Documentação relativa à tokenização e à sentenciação do corpus Petrolês[1]

Versão 1.1 - 07/08/2020

Aline Silveira

Elvis de Souza

Tatiana Cavalcanti

Wograine Evelyn

Cláudia Freitas

1. Dentro da frase:

- **hífen**: palavras separadas por hífen sempre contam como um único token, ou seja, ele não é critério de separação de palavras (Ver Critérios de separação de frases), e o espaço que houver entre o hífen e um dos termos da palavra composta deverá ser eliminado.
  Exemplo: "**auto- sustentável**" é um token só, e vira "**auto-sustentável"**

- **travessão**: se houver espaço entre as palavras e o travessão, separá-los como tokens distintos. Se não houver espaço, tudo forma um único token.
  Exemplo:
  Faixa de temperatura: 160oC – 220oC → tokens: 160oC; –; 220oC
  Faixa de temperatura: 80oC-120oC → token: 80oC-120oC

- **barra**: palavras separadas por barra são um token só
  Exemplo: "propriedades ***físico/químicas***"; "fator ***homem/hora***"; "**km/h**"; "**S/cm**"

[1] Documentação produzida no âmbito do projeto Petrolês.

→ **OBS.: UNIDADES e SÍMBOLOS** (60 km/h; 10-5 S/cm; 5V; 200°C; 10ml; 50%):

a) Se *houver* espaço entre o número e a unidade, temos *dois* tokens distintos.
   Exemplo: "condutividade de 10-5 S/cm e janela de estabilidade eletroquímica maior que 5 V." → tokens: **10-5** e **S/cm**
b) Se *não houver* espaço (100oC), tudo se configura como *um* token.
   Exemplo: perdem água a temperaturas próximas a 100oC" → token: **100oC**

- **parênteses**: o parênteses é um token por si só, como outros sinais de pontuação (ponto final, dois pontos, vírgula etc.), e portanto não deve estar atrelado a nenhuma palavra. Uma *exceção* a essa regra se encontra em compostos químicos como "Indeno(1,2,3-cd)pireno", "Benzo(g.h.i)perileno" ou "Dibenzo(a,h)antraceno", cada um deles contando como um token.

  Exemplo: O **poli(bisfenol A-co-epicloridrina)** (PBE) é uma resina epóxi contendo grupos éter que podem coordenar cátions (Figura 4).

poli	nsubj(resina)
(	flat:name(poli)
bisfenol	flat:name(poli)
A-co-epicloridrina	flat:name(poli)
)	flat:name(poli)

- ***et al.***: separado em dois tokens, "**et**" e "**al.**", que terão como pos NOUN.

→ **OBS.2:** Números referentes às notas de rodapé, assim como equações, não devem estar no txt.

2. Delimitação de frases; sentenciação

- **Títulos e subtítulos**: tirar a numeração e colocar ponto final, separando-os como uma sentença (Ver SEPARADORES DE SENTENÇA – "." marca fim de frase).

- **PDF x TXT:**

3-Introdução:

3.1-Caracterização Geral:

Desde 1887, quando se teve o início da "era da propulsão mecânica" e posteriormente com o surgimento da indústria petroquímica em 1930, o petróleo tem tido importante função na sociedade, como fonte combustível e fornecendo matéria sintética para diversos produtos (CETESB, 2002).

```
# sent_id = 6-20140908-MONOGRAFIA_0-1
# text = Introdução.
Introdução      _       _       _       _
.       _       _       _       _       _

# sent_id = 6-20140908-MONOGRAFIA_0-2
# text = Caracterização Geral.
Caracterização  _       _       _       _
Geral   _       _       _       _       _
.       _       _       _       _       _
```

- **Listas itemizadas:**

Número + ponto final OU qualquer outro marcador (bolinha, tracinho) são eliminados.

- Quando separados por ***ponto e vírgula (ou vírgula)***: os itens formam uma única sentença (exemplos 1, 2 e 3).

- Quando ***não há pontuação*** sucedendo o item: adição de ponto e vírgula, colocando ponto final apenas no último item da lista (exemplo 3).

