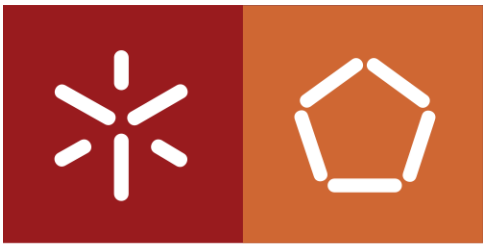

Universidade do Minho
Escola de Engenharia

Scripting no Processamento de Linguagem Natural
TP3 - 7. Conversor Fonético

Ricardo Neves A78764

Junho 2020

1 Introdução

Os sons das palavras de cada língua do mundo podem ser representados usando alfabetos fonéticos, como é o caso do IPA (International Phonetic Alphabet). Em português, a maior parte das letras pode resultar em diferentes sons, por exemplo, cara ou leve.

Assim, foi desenvolvido um script em Python capaz de traduzir uma palavra ou frase, de português para linguagem IPA.

```
[ricardo@neves] $ ~/Universidade/PLC/SPLN/ConversorFonetico
>> python3 conversor.py 'universidade'

Processing text...

Number of words: 1
Plain text: universidade
Stressed syllables: u·ni·ver·si·'da·de
IPA: u.ni.vɨɾ.si.'da.dɨ
Solution: u.ni.vɨɾ.si.d'a.dɨ
Comparison: 17/17, 100.00%
```

2 Desenvolvimento do sistema

Como podemos ver no exemplo acima, a palavra (ou frase) a ser traduzida é dada como argumento pelo utilizador. Se o input se tratar de mais do que uma palavra, é obrigatória a colocação da frase dentro de aspas. Caso contrário, o programa não aceitará o input introduzido, exibindo uma mensagem de erro.

2.1 Divisão silábica

A primeira etapa na codificação do programa foi dividir o input em sílabas. Recorrendo ao Dicionário de Divisão Silábica, disponível no Portal da Língua Portuguesa, foi possível transcrever as regras para código Python. Aqui, o script percorre toda a palavra, colocando pontos dividindo-a nas suas sílabas fonéticas.

Aqui, posso afirmar que o grau de acerto da divisão silábica está perto da perfeição, onde o resultado final encontra-se correto quase 100% das vezes (salvo situações extremamente específicas).

2.2 Encontrar sílaba tónica

Depois de dividir o input pelas suas sílabas, achei interessante encontrar a sílaba tónica de cada palavra. Ou seja, se a palavra não tiver nenhum acento em nenhuma das suas vogais, o script é capaz de identificar a sílaba tónica de cada palavra do input. Esta identificação será útil aquando da conversão da palavra portuguesa para caractéres IPA.

Assim, obtemos a divisão silábica e identificação da sílaba tónica em 'universidade':

```
Stressed syllables: u·ni·ver·si·'da·de
```

2.3 Conversão para IPA

A conversão das palavras portuguesas para IPA foi, sem dúvida, a parte mais complicada de toda a implementação. Isto deve-se à existência de dezenas de regras de conversão. Estas regras estão documentadas no portal 'Learn European Portuguese', que identifica as várias regras para as consoantes e vogais (acentuadas ou não) do alfabeto português. No entanto, ao longo da implementação deste sistema, pude reparar que existem algumas regras que não se encontram documentadas no portal, talvez devido ao facto de serem bastante específicas e raras.

Esta tradução não se encontra rigorosamente correta, pelos motivos identificados acima, mas posso afirmar que o resultado final é bastante positivo, dado o grande número e complexidade das regras.

Podemos ver o resultado da tradução da palavra 'universidade' para IPA:

```
IPA: u.ni.vɨɾ.si.'da.dɨ
```

2.4 Comparação com solução

Na reta final da implementação, o professor André disponibilizou um conjunto de palavras portuguesas e a sua tradução para IPA (em formato .json), de modo a comparar o meu resultado com a solução correta. Deste modo, foi fácil comparar os resultados e avaliar os mesmos.

Aqui, podemos ver a comparação entre o resultado obtido e a solução para a palavra 'universidade'. Neste caso, a avaliação final para este input é de 100%:

```
IPA: u.ni.vɨɾ.si.'da.dɨ
Solution: u.ni.vɨɾ.si.d'a.dɨ
Comparison: 17/17, 100.00%
```

De salientar que, apesar das milhares de palavras presentes no ficheiro fornecido, muitas outras palavras portuguesas não tem solução documentada. Nestes casos, o programa irá apenas apresentar o resultado final calculado pelo próprio.

3 Resultados obtidos

Explicado todo o raciocínio e codificação por detrás deste sistema, passemos então para apresentação e discussão dos resultados obtidos.

