UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO
Departamento de Ciências de Computação

SCC5908 Introdução ao Processamento de Língua Natural
SCC0633 Processamento de Linguagem Natural

Lista de exercícios: dos mais simples aos desafiadores

Para fixar o conteúdo visto nas aulas e nas leituras, assim como aprofundar em questões avançadas e relacionadas (para os mais interessados ☺)

1) Quais são os desafios de fazer uma máquina interagir fluentemente via linguagem natural como ocorre em filmes de ficção científica?

2) O que são os termos "língua" e "linguagem"? Por que a área no Brasil se chama Processamento de "Linguagem" Natural? Esse termo é apropriado? Explique.

3) O que é a área chamada de Linguística Computacional e qual sua relação com PLN?

4) Em que locais do mundo o português é falado? Como ele se posiciona entre as línguas mais faladas atualmente?

5) Qual a importância de preservar as línguas em extinção para as diferentes áreas (Computação, Linguística, Ciências Sociais, História, etc.)?

6) Pensamos em língua natural. Essa afirmação é verdadeira ou falsa? Justifique.

7) Quais são os principais marcos históricos do PLN e como eles se relacionam com a história da Inteligência Artificial?

8) A que se refere o termo "ambiguidade"? Em que níveis de tratamento linguístico ela pode ocorrer? Dê exemplos. Responda: como ela afeta o trabalho em PLN?

9) Responda: em sua opinião, é viável almejar falar uma língua sem ambiguidade?

10) Em termos dos interesses de PLN, trace paralelos entre os filmes "Prometheus" e "2001: Uma Odisseia no Espaço".

11) Quais são as particularidades da língua portuguesa em termos de origem, estruturação sentencial e paradigma flexional? Dica: veja o livro branco da língua portuguesa.

12) Quais são os níveis de conhecimento linguístico tradicionalmente tratados em PLN? Cite exemplos de análises linguísticas de cada nível e de como elas podem ser utilizadas em aplicações reais.

13) Os níveis de conhecimento linguístico tradicionalmente distinguidos em PLN são estanques. Essa afirmação é verdadeira ou falsa? Justifique e dê exemplos que evidenciem sua resposta.

14) Imagine que seu chefe lhe deu uma ordem: "entregue o relatório até o fim do dia!". Faça a análise dessa sentença de acordo com cada um dos níveis linguísticos de PLN.

15) Todo sistema tem suposições pragmáticas, mesmo que não modele explicitamente o conhecimento pragmático. Essa afirmação é verdadeira ou falsa? Dê exemplos que evidenciem sua resposta.

16) O que significa o termo "n-grama"? E "skip n-grama"? Qual a relação do termo com a representação tradicional de *bag of words* em aprendizado de máquina?

17) Quais são as 3 etapas do trabalho em PLN? Explique o que se faz em cada uma.

18) Suponha que você foi contratado para desenvolver um sistema de revisão ortográfica e gramatical para o português. O que deve ser feito em cada uma das etapas de PLN?

19) O que propõe a corrente "simbólica" do PLN? Quais suas vantagens e desvantagens?

20) Explique o "conflito" existente entre abordagens racionalistas e empiristas em PLN, destacando seus autores e principais pensamentos.

21) A que se referem os termos "competência" e "desempenho" linguísticos? Como eles afetaram o desenvolvimento de sistemas de PLN na história?

22) O que era o "mundo dos blocos" e qual sua importância em PLN?

23) Em PLN, o que se entende por "recurso", "ferramenta" e "aplicação"? Dê um exemplo de cada no contexto de tradução automática.

24) O que são e quais as diferenças entre sistemas de extração de informação, sumarização de textos e recuperação de informação? Exemplifique o que cada um faz.

25) O que é a tarefa de resolução de correferência? Qual sua relação com anáforas? Dê exemplos de sentenças com correferências e cite possíveis estratégias para identifica-las automaticamente.

26) O que é o modelo *noisy-channel*? Desenhe seu esquema e explique como ele poderia ser instanciado para a tarefa de sumarização automática.

27) O que são córpus? Do ponto de vista da Linguística, como eles podem ser usados? E do ponto de vista da Computação?

28) Cite as principais questões de pesquisa relacionadas à construção e anotação de córpus.

29) Qual a importância da concordância na anotação de um córpus em termos teóricos e práticos?

30) Por que o uso do Google pode ser ruim para a construção de córpus?

31) Explique brevemente a tipologia relacionada a córpus. Como o conhecido córpus MAC-MORPHO poderia ser classificado segundo essa tipologia?