- Quando separados por ***ponto final***: cada item forma uma sentença própria (ponto final é delimitador de sentença SEMPRE).

Exemplo 1:

De acordo com CETESB (2002) as características mais relevantes em um derrame são:

1. Tipo e quantidade de petróleo, sendo os mais tóxicos os óleos leves devido á presença de uma quantidade maior de compostos aromáticos;
2. Amplitude de maré, podendo esta agravar o efeito do derrame ou mesmo contribuir para processo de limpeza;
3. Época do ano, por causar consideráveis variações na estrutura e composição das comunidades biológicas costeiras;

Como proceder: O que vem após número + ponto final é deslocado para logo após o dois pontos ou ponto e vírgula, seguido de espaço:

De acordo com CETESB (2002) as características mais relevantes em um derrame são: XXXX

(ponto final é separador de sentença, mas ; e : **não** são – para mais informações, ver "Critérios de separação de frases")

Resultado:

De acordo com CETESB (2002) as características mais relevantes em um derrame são: Tipo e quantidade de petróleo, sendo os mais tóxicos os óleos leves devido á presença de uma quantidade maior de compostos aromáticos; Amplitude de maré, podendo esta agravar o efeito do derrame ou mesmo contribuir para processo de limpeza; Época do ano, por causar consideráveis variações na estrutura e composição das comunidades biológicas costeiras;

Exemplo 2:

PEMFC (proton exchange membran fuel cell – célula a combustível de membrana de troca de próton):

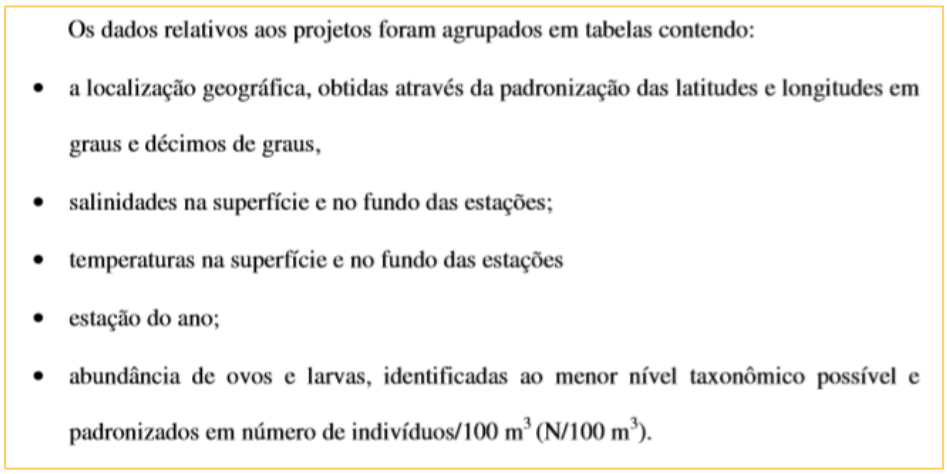
Eletrólito: membrana polimérica de condução protônica;
Faixa de temperatura: 80^{o}C-120^{o}C;
Vantagens: alta densidade de potência, operação flexível, mobilidade;
Desvantagens: custo da membrana e catalisador, contaminação do catalisador com monóxido de carbono;
Aplicações: veículos automotores, espaçonaves, unidades estacionárias.

Como proceder: PEMFC é subtítulo, portanto colocamos ponto final.

#text = PEMFC (proton exchange membran fuel cell – célula a combustível de membrana de troca de próton).

#text = Eletrólito: membrana polimérica de condução protônica; Faixa de temperatura: 80oC-120oC; Vantagens: alta densidade de potência, operação flexível, mobilidade; Desvantagens: custo da membrana e catalisador, contaminação do catalisador com monóxido de carbono; Aplicações: veículos automotores, espaçonaves, unidades estacionárias.

