Recuperação inteligente da informação
Relatório atividade E-1.8

André Pacheco

Doutorado em Ciência da Computação
Programa de pós-graduação em informática
Universidade Federal do Espírito Santo

Conteúdo

1 Introdução

A atividade E-1.8 tem como objetivo a identificação dos padrões do texto da base de dados *usielCarneiro* já trabalhada na atividade E-1.5. Neste caso a base segue um padrão de uma data, um título do artigo, uma passagem da bíblia e o texto do autor. Nesta atividade, todos os arquivos foram analisados e gerou-se novos arquivos com cada seção que foram salvos em pastas divididos por mês.

2 Identificação das seções

Para identificar as seções um modelo probabilístico pode ser utilizado. Todavia, existem apenas 176 arquivos e para gerar um treinamento com os mesmos talvez não seja o suficiente. Sendo assim, foi utilizado um modelo probabilístico "gerenciado". Os passos para obter as identificações são descritos na sequência.

2.1 Data

Para identificar a data foi utilizado a seguinte abordagem. Primeiramente existem palavras chaves que descrevem um data, como nome dos dias da semana e do ano. Além disso, o número de caracteres total da frase também tem que ser levado em consideração. Sendo assim, toda linha recuperada é quebrada por palavras e cada palavra é verificada se ela pertence ao conjunto de palavras chaves de uma data. Sendo assim, é realizado a divisão da quantidade de palavras chaves na frase pelo número de palavras total. Esse valor é a probabilidade da frase ser uma data. Além disso, se uma frase possui mais de 100 caracteres essa probabilidade diminui exponencialmente. Com essa aboradagem, foi possível identificar todas as frases que representam uma data no documento.

2.2 Título

Para identificar um título foi levado em consideração o número de caracteres da amostra e a informação do rótulo da frase anterior. Se o rótulo anterior for uma data a probabilidade de ser um título é potencializada. Além disso, se o conjunto de caracteres excede 200, a probabilidade de ser um título diminui exponencialmente.

2.3 Passagem da bíblia

Nesta identificação também é considerada as classes anteriores. Se forem Data e títulos a probabilidade de ser uma passagem da bíblia é potencializada. Além disso, é identificado os tokens dos versiculos e verificado se este pertence a um livro da bíblia, tanto como nome completo, quanto como abreviação (por isso que o modelo é probabilístico gerenciado, pois essas informações já são dadas para o sistema previamente, sem retirar do texto em si). Na sequencia é realizada a mesma verificação executada na data.

2.4 Texto do autor

Por fim é verificado se o conjunto de caracteres é um texto do autor. Da mesma forma que anteriormente, para determinar a probabilidade de ser um texto também é verificado as 3 classes anteriores. Como o padrão dos arquivos é sempre data, título, passagem e texto, a probabilidade é potencializada caso o padrão seja mantido. Além disso, é verificado se o número de excede 300, a probabilidade de ser um texto do autor cresce.

2.5 Problemas encontrados

Alguns dos problemas encontrados são discutidos nesta subseção. Primeiramente, em alguns arquivos existe a data mas o autor não escreveu nada na sequência. Esse texto foi descartado. Em outros casos, existem texto no início do arquivo,

antes dos textos diários. Esses textos também foram descartados. Outro problema são erros de português na escrita de dias da semana e/ou meses. Em alguns documentos o autor escreve, por exemplo, STEMBRO ou SETEBRO ou invés de SETEMBRO. Para solucionar isso, foi utilizado um regex de padrão para verificar se a palavra escrita erroneamente casa com a corretamente. Com isso foi possível contornar essa adversidade.

2.6 Exemplo de um texto minerado

Todos os textos extraídos são savos nas pastas de acordo com o mês que ele foi escrito. Não foi separado por ano pois os anos são escritos no nome do arquivo/pasta original. Um exemplo de texto estraído é descrito na sequência:

```
<DATA_DO_DOCUMENTO>
SÁBADO, 23 DE ABRIL

<TITULO_DO_DOCUMENTO>
A CRUZ DE CRISTO

<CITACAO_DA_BIBLIA>
"Quanto a mim, que eu jamais me glorie, a não ser na cruz de nosso Senhor
Jesus Cristo, por meio da qual o mundo foi crucificado para mim, e eu para
o mundo. De nada vale ser circuncidado ou não. O que importa é ser uma nova
criação." (Gálatas 6.14-15)

<TEXTO_DO_PASTOR>
Quanto mais organizamos nossa vida religiosa e formalizamos nossa fé, criando
símbolos para a nossa espiritualidade, mais corremos o risco de abrir mão da
humildade cristã pelo orgulho religioso. Na religião, as vezes acontece como
na sociedade: há símbolos de grandeza e indicadores de superioridade.
Por exemplo: conhecimento bíblico, cargos e funções na igreja, dons
especiais, encaixar-se nos padrões, etc. Na carta aos cristãos da Galacia
Paulo criticou os que tentavam impressionar e declarou sua escolha:
gloriar-se na cruz de Cristo.
```