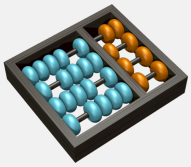
Instituto de Computação - UNICAMP

MC102 - Algoritmos e Programação de Computadores



Laboratório 07: Hipótese de Sapir–Whorf

Segunto Semestre de 2017 - Turmas Coordenadas

Peso da Atividade: 2

Prazo de Entrega: 20 de Outubro de 2017 às 23:59:59

Conteúdo

- Contexto
- Tarefa
- Observações da Tarefa
- Exemplos
- Observações Gerais
- Critérios Importantes

Contexto

"The Sapir-Whorf hypothesis is the theory that the language you speak determines how you think"

Dr. Louise Banks em *A Chegada - 2016*.

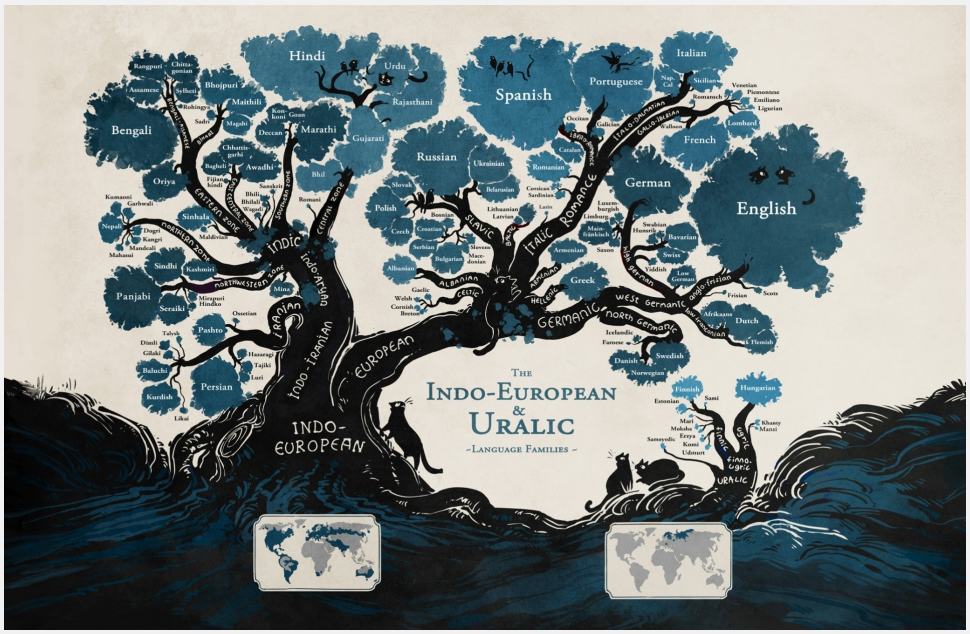


Figura 1: Mapa ilustrado por Minna Sundberg's que retrata a relação entre línguas Indo-Européias e Urálicas. Extraído daqui.

A Professora Louise Banks, uma renomada pesquisadora do Instituto de Estudos da Linguagem da UNICAMP, terminou sua aula daquele primeiro de dezembro de 2048 bem cansada. Apesar daquela aula ter sido sobre sua linha de pesquisa - estudo da expansão das línguas românicas (por exemplo, o Português, Francês, Romeno, Espanhol e Italiano) e a influência da linguagem na evolução da nossa sociedade - a turma grande e as quatro horas de aula já no final do dia a tinham deixado esgotada.

Após a aula, na sua sala, ela voltou-se para o seu estudo da hipótese de Sapir-Whorf (ou relativismo linguístico). Esta teoria foi proposta por volta de 1930 por Edward Sapir e Benjamin Lee Whorf e tem por princípio a hipótese de que a estrutura da linguagem afeta a cognição dos seus falantes. Uma versão fraca da hipótese apenas afirma que a linguagem afeta o pensamento e decisões, enquanto a versão forte afirma que a linguagem determina a forma que as pessoas pensam.

