Recuperação inteligente da informação
Relatório atividade E-1.4

André Pacheco

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

Conteúdo

Recuperação inteligente da informação Relatório atividade E-1.4 1

1 Introdução 1
2 Indexação da base dados 1
3 Similaridade 4
4 Página HTML 6
5 Conclusão 7

1 Introdução

A atividade E-1.4 tem como objetivo a realização de uma análise da base de dados `UsielCarneiro` [1]. Esta base possui 176 documentos na extensão `.docx` contendo reflexões pessoais do pastor Usiel Carneiro. Para sua análise foi utilzado o aLine juntamente com outras bibliotecas Python, como a NLTK [2], e implementações próprias. Ao longo deste relatório serão discutidos os resultados em relação a frequência e sintática das palavras, n-grams, similaridade entre documentos e tempo computacional.

2 Indexação da base dados

Para realizar a indexação da base de dados foi utilizado o aLine da seguinte forma:

```
$ aLine -i -l lista-arquivos.txt -d usielCarneiro
```

[1] Disponível em http://www.inf.ufes.br/ elias/dataSets/usielCarneiro.tar.gz
[2] http://www.nltk.org/

Na qual, todos os resultados do framework são armazenados na pasta `usielCarneiro`. Nesta atividade, será utilizado o arquivo `dictionary.txt`, onde fica armazeando a indexação propriamente dita, e o arquivo `cache.txt`, que contém a vetorização dos documentos. Este último será utilizado para o cálculo de similaridade dos documentos.

No entanto, o aLine não é capaz de lidar com arquivos de extensão `.docx`, portanto o primeiro passo do algoritmo é converter todos os documentos para `.txt`. Na sequência, foi gerado a lista de arquivos para que o aLine possa indexá-los. O resultado da indexação apresentaram alguns problemas. Tomando como exemplo o *token* `deus`, o mesmo foi indexado de diferente maneiras: `deus`, `deus"`, `"deus`, `deus'`, `'deus`, `deus],` `"deus,` `deus`. Esse mesmo problema ocorre com diversas palavras na base de dados. Como não se tem acesso ao código fonte da ferramenta, não é possível corrigir esses problemas diretamente na indexação. Todavia, no script Python, foi retirado esses caracteres. Por este motivo, na indexação do aLine, `deus` aparece 9.141 vezes, todavia, mas na verdade, retiradas todas essa anormalidades, o *token* possui frequência de 9.318.

Realizada a indexação, a primeira análise é relacionada a frequência das palavras, substantivos, adjetivos e [bi-tri]-grams. Como a aLine não é capaz de realizar tais tarefas, foi necessários o uso de outras ferramentas e estratégias. As palavras mais frequentes da base de dados foi facilmente calculada por meio do método `Counter()`, nativo do python. Os bi e tri grams mais frequentes foram obtidos por meio da biblioteca NLTK. A análise de substantivos e adjetivos necessitou de uma estratégia mais elaborada. É difícil encontrar ferramentas para este tipo de análise na língua portuguesa. Portanto, a melhor estratégia encontrada foi incorporar um dicionário e verificar a a classe gramatical da palavra. Foi realizado o download de um dicionário aberto [3] em português brasileiro em XML. Foi realizado um *parse* neste arquivo para transformá-lo em um dicionários (hash) em Python. Essa estrutura contém 123.390 palavras e suas respectivas classes gramaticais. Esse é um projeto interessante e por conta disso, ele próprio foi disponibilizado no meu github pessoal [4]. Todavia, como será discutido na sequência, ainda tem muito o que ser melhorado nessa abordagem.

