Proposta de projecto de colaboração com o Arquivo da Web Portuguesa

Prospecção de texto suportada em colecções de n-gramas

A FCCN tem em curso o projecto de Arquivo da Web Portuguesa e procura colaborar com entidades de Investigação e Desenvolvimento, que tenham interesse em participar na realização de actividades inovadoras. Este documento apresenta uma proposta de um projecto com a duração estimada de 1 ano, que poderia fazer parte de um trabalho de mestrado ou de iniciação à investigação.

Os modelos baseados em n-gramas são usados em várias tarefas de processamento estatístico de língua natural, tais como o reconhecimento de entidades em texto, a correcção ortográfica ou extracção de informação. Em 2006, a Google disponibilizou uma colecção de dados que veio a ganhar uma imensa popularidade entre a comunidade de investigadores, a qual consiste ficheiros de texto com contagens de frequência para os n-gramas de palavras (com n variando entre 1 e 5) extraídos da recolha da Web feita pelo *crawler* do Google (http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html). Embora muito útil, esta colecção de n-gramas apresenta a desvantagem de apenas incluir palavras de documentos Web escritos em inglês. Para tarefas envolvendo o processamento de outras línguas, seria útil ter uma colecção de n-gramas semelhante à disponibilizada pelo Google.

Neste projecto pretende-se:

- abordar a construção de um corpus de n-gramas, com base em textos da Web recolhidos no contexto do projecto do Arquivo da Web Portuguesa;
- estudar a aplicação da colecção de n-gramas no contexto de problemas de prospecção de textos em português, nomeadamente em problemas de reconhecimento de entidades e de correcção ortográfica.

Neste trabalho será estudada a utilização da plataforma de computação distribuída e das recolhas de documentos associadas ao projecto do Arquivo da Web Portuguesa, com vista à construção de um corpus de n-gramas para a língua Portuguesa semelhante à colecção disponibilizada pelo Google. O Arquivo da Web Portuguesa usa uma plataforma de computação distribuída denominada Hadoop,

a qual consiste essencialmente de uma implementação *open-source* da plataforma MapReduce proposta pelo Google.

Este projecto apresenta potencialidade para atingir os seguintes objectivos:

I. Construção da aplicação

- Construir uma colecção de n-gramas, semelhante à disponibilizada pelo Google, com base nos documentos recolhidos no projecto do arquivo da Web Portuguesa. Nesta tarefa, poderá vir a ser reutilizado algum *software* já existente (http://www.qnan.org/~pmw/software/hadoop-ngram/doc/).
- Estudar possíveis extensões ao formato utilizado pelo Google na sua colecção de n-gramas, por exemplo armazenando a informação temporal associada às recolhas dos documentos (i.e., construir colecções de n-gramas correspondentes a vários snapshots tirados ao longo do tempo).

II. Análise e aplicação

- Avaliar a aplicação da colecção de n-gramas em problemas de reconhecimento de entidades em textos na língua Portuguesa, reutilizando por exemplo os dados do evento HAREM (http://poloxldb.linguateca.pt/harem.php). Nesta tarefa, poderá vir a ser reutilizado algum software já existente (http://turing.cs.washington.edu/lex/telex_download.zip).
- Avaliar a aplicação da colecção de n-gramas em problemas de correcção ortográfica, mais concretamente no problema da correcção ortográfica de *queries* introduzidas num motor de busca.

Apontam-se como vantajosos conhecimentos de programação em Java e algoritmia, assim como conhecimentos básicos (a desenvolver) nas áreas de *machine learning* e processamento de língua natural.

Bibliografia

- Xiaoyang Yu (2008) Estimating Language Models Using Hadoop and Hbase. MSc Thesis.
- F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber (2006) Bigtable: A distributed storage system for structured data. In OSDI '06, pages 205–218.
- J. Dean and S. Ghemawat (2004) Mapreduce: Simplified data processing on large clusters. In OSDI '04, pages 137–150.

- Ian Fette (2007) Combining n-gram based statistics with traditional methods for named entity recognition. School of Computer Science, Carnegie Mellon University.
- Downey, D., Broadhead, M., & Etzioni, O. (2007). Locating complex named entities in web text. IJCAI.
- Carlson, A.; Fette, I. (2007) Memory-based context-sensitive spelling correction at web scale. Sixth International Conference on Machine Learning and Applications
- Farag Ahmed, Ernesto William De Luca and Andreas Nürnberger (2008) MultiSpell: an N-Gram Based Language-Independent Spell Checker, In: Poster Postproc of Eighth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2007), Mexico City, Mexico, IEEE CS Press.
- Darshan Paranjape, Bin Lan, Vishnu Pedireddi, Anurag Jain (2007) Google N-gram Patterns. Department of Computer Science University of Minnesota, Duluth
- Satoshi Sekine (2008) A Linguistic Knowledge Discovery Tool: Very Large Ngram Database Search with Arbitrary Wildcards. Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)