Aquela tarde do dia primeiro de dezembro foi o dia em que a vida da Professora Louise mudaria para sempre. Após iniciar a correção das provas finais da disciplina HL917 - Escrita Extraterreste Não Linear, Louise é abordada por Karen, uma professora do Instituto de Física.

– Professora Louise?
– Sim, por favor, entre.
– Precisamos de sua ajuda. Acreditamos ter recebido um contato extraterreste.

Louise fica em choque. *Isso não é possível.* Ela pensa. *Um primeiro contato...*

– Procuramos a senhora por já ter estudado este tipo de linguagem. Acreditamos que conseguirá decifrar facilmente o que recebemos.
– Mas... eu nunca... como assim receberam um contato extraterrestre?
– Trouxemos para a senhora pois acreditamos que a senhora já tenha recebido outros contatos desta forma. Por meio da interceptação de ondas eletromagnéticas vindas do espaço, observamos um padrão circular de interferências.

Karen mostra para Louise um papel com desenhos, como o exibido na imagem abaixo.

Figura 2: Exemplo de símbolo exibido por Karen para a Louise.

Louise pega o papel, olha atentamente e, enquanto lágrimas caem dos seus olhos, numa mistura de emoção e ansiedade, diz:

– É tão lindo. Esta é a primeira vez que vejo estes símbolos. Nunca antes um contato foi estabelecido. Nossos estudos são apenas teóricos, mas, olhando para isso, não tenho dúvidas...

Karen passa a não entender nada e pergunta:
– Este é o primeiro contato? Vocês nunca receberam isso antes? O que estes símbolos dizem?
Louise levanta a cabeça, olha nos olhos de Karen e diz:
– Nossa sociedade ainda precisa evoluir muito para conseguir entender...

Essa tarefa é baseada no filme 'A Chegada - 2016' de *Eric Hisserer*, que por sua vez foi baseado no conto *Story of Your Life* (1998) de Ted Chiang.

Referências:

- How Realistic Is the Way Amy Adams' Character Hacks the Alien Language In Arrival? We Asked a Linguist. Link.
- What is the Sapir-Whorf hypothesis?. Link.
- Linguistic relativity. Link.

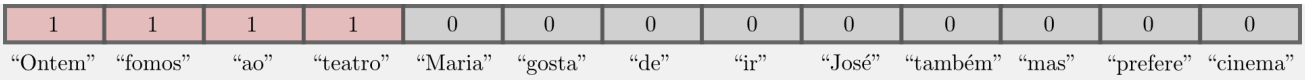
Tarefa

Louise está longe de conseguir decifrar essa linguagem extraterrestre. Contudo, ela vem tentando analisar as relações entre as palavras de uma linguagem para determinar a frequência de ocorrência simultânea entre as palavras, principalmente entre marcadores temporais e outras palavras. Ela acredita que essa análise pode posteriormente ajudar a decifrar a semântica dos símbolos do primeiro contato extraterrestre. Neste laboratório, iremos construir um protótipo para essa análise que investiga a relação entre marcadores temporais e outras palavras em um texto.

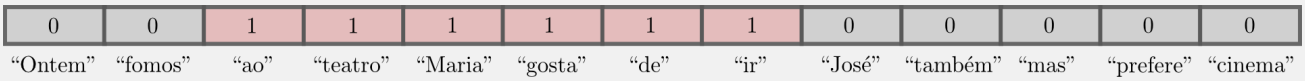
Dado um texto e algumas palavras a serem analisadas, chamadas de palavras-de-busca, iremos verificar qual marcador temporal se relaciona mais com cada uma das palavras sendo analisadas. Os marcadores temporais que serão utilizados são os seguintes: *ontem*, *hoje*, *amanhã*, *agora*, *logo*, *cedo*, *tarde*, *breve*, *nunca*, *sempre*, *jamais*.