Na Tabela 1 são descritas as 30 palavras, os 30 substantivos e os 30 adjetivos mais frequentes em toda a base de dados. Nas palavras mais frequentes, podem ser observadas diversas *stopwords* e, por se tratar de documentos escritos por um pastor, palavras de um vocabulário religioso, como `deus`, `jesus` e `vida`. Na frequência de adjetivos, pode-se notar palavras como `amoroso`, `saudável`, `feliz` etc. Todavia, muitas palavras são consideradas como adjetivo por este dicionário, mas depende de um contexto para serem interpretadas como tal, por exemplo, `mundo` e `filho`. Por fim, nos substantivos mais comuns também aparecem palavras de cunho religioso como `pecado`, `viver`, `cristo`, `evangelho` etc. É importante ressaltar, que nem todas as palavras escritas pelo pastor estão disponíveis neste dicionário aberto. Palavras como `chorarem`, `talhados`, `iludiremos`, dentre outras, não estão presentes. Portanto, é possível perceber a dificuldade

[3] http://www.dicionario-aberto.net/
[4] http://www.github.com/paaatcha/dicionario-ptbr

da análise sintática, e que nessa abordagem adotada, é muito sensível ao dicionário utilizado. Essa é uma abordagem interessante para usar um sistema de *crowdsourcing* para aumentar e melhorar o dicionário. Essa é uma ideia na qual continuarei trabalhando.

Tabela 1. Palavras, substantivos e adjetivos mais frequentes

Rank	Palavra	Freq.	Adjetivo	Freq.	Substantivo	Freq.
1	e	23073	tudo	1606	deus	9318
2	de	22062	mundo	927	mas	5011
3	que	21106	filho	540	um	4554
4	a	19935	meu	427	sua	2372
5	o	17243	preciso	356	cristo	2278
6	não	9923	qualquer	290	fé	1996
7	deus	9318	mestre	235	seu	1985
8	é	9253	momento	228	são	1449
9	nos	7690	qual	227	reino	1301
10	se	7354	algum	209	pela	1001
11	para	7282	espiritual	206	viver	959
12	em	6963	diferente	189	sempre	947
13	com	5373	irmão	181	senhor	920
14	do	5251	difícil	174	porque	880
15	mas	5011	feliz	170	eu	858
16	ele	4785	duas	154	algo	853
17	por	4765	maio	147	pai	779
18	vida	4742	saudável	138	coração	685
19	um	4554	fundamental	123	era	619
20	como	4284	capaz	122	hoje	618
21	da	4006	nenhum	119	pecado	559
22	jesus	3800	perto	115	presença	518
23	uma	3757	pastor	106	verdade	515
24	os	3661	espaço	96	história	510
25	mais	3267	eterno	89	deve	506
26	ser	3079	futuro	87	cristão	498
27	nossa	2978	amoroso	86	grande	467
28	as	2854	servo	86	próximo	445
29	nós	2745	teu	83	evangelho	433
30	isso	2719	único	80	estar	423

Na Tabela 2 são apresentados os 30 bi e tri-grams mais frequentes na base de dados. Novamente observamos uma linguagem bem típica de uma pessoa religiosa, com por exemplo, `reino-de-deus`, `de-deus`, `a-fé-cristã` etc. Além disso, como os arquivos possuem as datas por extenso, também aparece o bi-gram `de-dezembro`, remetendo ao mês que o pastor escreveu esta reflexão.

Em relação ao tempo computacional da indexação realizada pelo aLine, considerando todos os 176 documento da base, a média de execução é menor do que 0.23 seg. Portanto, para essa base de dados, este não é um fator limitante.