32) Quais os desafios em obter representatividade e balanceamento em um córpus? Todo córpus balanceado é representativo? O inverso é verdade?

33) O que é a área chamada de Linguística de Córpus e como ela se relaciona com PLN?

34) Quem foi Zipf e o que sua principal lei diz? Crie e analise a curva de Zipf para o poema "No Meio do Caminho" de Carlos Drummond de Andrade, reproduzido abaixo:

No meio do caminho tinha uma pedra
tinha uma pedra no meio do caminho
tinha uma pedra
no meio do caminho tinha uma pedra.

Nunca me esquecerei desse acontecimento
na vida de minhas retinas tão fatigadas.
Nunca me esquecerei que no meio do caminho
tinha uma pedra
tinha uma pedra no meio do caminho
no meio do caminho tinha uma pedra.

35) Para que servem os chamados "cortes de Luhn"? Como eles podem ser determinados? Exemplifique para o poema de Carlos Drummond de Andrade.

36) Construa a matriz termo-contexto para as sentenças abaixo. Explique e exemplifique como esse tipo de matriz pode ser usado.

Uma garrafa de tesguino está aberta na mesa.
Todos gostam de tesguino.
Tesguino te deixa bêbado.
Cerveja é melhor em garrafa.
É fácil ficar bêbado com uma boa cerveja.

37) Quais as diferenças entre vetores esparsos e densos em PLN? Como eles são produzidos e quais são mais usados atualmente?

38) O que é um thesaurus? E uma ontologia? Como eles se diferem? Dê exemplos de conhecimentos codificados nesses recursos.

39) Como *word embeddings* podem auxiliar na construção de ontologias? Dê exemplos.

40) O que é entropia e qual sua relação com a dificuldade das tarefas de PLN?

41) O que são expressões regulares e como elas podem ser utilizadas em PLN? Construa expressões que reconheçam citações a datas em documentos, considerando suas variadas formas de ocorrência.

42) O que é uma "máquina morfológica"? Dê exemplos do que ela poderia fazer e como seria útil em aplicações de PLN.

43) O que é um transdutor? E uma rede de transição? Dê exemplos.

44) Construa um transdutor para fazer análise morfológica das palavras "modelo", "modelagem", "modelamento", "modelação" e "modelei", produzindo seus lemas, etiquetas morfossintáticas e informações associadas apropriadas (como gênero, número, pessoa, tempo, etc.).

45) Por que as classes gramaticais são ditas morfossintáticas e não morfológicas apenas? Forneça exemplos que evidenciem sua resposta.

46) O que é a iniciativa de *Universal Dependencies* e como ela se distingue dos demais tagsets existentes?

47) Quais as estratégias usuais para desenvolvimento de um tagger?

48) Construa o modelo de Markov oculto completo para a receita de chantilly abaixo:

> *Coloque o creme de leite bem gelado e o açúcar na batedeira e bata até que o creme comece a fazer ondas.*
> *Bata mais um pouquinho na velocidade mínima até chegar ao ponto (com ondas mais firmes).*
> *Quando estiver bem firme (sem cair da colher), é hora de parar de bater. O chantilly está pronto.*

49) Usando TBL (*Transformation-Based Learning*), construa um tagger completo usando como córpus a receita de chantilly. Faça as suposições que julgar necessárias e assuma que os recursos necessários para isso estão disponíveis.

50) O que são léxicos e quais as diferenças para um dicionário tradicional? O que são léxicos "legíveis por máquina" e "tratáveis por máquina"?

51) Faça a análise sintática de constituintes e de dependência da sentença "O pai abraçou o filho após o desastre". Responda: qual a diferença entre os dois tipos de análise? E por que ambas são consideradas de nível sintático?

52) Cite algumas aplicações da sintaxe na área de PLN, justificando sua importância nas aplicações.

53) O que é uma gramática livre de contexto? A língua portuguesa é livre de contexto?

54) Construa uma gramática completa para as instruções abaixo. Use o bom senso para tomar as decisões de representação relevantes, tentando manter a generalidade, legibilidade e elegância da gramática.

    Pressione ctrl+alt+delete.
    Selecione a opção de alterar uma senha.
    Digite sua senha antiga seguida de uma nova senha.
    Em seguida, digite a nova senha para confirmá-la.
    Pressione enter.

55) Implemente sua gramática do exercício anterior em DCG em algum ambiente de Prolog e teste se ela aceita a sentença inapropriada "pressione senha". Explique o comportamento resultante desse teste. Se necessário, usando atributos, altere sua gramática para que somente sentenças apropriadas sejam reconhecidas.