A entrada do seu programa é composta por: o número de frases do texto a ser analisado; as frases do texto, uma em cada linha; o número de palavras-de-busca seguido destas palavras, uma por linha.

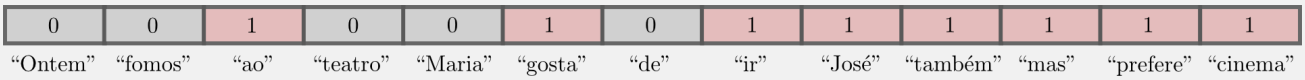
Dado um texto de entrada com n frases, iremos determinar a frequência das palavras em cada frase do texto de entrada. Para representar a frequência das palavras utilizaremos uma representação baseada em vetor, em que cada posição do vetor representa uma palavra do texto e o valor daquela posição a frequência da palavra correspondente em uma determinada frase. Por exemplo, considere o seguinte texto: "*Ontem fomos ao teatro. Maria gosta de ir ao teatro. José também gosta, mas prefere ir ao cinema.*". Neste caso, temos o seguinte conjunto de palavras: *Ontem*, *fomos*, *ao*, *teatro*, *Maria*, *gosta*, *de*, *ir*, *José*, *também*, *mas*, *prefere*, *cinema*. Dessa forma, a frase "*Ontem fomos ao teatro.*" é representada pelo seguinte vetor de frequência:



a frase "*Maria gosta de ir ao teatro*" por:



e a frase "*José também gosta, mas prefere ir ao cinema.*" por:



Você deve analisar a ocorrência de `p` palavras-de-busca com os marcadores temporais definidos previamente, onde 0 ≤ `p` ≤ 10. Depois de determinar as frequências das palavras para cada frase do texto, você irá calcular, para cada palavra-de-busca, a taxa de ocorrência em que aquela palavra ocorreu simultaneamente com algum marcador temporal, considerando todas as ocorrências da palavra no texto.

Considere que `palavra_1` seja uma das `p` palavras-de-busca e que estamos analisando essa palavra com o marcador *ontem*. Você irá determinar a quantidade de vezes que a `palavra_1` ocorre no texto, denominada *total_ocorrencias*; e a quantidade de vezes em que ela ocorre simultaneamente em uma frase com o marcador *ontem*, denominada *ocorrencia_simultanea*. A taxa de ocorrência para `palavra_1` e o marcador *ontem* é então dada por *(ocorrencia_simultanea*100)/total_ocorrencias*.

A saída do seu programa será o marcador temporal que mais se relaciona com cada uma das `p` palavras-de-busca no texto analisado e a frequência das palavras considerando todas as frases do texto. A primeira parte da saída deve ser impressa no seguinte formato:
<`palavra_1`> se relaciona com <`marcador temporal`> em `X` % das ocorrencias.

Por exemplo, se analisarmos a palavra *teatro* no texto acima, teríamos como saída:
<teatro> se relaciona com <ontem> em 50 % das ocorrencias.

Se houver mais de um marcador temporal com a mesma taxa de ocorrência, imprima todos os casos considerando a ordem dos marcadores temporais apresentada anteriormente. Se não houver ocorrência simultânea da palavra-de-busca com algum marcador temporal, imprima a seguinte mensagem para a palavra correspondente:
<`palavra_1`> nao se relacionou com nenhum marcador temporal.
Consulte a primeira parte da saída nos exemplos de execução para ver um exemplo deste caso.

Para a segunda parte da saída, você deve imprimir a frequência das palavras do texto (isto é, nas frases) utilizando a representação baseada em vetor. A primeira linha representa todas as palavras do texto, em minúsculo, na ordem em que foram apresentadas na entrada e a segunda linha as frequências correspondentes. Consulte os exemplos de execução para ver o formato de saída.

