CAMELEON Tutorials

Carlos Ramisch
2013-12-23

Table of Contents

Os tutoriais serão ministrados em português.

Tutorial 1 - Carlos Ramisch - LIF, Aix-Marseille Université (France)

```
Instituto de Informática da UFRGS, prédio 43413(67), sala 101
Segunda-feira, 16 de dezembro de 2013, 13:30 às 16:30
```

Statistical corpus processing: from raw text to lexical resources

Corpora (very large textual bases) are valuable sources of information. However, in order to make this information useful for computers, one needs to transform non-structured natural language sentences into more structured representations (e.g. lexicons, ontologies) that can be dealt with automatically. Corpus-based methods for natural language processing are the dominant trend in research and industry, and a powerful tool for building applications such as machine translation, information extraction and question answering. In this tutorial, I will use multiword expressions (recurrent word combinations like "big deal", "give up" and "once in a blue moon") as a case study for exploiting large text collections. The tutorial will contain two main parts. First, I will provide a theoretical overview of common techniques for dealing with corpora and extracting lexical information, specially multiword expressions. Second, I will describe step-by-step practical examples for common activities such as tokenization, case homogenization, part-of-speech tagging, word and n-gram counting, association measures and so on. This second part will use freely available tools (e.g. mwetoolkit, TreeTagger) and corpora (e.g. CAMELEON corpus) for the examples and exercises. This tutorial will be addressed at anyone interested in corpora and natural language processing. Basic knowledge about how to use the Unix terminal is recommended, but not strictly necessary. The target public includes computer science and applied linguistics students (undergraduate, graduate...) and researchers. Thus, I will not go into deep technical and statistical details, but make it as a general, broad overview. The main goal of the tutorial is to share tools and methods that allow intelligent exploitation of textual information, which can be further refined and adapted for many applications.

- Tutorial slides
- Tutorial materials (download and unzip in Desktop)

Processamento estatístico de corpus : do texto bruto aos recursos lexicais

Corpora (grandes bases de texto) são fontes valiosas de informação. No entanto, se quisermos tornar esta informação interpretável para computadores, é preciso transformar sentenças não estruturadas em linguagem natural em representações mais estruturados (por exemplo, léxicos e ontologias), que podem ser processadas automaticamente. Métodos baseados em corpus para processamento de linguagem natural são atualmente a tendência dominante na pesquisa e na indústria, e representam ferramentas poderosas para a criação de aplicações como tradução automática, extração de informações e sistemas de pergunta-resposta. Neste tutorial, vou usar expressões multipalavra (combinações recorrentes de palavras como "chutar o pau da barraca" , "correr um risco" e "ferro de passar"), como um estudo de caso para a exploração de grandes coleções de textos. O tutorial irá conter duas partes principais. Em primeiro lugar, vou apresentar uma visão teórica das técnicas comuns para lidar com corpora e para extrair informações lexicais, especialmente expressões multipalavras. Em segundo lugar, vou descrever passo-a-passo exemplos práticos para atividades comuns, como tokenização, homogeneização de maiúsculas, etiquetagem morfo-sintática, contagem de palavras e de n-gramas, medidas de associação e assim por diante. Esta segunda parte irá utilizar ferramentas (por exemplo, mwetoolkit e TreeTagger) e corpora (por exemplo, corpus CAMELEON) disponíveis gratuitamente para os exemplos e exercícios. O tutorial poderá interessar pessoas que desejam saber mais sobre corpora e processamento de linguagem natural. Conhecimento básico sobre como usar o terminal Unix é recomendado, mas não é estritamente necessário. O público-alvo inclui alunos (graduação e pós) e pesquisadores em ciência da computação e em linguística aplicada. Assim, eu não descreverei detalhes técnicos e estatísticos, mas privilegiarei uma introdução geral e ampla. O principal objetivo do tutorial é compartilhar ferramentas e métodos que permitem a exploração inteligente da informação textual, que poderão ser refinados e adaptados para diversas aplicações.

Tutorial 2 - Cassia Trojahn - IRIT, Université Toulouse II - Le Mirail (France)

```
Instituto de Informática da UFRGS, prédio 43413(67), sala 101
Terça-feira, 17 de dezembro de 2013, 13:30 às 16:30
```

An introduction to ontology matching

Ontology matching is seen as the solution to data heterogeneity in ontology-based applications, enabling the interoperability between them. Matching ontologies consists of finding corresponding entities (i.e., classes, properties, or instances) in different ontologies (usually one source ontology and one target ontology). It is a key process in the Linked Data age. Many different techniques implementing this process have been proposed, which are surveyed from different perspectives. The distinction between them is accentuated by the manner in which they exploit the features within an ontology. Whereas syntactic techniques consider measures of string similarity; semantic ones consider semantic relations usually on the basis of lexical oriented linguistic resources; while structural techniques consider term positions in the ontology hierarchy. In this tutorial, an introduction to the matching process and the different approaches for alignment will be presented together with a practical demonstration of the main functionalities of the Alignment API.

- Tutorial slides
- Tutorial material (download and unzip in Desktop)

Introdução ao Alinhamento de Ontologias

O processo de alinhamento de ontologias é visto como a solução para a heterogeneidade semântica em sistemas baseados em ontologias, permitindo a interoperabilidade entre os mesmos. O processo de alinhamento consiste em identificar correspondências entre as entidades (classes, propriedades, ou instâncias) de duas ontologias. Este processo é fundamental na era dos dados abertos vinculados (Linked Open Data). Diferentes técnicas de alinhamento têm sido propostas na literatura. A distinção entre elas é acentuada pela forma em que elas exploram os recursos e características de uma ontologia. Enquanto as técnicas sintáticas consideram medidas de similaridade entre cadeias de caracteres; as técnicas semânticas consideram as relações semânticas entre as entidades, geralmente com a ajuda de recursos lexicais; e técnicas estruturais exploram as posições das entidades nas hierarquias das ontologias. Neste tutorial, uma introdução ao processo de alinhamento sera apresentada, as diferentes abordagens para o problema serão discutidas e uma demonstração prática das principais funcionalidades da Alignment API sera realizada.

Some pictures