56) O que são métodos ditos *top-down* e *bottom-up* de parsing? Dê exemplos de como essas análises são feitas.

57) Com o método CKY e usando a sua gramática do exercício (4), faça o reconhecimento passo a passo da sentença "Selecione a senha nova". Faça as modificações necessárias na gramática.

58) O que é um *treebank* e como ele pode ser usado em PLN?

59) Como a ambiguidade semântica afeta a análise sintática? Dê exemplos que evidenciem sua resposta.

60) Com base no *treebank* Bosque, que compõe a Floresta Sintá(c)tica, e em seus conhecimentos adquiridos de expressões regulares, enriqueça suas regras gramaticais do exercício (4) com probabilidades, tornando-a uma gramática livre de contexto probabilística. Explique como você calculou as probabilidades necessárias.

61) Em relação ao exercício anterior, você acredita que seus cálculos sofreram com a esparsidade de dados? Justifique sua resposta e faça considerações sobre a uso de suavização nos cálculos.

62) Faça a análise semântica completa do trecho de texto a seguir, abrangendo desde as questões de semântica lexical até textual.

    Quero muito jogar tênis hoje. Vai depender da chuva.

63) Por que a representação AMR (*Abstract Meaning Representation*) tem tido destaque na área?

64) Que tipos de fenômenos são abrangidos pela semântica lexical? Dê exemplos e cite possíveis situações de uso desse conhecimento em sistemas de PLN.

65) *Word embeddings* capturam semântica lexical. Essa afirmação é verdadeira? Justifique sua resposta.

66) O que é uma *wordnet*? Por que ela é relevante? E o que seria uma *wordnet* terminológica?

67) O verbo "cantar" tem vários significados. Consulte alguma *wordnet* para o português para elencar seus significados possíveis. A partir disso, crie uma gramática enriquecida com atributos semânticos que permitam a interpretação semântica das sentenças abaixo, produzindo as fórmulas lógicas indicadas.

O tenor cantou no teatro da cidade.
∃e agente(e,tenor) ^ local(e,teatro_da_cidade)

O tenor cantou o hino da cidade.
∃e agente(e,tenor) ^ tema(e,hino_da_cidade)

A moça cantou o número do bingo.
∃e agente(e,moça) ^ tema(e,número_do_bingo)

O garoto cantou a menina na balada.
∃e agente(e,garoto) ^ experienciador(e,menina) ^ local(e,balada)

68) O que é uma *framenet* e qual sua diferença para uma *wordnet*?

69) O que é o PropBank? Faça considerações sobre os desafios de construir um recurso como esse para uma língua qualquer.

70) Considere o domínio dos dispositivos tecnológicos móveis, que inclui, por exemplo, smartphones, notebooks, netbooks, tablets, tocadores de mp3, etc. Construa uma ontologia para este domínio. Se desejar, projete um pequeno córpus para lhe apoiar nessa tarefa e/ou consulte dicionários, léxicos e outros recursos linguístico-computacionais que julgar necessários.

71) O que são entidades nomeadas e quais seus tipos? Entidades nomeadas são o mesmo que classes semânticas em PLN? Justifique sua resposta.

72) No repositório REPENTINO (acesse http://labclup.letras.up.pt/repentino/), faz-se a diferença entre os termos "entidade nomeada" e "entidade mencionada". Explore esse repositório e explique (i) como ele foi construído, (ii) para que ele foi aplicado e (iii) qual a diferença entre os termos citados?

73) Atualmente, um dos temas mais pesquisados em PLN é a análise de sentimentos, que, em seu sentido mais amplo, inclui desde tarefas de detecção de polaridade de trechos textuais a traçar perfil psicológico de pessoas com base nos materiais escritos que produzem. Explique que níveis de conhecimento tratados em PLN estão envolvidos nesse tipo de aplicação.

74) Suponha que você foi contratado para desenvolver um sistema de revisão ortográfica, gramatical e estilística para um aplicativo de mensagens de smartphones. Conhecendo as três etapas já clássicas de PLN e os conceitos e processos linguístico-computacionais de cada nível de conhecimento da área, faça uma proposta de como projetar e desenvolver um sistema desse tipo.

75) O Google Tradutor é, sem dúvida alguma, uma aplicação linguístico-computacional de muito destaque. Entretanto, como todo sistema que lida com língua natural, tem limitações. Com todo seu conhecimento adquirido em PLN, faça alguns testes no sistema online de tradução do Google e, com base nos exemplos testados, indique pontos em que ele poderia melhorar, explicitando em que nível de conhecimento linguístico investir e possíveis estratégias para isso.