Exemplo 3:



Os dados relativos aos projetos foram agrupados em tabelas contendo:

- a localização geográfica, obtidas através da padronização das latitudes e longitudes em graus e décimos de graus,
- salinidades na superfície e no fundo das estações;
- temperaturas na superfície e no fundo das estações
- estação do ano;
- abundância de ovos e larvas, identificadas ao menor nível taxonômico possível e padronizados em número de indivíduos/100 m^3 (N/100 m^3).

text = Os dados relativos aos projetos foram agrupados em tabelas contendo: a localização geográfica, obtidas através da padronização das latitudes e longitudes em graus e décimos de graus, salinidades na superfície e no fundo das estações; temperaturas na superfície e no fundo das estações; estação do ano; abundância de ovos e larvas, identificadas ao menor nível taxonômico possível e padronizados em número de indivíduos/100 m3 (N/100 m3).

Critérios de separação de frases

1. SEPARADORES DE SENTENÇA

a. ponto final

text = Nesta reação de adição, um mol de ligações duplas conjugadas sempre consumirá dois mols de iodo.
sent_id = 0-20150121-TESEMSC_0-11

OBS.: O ponto final se diferencia do ponto marcador de abreviações, como aquele encontrado na expressão "et al." **–** tokenizada como "et" e "al."

2. NÃO SEPARADORES DE SENTENÇA

a. vírgula (,)

text = Estes compostos diminuem a qualidade dos produtos petrolíferos devido à sua fácil polimerização, já que as suas ligações duplas conjugadas apresentam alta reatividade.

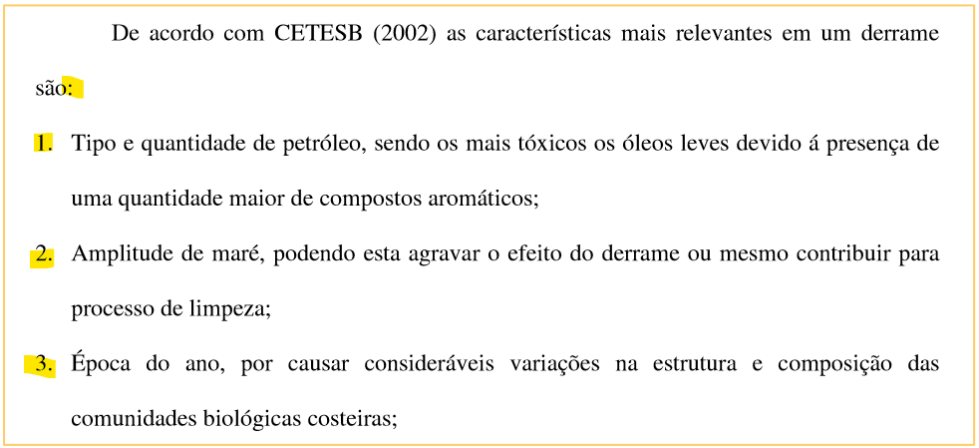
sent_id = 0-20150121-TESEMSC_0-3

b. ponto e vírgula (;)

text = Adicionalmente, tem limitações relacionadas com a concentração e a natureza do fenol (Smith, 1987; Spiker, 1992; Aitken, 1993 e Wada, 1994).

sent_id = 2-20150126-TESEDSC_0-9

OBS.: Os itens de uma lista itemizada, quando terminados em ponto e vírgula, formam uma única sentença (ver tópico "Listas itemizadas" mais acima).



De acordo com CETESB (2002) as características mais relevantes em um derrame são:

1. Tipo e quantidade de petróleo, sendo os mais tóxicos os óleos leves devido á presença de uma quantidade maior de compostos aromáticos;
2. Amplitude de maré, podendo esta agravar o efeito do derrame ou mesmo contribuir para processo de limpeza;
3. Época do ano, por causar consideráveis variações na estrutura e composição das comunidades biológicas costeiras;

No TXT:

text = De acordo com CETESB (2002) as características mais relevantes em um derrame são: Tipo e quantidade de petróleo, sendo os mais tóxicos os óleos leves devido á presença de uma quantidade maior de compostos aromáticos; Amplitude de maré, podendo esta agravar o efeito do derrame ou mesmo contribuir para processo de limpeza; Época do ano, por causar consideráveis variações na estrutura e composição das comunidades biológicas costeiras; Grau de hidrodinamismo, determinado pela quantidade, intensidade e força das ondas e correntes locais; Ciclo construtivo/destrutivo do ambiente: determinado pelo grau de erosão e deposição das praias; Tipo de substrato; Tipo de comunidade; Exposição prévia a outros impactos; Formas de limpeza aplicadas ao derrame.