Observações da Tarefa

- O texto de entrada não contém acentos e os números estão apresentados por extenso.
- Note que a saída não contém acentos e é apresentada em letras minúsculas.
- Note que "Banana", "banana" e "BaNaNa" são as mesmas palavras. Utilize a função disponibilizada para transformar todas as palavras em minúsculas e remover os caracteres especiais da entrada.
- Você *deve* inserir um espaço após cada palavra da lista de palavras do texto e após cada frequência.
- Você *deve* inserir uma quebra de linha `\n` após cada relação da palavra-de-busca com algum marcador temporal, após a lista de palavras do texto, e a frequência das palavras no texto.
- Uma palavra tem no máximo `20` caracteres.
- Uma frase tem no máximo `250` caracteres.
- Um texto de entrada tem no máximo `50` frases e `250` palavras.
- Os cabeçalhos das funções a serem implementadas estão descritas no arquivo auxiliar `lab07.h`. Não é permitido alterar este arquivo, nem os parâmetros das funções.
- Você pode implementar outras funções que desejar, além das que estão declaradas no arquivo `lab07.h`.

Exemplos

Notas:

Textos em `azul` designam dados de entrada, isto é, que devem ser lidos pelo seu programa.
Textos em `preto` designam dados de saída, ou seja, que devem ser impressos pelo seu programa.

Exemplo de execução 1:

```
3
Ontem fomos ao teatro.
Maria gosta de ir ao teatro.
Jose tambem gosta, mas prefere ir ao cinema.
1
teatro

<teatro> se relaciona com <ontem> em 50 % das ocorrencias.
ontem fomos ao teatro maria gosta de ir jose tambem mas prefere cinema
1 1 3 2 1 2 1 2 1 1 1 1 1
```

Exemplo de execução 2:

```
5
Amanha vou acordar cedo.
Logo cedo vou tomar meu cha.
Ao meio dia vou almocar.
De tarde lanchar.
E de noite jantar.
3
cha
lanchar
jantar

<cha> se relaciona com <logo> em 100 % das ocorrencias.
<cha> se relaciona com <cedo> em 100 % das ocorrencias.
<lanchar> se relaciona com <tarde> em 100 % das ocorrencias.
<jantar> nao se relacionou com nenhum marcador temporal.
amanha vou acordar cedo logo tomar meu cha ao meio dia almocar de tarde lanchar e noite jantar
1 3 1 2 1 1 1 1 1 1 1 2 1 1 1 1 1
```

Observações Gerais

- O número máximo de submissões é 20.
- O arquivo `lab07.c` deve conter todo o seu programa, incluindo o corpo de todas as funções.
- Para a realização dos testes automáticos, a compilação se dará da seguinte forma: `gcc lab07.c -o lab07 -Wall -Werror -ansi -pedantic` .
- Não se esqueça de incluir no início do programa uma breve descrição dos objetivos, da entrada, da saída, seu nome, RA e turma.
- Após cada submissão, você deve aguardar um minuto até poder submeter seu trabalho novamente.
- Ao final deste laboratório, você terá aprendido como utilizar cadeias de caracteres (*strings*) e funções.

Critérios Importantes

O **não** cumprimento dos critérios abaixo acarretará em **nota zero na atividade**, independentemente dos resultados dos testes do SuSy.

- Sua solução deve atender todos os requisitos definidos no enunciado.
- Não serão aceitas soluções contendo estruturas não vistas em sala (para este laboratório, poderão ser utilizadas apenas variáveis simples, vetores, matrizes, operações de entrada e saída, operações aritméticas, desvios condicionais, estruturas de repetição, strings e funções).
- Não é permitido o uso de `continue` e `break` (exceto em estruturas do tipo *switch-case*).
- Não é permitido o uso de variáveis globais.
- Cada função deve conter apenas um único `return`.
- Os únicos cabeçalhos aceitos para inclusão são `stdio.h`, `string.h` e `lab07.h`.