Capitulo_4

Flavio Codeço Coelho
(Dated: November 25, 2019)

CONTENTS

I. APLICANDO MODELAGEM DE ASSUNTOS AO DHBB

Neste capítulo vamos explorar ferramentas de modelagem de assuntos e explorar aplicações ao DHBB. Como sempre, começamos com alguns imports familiares.

Vamos também carregar o modelo de NLP para a língua portuguesa do Spacy:

Agora faremos alguns imports novos, particularmente da biblioteca Gensim, que nos oferece as ferramentas que necessitamos para modelagem de assuntos.

Para minimizar o uso de memória, vamos construir uma classe para representar o nosso corpus como um iterador, operando diretamente do banco de dados. Desta forma, ao fazer nossas análises, podemos carregar um documento por vez para alimentar os modelos, sem a necessidade de manter todo o corpus na memória, economizando memória RAM.

Abaixo um pequeno exemplo de como a classe `DHBBCorpus` funciona:

```
['ńJosé', 'Machado', 'Coelho', 'de', 'Castroż', 'nasceu', 'em', 'Lorena', 'SP']
```

I.1. Word2vec

Vamos começar pelo treinamento de um modelo word2vec. Este modelo itera 6 vezez sobre o corpus logo, devemos ver o contador atingir 46122. Estas repetições são necessárias para permitir a

I.1.1. Explorando o modelo

```
ńJosé
Machado
Coelho
de
Castroż
nasceu
em
Lorena
SP
Estudou
```

```
array([ 1.213255  ,  0.8772738 , -0.06480797,  3.2624714 ,  2.7141786 ,
       -1.8534654 ,  2.2568564 ,  2.9633396 ,  2.872959  ,  0.17697811,
        0.12649159,  2.1954768 ,  0.69300246,  0.7808246 , -1.8608911 ,
```

```
       -1.9773395 , -6.7470145 ,  1.3309088 , -1.8495115 ,  1.4977074 ,
        4.106705  ,  0.18757463,  0.976982  ,  0.50778925, -1.3095387 ,
       -0.46490836,  1.3230817 , -1.0534579 , -3.3441863 , -4.0759525 ,
        0.20254904,  2.8288093 ,  0.0304382 ,  0.79879284, -0.27596825,
        1.0034108 ,  2.1574116 , -0.84378964, -2.6574318 ,  0.725499  ,
       -2.5952895 ,  0.6314094 , -2.789002  ,  5.0512304 , -1.6781955 ,
       -3.2558453 , -2.3100727 , -2.2926745 , -4.6876025 , -3.1457222 ,
        0.19902074, -0.6323914 ,  1.0432411 , -1.1918347 ,  4.250325  ,
       -2.759753  ,  2.7777524 ,  2.8528945 , -0.3918448 , -1.9812953 ,
        4.549228  , -6.3611536 ,  0.09752683,  0.37324116,  0.77719754,
       -3.6438515 , -3.4890673 , -2.3472724 ,  0.6998824 ,  3.6128554 ,
       -2.5014155 , -1.9615922 , -1.4071759 ,  0.5612391 ,  1.0864763 ,
        4.145173  , -0.7451139 , -5.0320497 , -1.4627253 , -5.186103  ,
        1.4427031 ,  5.217514  , -2.8483255 ,  3.8392422 , -4.1885376 ,
        3.3657196 , -3.235564  , -3.951503  , -1.6820241 ,  3.6057582 ,
        5.227617  ,  1.9262496 , -1.3290535 ,  1.9148936 ,  5.4682064 ,
        2.8205392 , -0.93365663, -0.58919495,  1.3472942 ,  1.4066038 ],
      dtype=float32)
```

```
38762
```

```
[('cidadão', 0.7166314125061035),
 ('dever', 0.7115763425827026),
 ('universo', 0.708888053894043),
 ('pensamento', 0.7076699733734131),
 ('nacionalismo', 0.7048879861831665),
 ('senso', 0.7020969390869141),
 ('inimigo', 0.6951218843460083),
 ('espírito', 0.691114068031311),
 ('sentimento', 0.6901904940605164),
 ('desafio', 0.6858540773391724),
 ('fenômeno', 0.6775267720222473),
 ('leitor', 0.6737638115882874),
 ('retrato', 0.6691586971282959),
 ('veículo', 0.6683763265609741),
 ('trabalhador', 0.6611828804016113),
 ('erro', 0.6600130200386047),
 ('povo', 0.6572052836418152),
 ('negócio', 0.6528052091598511),
 ('homem', 0.650109052658081),
 ('liberalismo', 0.6475819945335388)]
```

```
8713256
```

I.1.2. Visualizando os vetores de palavras

```
[('pragmática', 0.6115255355834961),
 ('política', 0.6081854701042175),
 ('Resultante', 0.5997891426086426),
 ('político-institucional', 0.5906023979187012),
 ('político-social', 0.5800623893737793),
 ('inevitavelmente', 0.5741809010505676),
 ('nítida', 0.5726122856140137),
```

```
 ('persistente', 0.5723466873168945),
 ('induzida', 0.5718863010406494),
 ('vulnerabilidade', 0.5662615895271301)]

array([-1.1804812 ,  0.17904007,  2.7701726 ,  2.2788234 , -2.3939579 ,
       -1.6261489 ,  3.037264  , -1.5370936 ,  0.1686287 ,  1.7679174 ,
       -0.6206008 , -1.4379971 ,  0.9625195 ,  5.3786445 , -1.5540149 ,
       -0.00777411, -1.401691  , -4.2874527 , -0.82898766,  1.8527349 ,
        1.1073897 ,  0.4124229 , -1.7202759 , -1.9285159 ,  0.22249055,
       -0.1637733 ,  0.3505941 , -1.6477919 ,  3.5996528 , -1.884695  ,
        0.70941716,  2.4855223 ,  0.19784053,  1.0554858 , -2.0812526 ,
        0.25975254,  1.0439771 ,  0.40970242,  0.78404856, -2.2612848 ,
        1.7169695 ,  0.73526394,  1.4105549 ,  4.066178  , -0.40555423,
        1.8521545 ,  3.7054353 ,  1.3822607 , -4.056506  ,  1.9932911 ,
        0.89136565, -2.6281335 ,  3.6849546 ,  3.6759586 , -0.3893448 ,
       -0.93963474,  3.059989  ,  0.07065499, -0.69516784, -1.2888556 ,
        0.92977023, -1.6649121 ,  2.6862006 , -2.4784298 ,  1.8600826 ,
        0.779212  ,  2.1920981 , -1.9809456 , -2.688896  ,  0.60114765,
       -2.3174565 ,  0.08518887,  0.43477178,  1.2217847 ,  0.9667772 ,
        1.8949485 ,  1.2996533 ,  2.6153111 , -0.04217112,  0.72740364,
       -0.7117096 , -0.73151827,  0.16029513,  0.06895769, -0.17937306,
        0.48638743, -0.52041715, -2.0970736 , -1.1521842 , -3.9579785 ,
        0.51819354,  0.12455559,  2.2292786 ,  0.33785737, -0.4980077 ,
       -2.9939485 , -2.5539153 , -1.4805562 ,  3.1075735 ,  0.99266756],
      dtype=float32)
```