1ª. LISTA DE EXERCÍCIOS

1. Quais as vantagens e desvantagens dos tipos de anotação de corpora inline vs. stand-off?

Definição: Anotações stand-off são um tipo de anotação que são armazenadas em um arquivo diferente do local dos dados que estão sendo descritos por elas. É, portanto, o oposto da anotação inline, onde os dados e anotações são misturados em um único local ou arquivo (ECKART, 2012).

As vantagens da anotação **inline** é que ela consome menos processamento da máquina, por ser linear, cada tarefa será realizada após o término de uma outra. Por outro lado é mais lenta por definição.
As vantagens da anotação **stand-off** é em relação a velocidade de processamento que pode ser feito em um curto tempo, esse processo pode ser paralelizado, visto que: as anotações são armazenadas em um arquivo diferente do local dos dados que estão sendo descritos por elas. Por sua vez, essa técnica requer maior poder computacional.

Anotação Stand-off: flexibilidade

O texto primário pode ser usado sem anotações ou com anotações se necessário.
O usuário pode escolher trabalhar com uma anotação em particular independente do textos.
O córpus pode conter anotações de diferentes tipos, ou várias versões de um único tipo de anotação (por exemplo, múltiplas marcações de etiquetadores morfossintáticos (taggers)) sem problemas de compatibilidade.
O projeto pode distribuir anotações independentes do texto para download, porque as anotações possuem links para os dados originais (conteúdo), assim qualquer usuário que já fez download do córpus pode posteriormente somente baixar as novas anotações.
Os metadados são armazenados em um documento separado, usando âncoras de referência.
Alinhamento baseado em compensações de token ou caractere.
Dados primários são deixados intocados.

Inline:
Dados e metadados (anotação ou marcação) são combinados juntos em um único arquivo.

2. Ao seu ver, porque o formato híbrido, isto é, a anotação "em camadas" é o mais usado hoje em dia?

3. Quais tipos de anotação são realizadas pela ferramenta BRAT (http://brat.nlplab.org)?

Anota itens (palavras, sintagmas, trechos) e relações, anotação de entidade nomeada, anotação de entidade nomeada, Chunking, anotação de referência central, anotação de relação binária, anotação de evento.

Para quais problemas de mineração de textos ela pode ser usada?

Correferência, Dependência Sintática, Entidade Mencionada, Extração de Eventos, Identificação de Erros, Extração de Informação, Anotação de metáfora através de identificação ascendente, etc.

4. O que se entende por POS tagging? E qual a diferença entre POS tagging e parsing?

Uma tag POS (ou tag de parte da fala) é uma etiqueta especial atribuída a cada token (palavra) em um corpo de texto para indicar a parte da fala e muitas outras categorias gramaticais, como tempo, número (plural / singular), case etc. Os tags POS são usados em pesquisas de corpus e em ferramentas e algoritmos de análise de texto.

O POS tagging é uma técnica que mapeia uma palavra para uma chave(token) e classifica as palavras em suas classes gramaticais, já o parsing vai além disso, análise sintática (também conhecida pelo termo em inglês *parsing*) é o processo de analisar uma sequência de entrada (lida de um arquivo de computador ou do teclado, por exemplo) para determinar sua estrutura gramatical segundo uma determinada gramática formal. Produz a árvore de análise sintática mais provável de uma frase.

5. Discuta sobre as propriedades do parsing constituinte e o parsing de dependências

Um Constituency Parser é um tipo de analisador que divide a frase em subgrupos localizados em diferentes níveis de uma árvore de derivação, representando as combinações de elementos e suas ligações, de acordo com a gramática usada. Outro tipo de analisador sintático, frequentemente utilizado, é o Dependency Parser. As análises geradas por analisadores deste tipo não representam a frase em níveis hierárquicos, mas através de relacionamentos de seus elementos.

9. Use o stemmer em português disponível no link http://www.nilc.icmc.usp.br/nilc/tools/stemmer.html para gerar os stemming (radicais) de todas as palavras do texto em anexo “texto_pt.txt”.
Pergunta:
este stemmer funciona bem ao seu ver? Isto é, existem palavras que ele reduziu demais seu radical e em outras ele não fez nada? Discuta sobre isso.

O arquivo .exe do site **http://www.nilc.icmc.usp.br/nilc/tools/stemmer.html** não rodou em minha máquina, abria e fechava rapidamente.