Tabela 2. Os bi e tri-grams mais frequêntes na base de dados

Rank	bi-grams	Freq.	tri-grams	Freq.
1	(de - deus)	138	(reino - de - deus)	92
2	(reino - de)	93	(o - reino - de)	66
3	(o - reino)	75	(em - primeiro - lugar)	22
4	(o - que)	70	(de - deus - e)	14
5	(a - deus)	52	(no - reino - de)	13
6	(a - vida)	44	(de - deus - em)	13
7	(deus - e)	40	(honrar - a - deus)	11
8	(năo - é)	39	(com - deus - e)	11
9	(e - o)	35	(do - reino - de)	11
10	(e - a)	35	(a - fé - cristă)	11
11	(com - o)	34	(a - deus - e)	11
12	(de - dezembro)	31	(deus - em - primeiro)	9
13	(é - o)	31	(deus - e - ao)	9
14	(que - nos)	31	(de - deus - é)	9
15	(nossa - vida)	28	(primeiro - lugar - o)	9
16	(que - năo)	27	(em - nossa - vida)	8
17	(do - que)	26	(de - deus - o)	8
18	(em - primeiro)	23	(deus - e - a)	8
19	(que - o)	22	(o - modo - como)	8
20	(primeiro - lugar)	22	(de - deus - chegou)	7
21	(a - fé)	22	(dízimos - e - ofertas)	7
22	(com - a)	21	(com - o - reino)	7
23	(e - năo)	20	(a - deus - com)	7
24	(é - a)	19	(a - vida - e)	7
25	(o - tempo)	19	(lugar - o - reino)	7
26	(com - deus)	19	(amor - a - deus)	6
27	(do - reino)	19	(e - ao - próximo)	6
28	(por - isso)	19	(amor - ao - próximo)	6
29	(que - a)	18	(uns - dos - outros)	6
30	(deus - em)	17	(servir - a - deus)	6

3 Similaridade

Para computar a similaridade entre os documentos, o aLine foi mais uma vez utilizado. Dessa vez, da seguinte forma:

```
$ aLine --similarity -d usielCarneiro --features usielCarneiro/cache.txt'
```

na qual a tag `-similarity` indica para o algoritmo para calcular a similaridade entre todos os documentos da base utilizado a vetorização obtida na indexação. Todo o resultado é armazenado no arquivo `similaridade.mtx`, que possui a métrica de similaridade entre todos os pares de documentos.

A medida de densidade espacial é definida da seguinte forma (Salton et al., 1975):

$$F = \sum_{n=1}^{n} \sum_{j=1}^{n} s(D_i, D_j) \forall i \neq j \quad (1)$$

na qual, n é o número de documentos da base, D_i e D_j são o i-ésimo e j-ésimo documento dessa mesma base e $s()$ é a medida de similaridade. Sendo assim, de acordo com os valores de similaridades obtidos pelo aLine a equação 1, para a base `usielCarneiro` a medida de densidade espacial é igual a 18201.91. Além disso, a similaridade média entre os pares de documentos é igual a 0.5932.

Ainda em relação a similaridade, os 10 documentos mais similares em toda a base de dos são:

1. **s = 1.0**
   - `devocionais-2013/Devocional Diário - 52 - 13 a 19 de setembro.docx`
   - `devocionais-2013/Devocional Diário - 59 - 01 a 07 de novembro.docx`
2. **s = 0.992975**
   - `devocionais-2013/Devocional Diário - 40 - 21 a 27 de junho.docx`
   - `devocionais-2013/Devocional Diário - 03 - 13 a 19 de janeiro.docx`
3. **s = 0.992717**
   - `devocionais-2013/Devocional Diário - 51 - 06 a 12 de setembro.docx`
   - `devocionais-2013/Devocional Diário - 16 - 15 e 16 de fevereiro.docx`
4. **s = 0.986623**
   - `devocionais-2013/Devocional Diário - 03 - 13 a 19 de janeiro.docx`
   - `devocionais-2013/Devocional Diário - 54 - 27 a 03 de outubro.docx`
5. **s = 0.986193**
   - `devocionais-2013/Devocional Diário - 47 - 09 a 15 de agosto.docx`
   - `devocionais-2013/Devocional Diário - 03 - 13 a 19 de janeiro.docx`
6. **s = 0.985991**
   - `devocionais-2013/Devocional Diário - 09 - 31 de janeiro.docx`
   - `devocionais-2013/Devocional Diário - 47 - 09 a 15 de agosto.docx`
7. **s = 0.985669**
   - `devocionais-2013/Devocional Diário - 62 - 22 a 28 de novembro.docx`
   - `devocionais-2013/Devocional Diário - 47 - 09 a 15 de agosto.docx`
8. **s = 0.985498**
   - `devocionais-2013/Devocional Diário - 09 - 31 de janeiro.docx`
   - `devocionais-2013/Devocional Diário - 03 - 13 a 19 de janeiro.docx`
9. **s = 0.985068**
   - `devocionais-2015/10 Outubro de 2015.docx`
   - `devocionais-2013/Devocional Diário - 03 - 13 a 19 de janeiro.docx`
10. **s = 0.984943**
    - `devocionais-2013/Devocional Diário - 54 - 27 a 03 de outubro.docx`
    - `devocionais-2013/Devocional Diário - 02 - 06 a 12 de janeiro.docx`

Pode se observar que dos dez documentos mais similares, nove foram escritos no mesmo ano. Todos eles são textos que citam muito a palavra `jesus`, `cristo` e `deus`.