Pode-se pensar que o tempo de execução desta implementação seja relativamente lento, devido às centenas de comparações efetuadas por cada caractér do input. No entanto, o resultado final é exibido quase imediatamente após o utilizador carregar na tecla 'Enter'.

Vamos a exemplos concretos:

Para as palavras 'universidade', 'castelo', 'padaria', 'deslumbramento' e 'ponta de sorte', a comparação final é de 100%. Aqui, o resultado final dado pelo programa é exatamente idêntico à solução disponibilizada pelo professor.

```
[ricardo@neves] $ ~/Universidade/PLC/SPLN/ConversorFonetico
>> python3 conversor.py 'castelo'

Processing text...

Number of words: 1
Plain text: castelo
Stressed syllables: cas·'te·lo
IPA: kɐʃ.'tɛ.lu
Solution: kɐʃ.t'ɛ.lu
Comparison: 9/9, 100.00%
```

```
[ricardo@neves] $ ~/Universidade/PLC/SPLN/ConversorFonetico
>> python3 conversor.py 'padaria'

Processing text...

Number of words: 1
Plain text: padaria
Stressed syllables: pa·da·'ri·a
IPA: pɐ.dɐ.'ɾi.ɐ
Solution: pɐ.dɐ.ɾ'i.ɐ
Comparison: 10/10, 100.00%
```

```
[ricardo@neves] $ ~/Universidade/PLC/SPLN/ConversorFonetico
>> python3 conversor.py 'deslumbramento'

Processing text...

Number of words: 1
Plain text: deslumbramento
Stressed syllables: des·lum·bra·'men·to
IPA: dɨʒ.lũ.bɾɐ.'mẽ.tu
Solution: dɨʒ.lũ.bɾɐ.m'ẽ.tu
Comparison: 16/16, 100.00%
```

```
[ricardo@neves] $ ~/Universidade/PLC/SPLN/ConversorFonetico
>> python3 conversor.py 'ponta de sorte'

Processing text...

Number of words: 3
Plain text: ponta de sorte
Stressed syllables: 'pon·ta de 'sor·te
IPA: 'põ.tɐ dɨ 'sɔɾ.tɨ
Solution: p'õ.tɐ dɨ s'ɔɾ.tɨ
Comparison: 15/15, 100.00%
```

O sistema também reconhece, divide e traduz nomes próprios. Aqui, não existe comparação possível, pelo que a avaliação final é de 0%.

```
[ricardo@neves] $ ~/Universidade/PLC/SPLN/ConversorFonetico
>> python3 conversor.py 'Ricardo Neves'

Processing text...

Number of words: 2
Plain text: Ricardo Neves
Stressed syllables: Ri·'car·do 'Ne·ves
IPA: ʁi.'kaɾ.du 'nɛ.vɨʃ
Solution:
Comparison: 0/1, 0.00%
```

Por fim, os exemplos seguintes mostram uma diferença de um caractér entre o resultado final e a solução disponível. No entanto, utilizado o conversor do Portal 'Learn European Portuguese', o resultado do programa é idêntico ao resultado do portal. Isto leva-me a concluir que um do Portal ou o ficheiro das soluções encontra-se errado (de referir que são raras as ocasiões em que esta descrepância se verifica).

```
[ricardo@neves] $ ~/Universidade/PLC/SPLN/ConversorFonetico
>> python3 conversor.py 'lixo ecológico'

Processing text...

Number of words: 2
Plain text: lixo ecológico
Stressed syllables: 'li·xo e·co·ló·gi·co
IPA: 'li.ʃu i.ku.lɔ.ʒi.ku
Solution: l'i.ʃu ɛ.ku.l'ɔ.ʒi.ku
Comparison: 18/19, 94.74%
```

```
[ricardo@neves] $ ~/Universidade/PLC/SPLN/ConversorFonetico
>> python3 conversor.py 'rapaz formiga que sabe tocar guitarra'

Processing text...

Number of words: 6
Plain text: rapaz formiga que sabe tocar guitarra
Stressed syllables: ra·'paz for·'mi·ga que 'sa·be to·'car gui·'tar·ra
IPA: ʁɐ.'paʃ fuɾ.'mi.gɐ kɨ 'sa.bɨ tu.'kaɾ gi.'ta.ʁɐ
Solution: ʁɐ.p'aʃ foɾ.m'i.gɐ kɨ s'a.bɨ tu.k'aɾ gi.t'a.ʁɐ
Comparison: 40/41, 97.56%
```

4 Publicação do módulo no TestPypi

Finalmente, este módulo foi publicado na biblioteca TestPypi, com o nome de 'conversorfonetico'. Pode ser acedido **aqui**. Através de um terminal, é possível fazer download e executar o módulo; no entanto, é exibido um erro que refere que o ficheiro das soluções 'ipa.json', disponibilizado pelo professor, não se encontra disponível. Sendo assim, não foi possível testar o módulo.