MC102W - Algoritmos e Programação de Computadores

Lab10: Respondenator 3000

Prazo: 31 de Maio de 2020 **Peso da Atividade:** 4

> *"Talvez até não fizesse muita diferença se eles soubessem exatamente quanto poder exercia o presidente da Galáxia: absolutamente nenhum. Apenas seis pessoas na Galáxia sabiam que a função do presidente não era exercer poder, e sim desviar a atenção do poder."* (*O guia do mochileiro das galáxias*)

O ano é 2050 e Eleanor foi eleita presidente do Brasil. Porém, sua vitória foi devido a um tremendo erro do sistema, pois ela sequer tinha se candidatado e não entende nada sobre liderar um país. Para não ter que recorrer a uma nova eleição e nem admitir a falha no sistema eleitoral, todos os outros poderes decidem deixar ela governar, mas sempre com absoluto controle das suas declarações para que o povo não perceba que ela não é capacitada para o cargo. As declarações são elaboradas pelo Chidi, doutor em uma importantíssima ciência chamada Filosofia que infelizmente não é mais estudada devido aos ~~cortes~~... digo, contingenciamentos dos recursos dos anos 20. Em todas as coletivas de imprensa, Eleanor sempre leva vários papéis de respostas ordenadas de acordo com o fluxo que a entrevista costuma seguir. Como a lei de Murphy parece ter sido feita para Eleanor, em um belo dia, em uma coletiva extremamente importante, ela derruba todos os papéis e tira todos da ordem. Por sorte ou sabedoria, ela acredita muito em ciência e fez fortes investimentos na área. Sua equipe técnica, da qual você faz parte, desenvolveu uma tecnologia revolucionária: o Respondenator 3000, prevendo esses momentos. O sistema recebe a pergunta feita pela imprensa e automaticamente retorna a resposta mais adequada para a questão. Ufa! Já imaginou a presidente respondendo sobre economia quando a pauta é saúde?

Tarefa

Sua tarefa neste laboratório é implementar o Respondenator 3000. O objetivo do sistema é comparar uma pergunta com **n** possíveis respostas e encontrar a resposta mais adequada. Para fazer a comparação, você precisa primeiramente processar os textos (tanto a pergunta quanto as respostas). O processamento tem as seguintes etapas:

1) **Padronização:** Na primeira etapa, todas as letras devem ser convertidas para minúsculas.
2) **Tokenização:** Em seguida, o texto é dividido em palavras ou *tokens*.

3) **Limpeza:** Nesta etapa você deve remover elementos que não caracterizam o texto, isto é, pontuações e palavras que não são específicas de um assunto como “a”, “um”, “quem”, entre outras. Essas palavras são chamadas de *stop words* ou palavras vazias e variam de acordo com o idioma e com a aplicação. As listas de palavras vazias e pontuações consideradas nesse laboratório estão nos arquivos auxiliares (PontStop.py) e devem ser importadas usando:

    from PontStop import *

4) **Reescrita:** Para facilitar a comparação, algumas palavras devem ser trocadas pelo sinônimo/representante padrão. Por exemplo, as palavras ‘é’, ‘são’ e ‘foram’ devem ser substituídas pelo representante ‘ser’; as palavras ‘moradia’ e ‘lar’ devem ser substituídas pelo sinônimo ‘casa’. O dicionário de sinônimos será informado em cada caso de teste.
5) **Representação:** Nesta etapa, você deve criar um descritor do texto que consiste do conjunto de palavras resultantes das etapas anteriores, **sem repetição**.

Após o processamento dos textos, o sistema julgará as respostas da seguinte forma:
Uma resposta será adequada para determinada pergunta se o descritor da pergunta estiver contido no descritor da resposta.

Exemplo:
Pergunta: Quem era o presidente do Brasil em 1990?

- Padronização: ‘quem era o presidente do brasil em 1990?’
- Tokenização: 'quem', 'era', 'o', 'presidente', 'do', 'brasil', 'em', '1990?'
- Limpeza - pontuação: 'quem', 'era', 'o', 'presidente', 'do', 'brasil', 'em', '1990'
- Limpeza - *stop words*: 'era, 'presidente', 'brasil', '1990'
- Reescrita: ‘ser', 'presidente', 'brasil', '1990'
- Descritor: 'ser', 'presidente', 'brasil', '1990'

Resposta 1: Fernando Affonso Collor de Mello, mais conhecido como Fernando Collor, foi o presidente do Brasil em 1990 e sofreu um processo de impeachment em 1992.

