Identificação e Organização de Fontes de Dados Textuais na Web

Diogo T. Ferreira, Luciano Antonio Digiampietri
Escola de Artes, Ciências e Humanidades, USP

Objetivos

Este projeto objetiva a identificação de bases textuais na Web e a especificação e o desenvolvimento de ferramentas para baixar e organizar automaticamente informações dessas bases Desta forma, faz-se necessário o desenvolvimento de ferramentas capazes de identificar fontes de dados confiáveis e sem repetição de informação, efetuar a recuperação de documentos em grande escala e filtrar as informações de modo a descartar anotações variadas, informações de formatação e outros[1].

Método/Procedimentos

Para este projeto foram realizados os seguintes métodos:
1. Identificação de fontes de dados textuais.
2. Desenvolvimento e implementação de uma ferramenta básica de busca de documentos.
3. Filtragem dos documentos para eliminação de conteúdos considerados desnecessários.
4. Cálculo de métricas: cálculo de diversas métricas sobre as informações extraídas na atividade anterior.

Resultados

Utilizando uma função da ferramenta wget, oriunda do Sistema Operacional Ubuntu, foi possível baixar inúmeros arquivos de portais de notícias. Em seguida, desenvolveu-se um script para filtragem das tags HTML e do conteúdo de forma que o corpo principal de cada documento foi mantido.
No passo seguinte, o conteúdo foi filtrado através da ferramenta PreText[2], para eliminação de sufixos, prefixos, *stopwords* e caracteres não-textuais. O PreText produziu duas saídas para cada documento, um com o texto filtrado com as condições já ditas e outra com o texto com 3-gramas[2].
Após a última filtragem foi criado mais um script, agora para contagem da frequência relativa das palavras e dos n-gramas em relação ao assunto da notícia e a frequência total.

Como última parte do projeto, desenvolvemos um programa que a partir dos dados de frequência relativa, identificamos o assunto provável de cada texto.
Abaixo seguem as quantidades de dados recolhidos dos sites escolhidos para o projeto:
-PortalG1(*g1.globo.com/*): 7,4GB(GigaBytes) de informações;
- R7(*www.r7.com/*): 14GB
-Época(*revistaepoca.globo.com/*): 3,7GB
- Veja(*veja.abril.com.br/*): 20GB
- Folha(*www.folha.uol.com.br/*): 8,8GB
Abaixo o gráfico demonstrando a frequência de repetição dos 3-gramas comparada com a base de documentos que obtivemos:

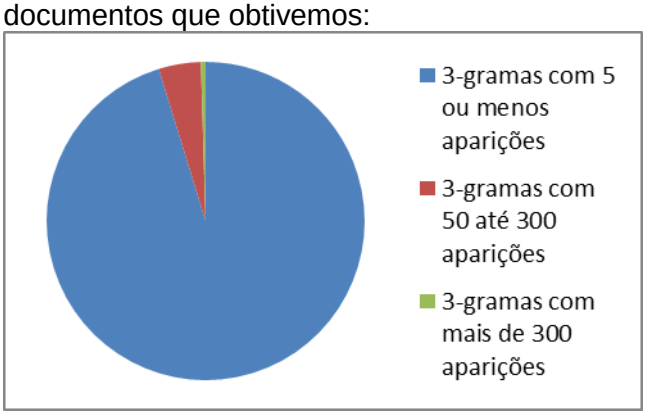

Figura 1 Gráfico demonstrando a frequência com que os 3-gramas aparecem.

Conclusão

Este projeto ajudará projetos futuros que necessitarão de um grande repositório de dados em forma textual. As ferramentas utilizadas e criadas se mostraram eficientes, dando-nos confiança para coletar uma quantidade imensa de conteúdo e posteriormente filtra-los. Havendo, inclusive, um classificador simples do assunto de notícias novas.

Referências Bibliográficas

[1]Kilgarriff, A. (2007) Googleology is Bad Science. Computational Linguistics (33) vol.1.
[2]Soares, M. V. B., Prati, R. C. Monard, M. C. (2008). PreTexT: A Reestruturação da Ferramenta de Pré-Processamento de Textos. São Paulo: Universidade de São Paulo - ICMC