorre muito em nuvens de palavras quando não se atenta pela remoção desses caracteres de pontuação. Logo, uma palavra com determinada frequência poderá ocorrer um número maior de vezes no texto, enviesando a análise.

Mais um detalhe a ser analisado é a forma da escrita que constitui o texto. Apesar de ocorrer em menor frequência, temos alguns itens a serem sempre analisados. O primeiro é o uso de separação silábica, pois a palavra "casa" é diferente de "ca-sa" quando inserida em um texto e processada pelo computador, pois o "-" é um símbolo.

Dependendo da situação, o texto ainda tem de ser pré-processado para eliminar palavras sem valor informacional, comumente denominadas *stopwords*. No caso de SA, nem sempre existe a eliminação dessa "classe" de palavras, mas vale chamar a atenção. Muitas vezes, essa lista de palavras a serem removidas consta de um léxico já padronizado.

Porém, em determinados domínios e públicos torna-se necessário atualizá-lo ou adaptá-lo. Como exemplo, considere um estudo que envolva a palavra-chave "COVID" a ser coletada em mensagens no Twitter. As mensagens que serão coletadas terão a palavra "COVID" em seu texto, sendo que não há adição de informação no seu uso, podendo ser removida do corpus. Todos os textos, de forma geral, terão essa palavra e dependendo da análise, ela é dispensável, ou seja, é interessante removê-la.

Referências

JARGAS, A. M. **Expressões Regulares - 5a edição: Uma Abordagem Divertida**. [s.l.] Novatec Editora, 2016.