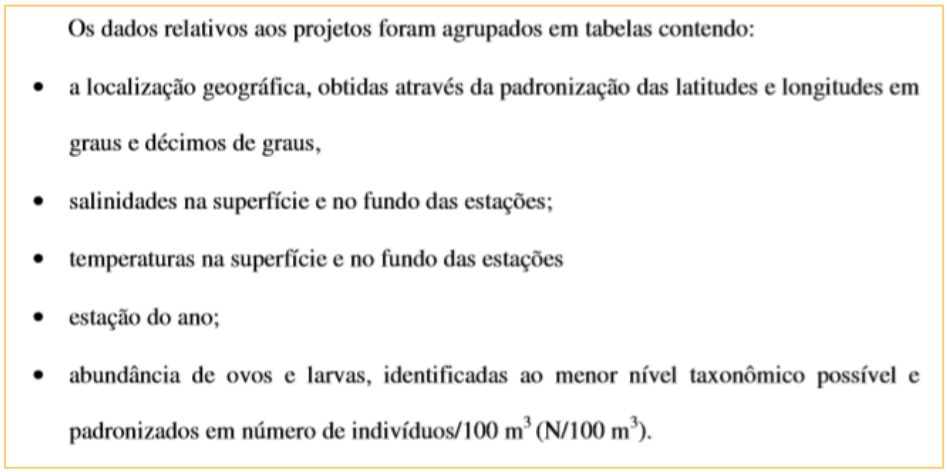
sent_id = 6-20140908-MONOGRAFIA_0-45

c. dois pontos (:)

text = Especificamente em organismos planctônicos, a contaminação pode vir de diferentes formas: através da fração solúvel, do contato direto com a mancha ou mesmo pela ingestão de alimentos contaminados por petróleo.
sent_id = 6-20140908-MONOGRAFIA_0-50

OBS.: Mesmo quando precede uma lista itemizada separada do texto, dois pontos não separam sentenças (ver tópico de "Listas itemizadas" mais acima).

No PDF:



Os dados relativos aos projetos foram agrupados em tabelas contendo:

- a localização geográfica, obtidas através da padronização das latitudes e longitudes em graus e décimos de graus,
- salinidades na superfície e no fundo das estações;
- temperaturas na superfície e no fundo das estações
- estação do ano;
- abundância de ovos e larvas, identificadas ao menor nível taxonômico possível e padronizados em número de indivíduos/100 m^3 (N/100 m^3).

No TXT:
text = Os dados relativos aos projetos foram agrupados em tabelas contendo: a localização geográfica, obtidas através da padronização das latitudes e longitudes em graus e décimos de graus, salinidades na superfície e no fundo das estações; temperaturas na superfície e no fundo das estações; estação do ano; abundância de ovos e larvas, identificadas ao menor nível taxonômico possível e padronizados em número de indivíduos/100 m3 (N/100 m3).

d. travessão (–)

#text = Neste trabalho estudou-se a degradação oxidativa usando enzima cloroperoxidase (CPO) de Caldariomyces fumago dos compostos fenólicos [...] e de compostos fenólicos presentes nas águas residuais de refinaria em efluente bruto – água acida (EB) e água de fundos de tanque de armazenamento de óleo cru (TA).

#sent_id: 2-20150126-TESEDSC_0_resumo-1

e. parênteses ()

text = Além dos dienos conjugados, o estireno e seus derivados (devido à conjugação da ligação dupla com o sistema aromático) também apresentam forte tendência à polimerização (POLÁK et al., 1986).

sent_id = 0-20150121-TESEMSC_0-4