FCCN
Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing

PORTUGUESE WEB ARCHIVE

Proposal for a collaborative project with the Portuguese Web Archive

Text retrieval using n-gram collections

FCCN is currently engaged in the Portuguese Web Archive project and seeks to cooperate with Research and Development organisations who are interested in participating in innovative activities. This document presents a proposal for a project with an estimated duration of 1 year, which could form part of a master's thesis or introduction to research.

Models based on n-grams are used in various tasks of natural language statistical processing, such as text entity recognition, spell checking and information extraction. In 2006, Google released a collection of data that has gained immense popularity among the research community, consisting of text files with frequency counts for n-grams of words (with n varying between 1 and 5) extracted from the Web compilation performed by Google's *crawler*. Although very useful, this collection of n-grams only includes words from Web documents written in English. For tasks involving the processing of other languages, it would be useful to have a collection of n-grams similar to that offered by Google.

This project aims to:

- tackle the construction of a corpus of n-grams, based on texts from the Web compiled in the context of the Portuguese Web Archive project;
- study the application of the n-gram collection in the context of problems of exploring texts in Portuguese, especially in problems of entity recognition and spelling correction.

There will be used the distributed computing platform and document collections associated belonging to the Portuguese Web Archive, for the construction of n-grams corpus for the Portuguese language which is similar to the collection provided by Google. The Portuguese Web Archive uses a distributed computing platform called Hadoop, which consists essentially of an *open-source* implementation of the MapReduce platform offered by Google.

This project aims to achieve the following objectives:

I. Development of an application to create the n-gram collection

- Generation of an n-gram collection, similar to the one provided by Google, based on documents held by the Portuguese Web archive. In this task, some existing software can be reused.
- Study possible extensions to the format used by Google in its collection of n-grams, for example by storing the temporal information associated with document compilations (i.e., build collections of n-grams corresponding to several snapshots taken over time).

II. Analysis and application

- Assess the application of the of n-grams collection in problems of entity recognition in Portuguese language texts, for example reusing the data of the HAREM event. Some existing software can be reused in this task.
- Assess the application of the n-gram collection in spell checking problems, specifically in the problem of checking the spelling of queries entered into the search engine.

Knowledge of Java programming and algorithms and interest in the areas of machine learning and natural language processing are required.

Bibliography

- 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)