- Padronização: ‘fernando affonso collor de mello, mais conhecido como fernando collor, foi o presidente do brasil em 1990 e sofreu um processo de impeachment em 1992.’
- Tokenização: 'fernando', 'affonso', 'collor', 'de', 'mello,', 'mais', 'conhecido', 'como', 'fernando', 'collor,', 'foi', 'o', 'presidente', 'do', 'brasil', 'em', '1990', 'e', 'sofreu', 'um', 'processo', 'de', 'impeachment', 'em', '1992.'
- Limpeza - pontuação: 'fernando', 'affonso', 'collor', 'de', 'mello', 'mais', 'conhecido', 'como', 'fernando', 'collor', 'foi', 'o', 'presidente', 'do', 'brasil', 'em', '1990', 'e', 'sofreu', 'um', 'processo', 'de', 'impeachment', 'em', '1992'
- Limpeza - *stop words*: 'fernando', 'affonso', 'collor', 'mello', 'conhecido', 'fernando', 'collor', 'foi', 'presidente', 'brasil', '1990', 'sofreu', 'processo', 'impeachment', '1992'
- Reescrita: 'fernando', 'affonso', 'collor', 'mello', 'conhecer', 'fernando', 'collor', 'ser', 'presidente', 'brasil', '1990', 'sofrer', 'processo', 'impeachment', '1992'
- Descritor: ‘fernando', 'affonso', '1990', 'conhecer', 'brasil', 'processo', 'impeachment', 'presidente', 'ser', 'sofrer', 'collor', '1992', 'mello'

Resposta 2: Itamar Augusto Cautiero Franco, mais conhecido como Itamar Franco, foi o presidente do Brasil entre 1992 (após o impeachment do presidente Fernando Collor) e 1995.

- Padronização: 'itamar augusto cautiero franco, mais conhecido como itamar franco, foi o presidente do brasil entre 1992 (após o impeachment do presidente fernando collor) e 1995.'
- Tokenização: 'itamar', 'augusto', 'cautiero', 'franco,', 'mais', 'conhecido', 'como', 'itamar', 'franco,', 'foi', 'o', 'presidente', 'do', 'brasil', 'entre', '1992', '(após', 'o', 'impeachment', 'do', 'presidente', 'fernando', 'collor)', 'e', '1995.'
- Limpeza - pontuação: 'itamar', 'augusto', 'cautiero', 'franco', 'mais', 'conhecido', 'como', 'itamar', 'franco', 'foi', 'o', 'presidente', 'do', 'brasil', 'entre', '1992', 'após', 'o', 'impeachment', 'do', 'presidente', 'fernando', 'collor', 'e', '1995'
- Limpeza - *stop words*: 'itamar', 'augusto', 'cautiero', 'franco', 'conhecido', 'itamar', 'franco', 'foi', 'presidente', 'brasil', '1992', 'impeachment', 'presidente', 'fernando', 'collor', '1995'
- Reescrita: 'itamar', 'augusto', 'cautiero', 'franco', 'conhecer', 'itamar', 'franco', 'ser', 'presidente', 'brasil', '1992', 'impeachment', 'presidente', 'fernando', 'collor', '1995'
- Descritor: 'itamar', 'impeachment', 'fernando', 'collor', 'franco', '1992', '1995', 'augusto', 'presidente', 'ser', 'brasil', 'cautiero', 'conhecer'

A resposta 1 é escolhida pois, diferente da resposta 2, ela contém o descritor da pergunta.

Observação:

- ❏ Palavras que não estiverem no dicionário de sinônimos devem permanecer na sua forma original.
- ❏ Não existem empates nos casos de teste, isto é, no máximo uma resposta será adequada para a pergunta dada.

O programa deve ser implementado em Python e deve utilizar os conceitos de listas, dicionários, conjuntos e tuplas.