O cálculo de similaridades par a par é bastante custoso computacionalmente. Sempre serão necessários calcular a métrica para um combinação de N tomados 2 a 2, sendo N o tamanho da base. A base `usielCarneiro` é pequena, possui apenas 176 documentos. Ainda assim, foi calculado o tempo computacional considerando lotes de 25 documentos. Observando a Figura 1, é possível perceber que o custo computacional para calcular a métrica é exponencial. Sendo assim, para bases com número grande de arquivos, o cálculo par a par é extremamente lento. Dependendo da base, até inviável.

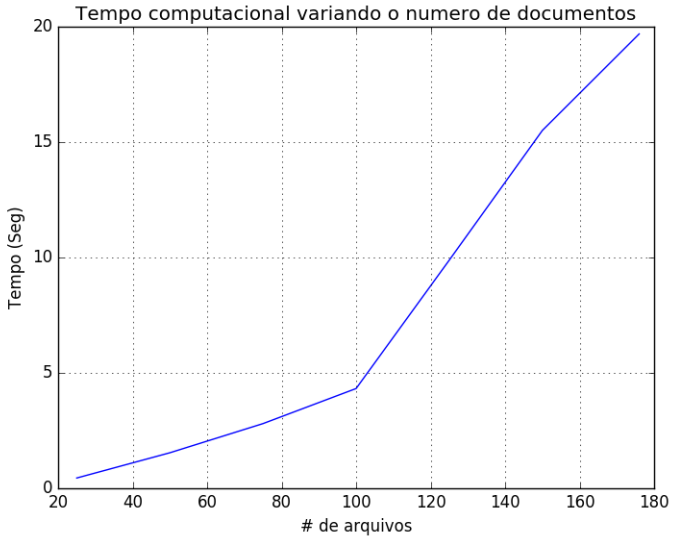

Figura 1. Variação do tempo computacional para calcular a similaridade par a par entre os documentos

4 Página HTML

Por fim, todos os documentos da base de dados são convertidos para PDF e uma página HTML é gerada, na qual todos os arquivos são *likandos*. Ao fim da execução do script Python, o usuário decide, repondendo sim ou não, se deseja gerar essa página. Caso sim, o processo de conversão dos dados será iniciado. Ele pode levar alguns minutos para converter todos os documentos. Em seguida, uma página HTML de com nome `usielCarneiro.html` é gerada no diretório raíz da aplicação.

5 Conclusão

Nesta atividade a base de dados `usielCarneiro` foi processada utilizando o aLine e outros utilitários em python. Além da indexação dos dados, foram apresentadas as palavras mais comuns e uma pequena análise sintática. Os resultados são coerentes com os tipos de dados apresentados, uma vez que os termos mais frequentes (sem contar *stopwords*), são típicos de um vocabulário religioso. A análise sintática se mostrou uma tarefa complicada, uma vez que o contexto altera o o grupo gramatical das palavras. Com isso, a solução apresentada possui bastante espaço para melhorias no futuro. Além disso, o aLine foi utilizado para calcular a similaridade entre os documentos. Foram apresentados os dez documentos mais similares bem como uma breve análise do tempo computacional para o cálculo dos mesmos. É interessante ressaltar, que para bases muito grandes, o calculo da similaridade par a par é praticamente inviável devido ao alto custo computacional.

Bibliografia

Salton, G., Wong, A., Yang, C.S., 1975. A vector space model for automatic indexing. Communications of the ACM 18, 613–620.