Entrada

A entrada do programa é dividida em dois blocos. O primeiro bloco consiste do dicionário de sinônimos e é delimitado pelos caracteres “{“ e “}”. Cada linha entre as chaves contém um representante e sua lista de sinônimos:

{
<representante_1>: [<sinonimo1>,<sinonimo2>,...]
...
}

O segundo bloco é composto de uma pergunta, o número **n** de resposta e as respostas:
<pergunta>
n
<resposta_1>
<resposta_2>
...
<resposta_n>

Observação:

- Cada representante no dicionário tem pelo menos um sinônimo na lista.
- n >= 2 (garantido em todos os casos de teste).
- Não existem quebras de linha nos textos (perguntas e respostas), apenas entre eles.

Saída

O seu programa deve imprimir os descritores da pergunta e de cada resposta, seguindo a ordem de entrada. Essa impressão deve ser no formato "Descritor pergunta: <conjunto_separado_por_virgula>" e "Descritor resposta i: <conjunto_separado_por_virgula>", com i começando em 1. O descritor deve ser impresso de forma ordenada. Para isso, utilize a função **sorted(conjunto)** que retorna uma lista**.** Ao final do processo, deve imprimir 'A resposta para a pergunta "<pergunta>" é: "<resposta>"', se encontrar uma resposta adequada. Caso contrário, o programa deve imprimir 'A resposta para a pergunta "<pergunta>" é: "42"'. Uma linha em branco deve ser impressa entre a impressão dos descritores e a mensagem final.

Exemplo

Exemplo 1:

Entrada

```
{
ser:era,foi
conhecer:conhecido
sofrer:sofreu
}
Quem era o presidente do Brasil em 1990?
2
Fernando Affonso Collor de Mello, mais conhecido como Fernando Collor, foi o presidente do Brasil em 1990 e sofreu um processo de impeachment em 1992.
Itamar Augusto Cautiero Franco, mais conhecido como Itamar Franco, foi o presidente do Brasil entre 1992 (após o impeachment do presidente Fernando Collor) e 1995.
```

Saída

```
Descritor pergunta: 1990,brasil,presidente,ser
Descritor resposta 1:
1990,1992,affonso,brasil,collor,conhecer,fernando,impeachment,mel
lo,presidente,processo,ser,sofrer
Descritor resposta 2:
1992,1995,augusto,brasil,cautiero,collor,conhecer,fernando,franco
,impeachment,itamar,presidente,ser

A resposta para a pergunta "Quem era o presidente do Brasil em
1990?" é "Fernando Affonso Collor de Mello, mais conhecido como
Fernando Collor, foi o presidente do Brasil em 1990 e sofreu um
processo de impeachment em 1992."
```

Exemplo 2:

Entrada

```
{
ser:é,era,são
expulsar:expulsou
ir:vai
acreditar:acredita
acidente:acidentes
banhista:banhistas
brasileir:brasileiro,brasileira,brasileiros,brasileiras
carro:carros
cidade:cidades
cidadão:cidadãos,cidadãs,cidadã
cumprir:cumpra
destino:destinos
engenheiro:engenheiros
lancha:lanchas
morrer:morra
moto:motos
deixar:deixe
ocorrer:ocorre
pedestre:pedestres
fazer:fará
gerar:gerando
permitir:permitirá
poder:possa
urbano:urbanos
alien:aliens
}
Qual é sua proposta para o transporte público?
4
Tudo vai na hora certa. Uma pergunta para o senhor: o senhor era
do partido comunista, que não acredita em Deus. O PT expulsou
quem é... quem não é... contra o aborto. O senhor é contra ou
```

```
favor do aborto?
Os carros são como as lanchas, as motos são como os Jet Skis e os
pedrestes são como os banhistas. E, assim como no trânsito das
cidades, no mar também ocorre acidentes. Nesse verão, cumpra a
norma de segurança no mar. Não deixe que a alegria de alguém
morra na praia.
O transporte público é um serviço fundamental na vida dos
cidadãos e cidadãs brasileiros. A nossa proposta é implantar nos
ônibus urbanos uma nova tecnologia desenvolvida pelos nossos
engenheiros que permitirá que o ônibus se teletransporte. Isso
fará com que o povo brasileiro possa chegar muito mais rápido aos
seus destinos, gerando mais qualidade de vida.
Aliens.
```

Saída

```
Descritor pergunta: proposta,público,ser,transporte
Descritor resposta 1:
aborto,acreditar,certa,comunista,contra,deus,expulsar,favor,hora,
ir,partido,pergunta,pt,senhor,ser
Descritor resposta 2:
acidente,alegria,banhista,carro,cidade,cumprir,deixar,jet,lancha,
mar,morrer,moto,norma,ocorrer,pedrestes,praia,segurança,ser,skis,
trânsito,verão
Descritor resposta 3:
brasileir,chegar,cidadão,desenvolvida,destino,engenheiro,fazer,fu
ndamental,gerar,implantar,nova,permitir,poder,povo,proposta,públi
co,qualidade,rápido,ser,serviço,tecnologia,teletransporte,transpo
rte,urbano,vida,ônibus
Descritor resposta 4: alien

A resposta para a pergunta "Qual é sua proposta para o transporte
público?" é "O transporte público é um serviço fundamental na
vida dos cidadãos e cidadãs brasileiros. A nossa proposta é
implantar nos ônibus urbanos uma nova tecnologia desenvolvida
pelos nossos engenheiros que permitirá que o ônibus se
teletransporte. Isso fará com que o povo brasileiro possa chegar
muito mais rápido aos seus destinos, gerando mais qualidade de
vida."
```

Critérios específicos

Os seguintes critérios específicos sobre o envio, implementação e execução devem ser satisfeitos.

i. Submeter no *SuSy* o arquivo:

⇒ `lab10.py`: Arquivo contendo todo o seu programa.

ii. Não serão aceitas soluções contendo estruturas não vistas em sala, exceto as indicadas neste enunciado.

Observações gerais

No decorrer do semestre haverá 3 tipos de tarefas no *SuSy* (descritas logo abaixo). As tarefas possuirão os mesmos casos de testes abertos e fechados, no entanto o número de submissões permitidas e prazos são diferentes. As seguintes tarefas estão disponíveis no *SuSy*:

- **Lab10-AmbienteDeTeste**: Esta tarefa serve para testar seu programa no *SuSy* antes de submeter a versão final. Nessa tarefa, tanto o prazo quanto o número de submissões são ilimitados, porém os arquivos submetidos aqui **não serão corrigidos**.
- **Lab10-Entrega**: Esta tarefa tem limite de uma **única** submissão e serve para entregar a **versão final** dentro do prazo estabelecido para o laboratório. Não use essa tarefa para testar o seu programa e submeta aqui apenas quando não for mais fazer alterações no seu programa.
- **Lab10-ForaDoPrazo**: Esta tarefa tem limite de uma **única** submissão e serve para entregar a versão final fora prazo estabelecido para o laboratório. Esta tarefa irá substituir a nota obtida na tarefa **Lab10-Entrega** apenas se o aluno tiver realizado as correções sugeridas no *feedback* ou caso não tenha enviado anteriormente na tarefa **Lab10-Entrega**.

Avaliação

Este laboratório será avaliado da seguinte maneira: a nota será proporcional ao número de casos **fechados** para os quais o seu programa gerou a resposta correta, **desde que os critérios indicados neste enunciado tenham sido atendidos**. Se o programa apresentou resposta correta para todos os casos, a nota será 10; caso contrário será p*10, onde p é o percentual de respostas corretas. A nota também poderá sofrer descontos de acordo com a qualidade do programa apresentado. Assim, mesmo que o código seja capaz de resolver todos os casos de teste fechados, a nota final ainda pode ser menor do que 10. Por isso, acrescente comentários explicativos, utilize variáveis sugestivas e faça um código claro e de acordo com o que foi solicitado.

Testando seu programa

Para testar se a solução do seu programa está correta, basta seguir o exemplo abaixo no terminal do Linux.

```
python lab10.py < arq01.in > arq01.out
diff arq01.out arq01.res
```

O `arq01.in` é a entrada e `arq01.res` é a saída esperada, ambos disponíveis no *SuSy*. O `arq01.out` é a saída gerada pelo seu programa. Após o prazo, os casos de teste fechados serão liberados e podem ser baixados e testados da mesma forma que os testes abertos.