# Palavras pulverizadas

Diana Santos

31 de janeiro de 2022

Neste texto descrevemos algumas experiências de avaliação de palavras pulverizadas (a nossa tradução de "word embeddings"), assim como uma comparação entre as que existem disponíveis para o português. É trabalho conjunto com Alberto Simões, Cristina Mota e Hugo Gonçalo Oliveira.

## 1 Criação

Comandos para criação das palavras pulverizadas:

Para word2vec

```
./word2vec -train literatecabase -output literatecabase.bin -cbow 1
-size 300 -window 8 -negative 25 -hs 0 -sample 1e-4 -threads 20 -binary
1 -iter 15
```

para glove

```
build/vocab_count -min-count 5 -verbose 2 < literatecabase >
literatecabase_vocab
build/cooccur -memory 4.0 -vocab-file literatecabase_vocab -verbose 2
-window-size 15 < literatecabase > literatecabase-cooccur.bin
build/shuffle -memory 4.0 -verbose 2 < literatecabase-cooccur.bin
> literatecabase_coocur.shuf.bin
build/glove -save-file literatecabase_vectors -threads 8 -input-file
literatecabase_coocur.shuf.bin -x-max 10 -iter 15 -vector-size 300
-binary 2 -vocab-file literatecabase_vocab -verbose 2
```

para fasttext

```
./fasttext cbow -input todossimples -output todossimples.bin -dim 300
```

para fasttext word2phrase

```
./word2phrase -train ~/Documents/palpulv/fastText/todosbase
-output todosbase-phrase0 -threshold 200 -debug 2
./word2phrase -train todosbase-phrase0 -output todosbase-phrase1
-threshold 100 -debug 2
./word2vec -train todosbase-phrase1 -output todosbase-phrase.bin
-cbow 1 -size 300 -window 10 -negative 25 -hs 0 -sample 1e-5
-threads 20 -binary 0 -iter 15
```

# 2 Pré-processamento

Por uma questão de facilidade de comparação, seguimos algumas das propostas do NILC, mas experimentámos também outras. Usámos os corpos do AC/DC (junho de 2021), que estão analisados sintacticamente, por isso pudemos dispor de várias informações sobre as palavras que outros projetos eventualmente não teriam.

Em todo os casos passámos todos os números para a sequência (arbitrária) "123", e mantivemos $< s >$ e $< /s >$ como separadores/indicadores de princípio e fim de frase.

**base** todas as palavras passadas para minúsculas

**simples** só palavras

**maiusculas** todas as palavras passadas para minúsculas exceto os nomes próprios

**mwe** palavras com mwe

**lemas** só lemas

**lemasmwe** lemas e mwe

**pallemas** palavra:lema

**pallemasmwe** palavra:lema com mwe

Criámos esta forma de corpos de entrada para o corpo TODOS, e para o corpo LITERATECA (apenas literatura).

# 3 Outros projetos

Muito brevemente, identificamos facilmente os seguintes repositórios, com base no blogue do David Batista [**?** ], e medimos o seu tamanho em número de palavras, e em percentagem de sobreposição. Também fizemos a mesma contagem não incluindo números, ou melhor, não incluindo "palavras" que contenham um ou mais algarismos.

| | Tamanho | sem algarismos |
|---|---|---|
| nilc | 929.606 | 910.215 |
| nlx | 873.910 | 752.001 |
| pt-lkb | 202.001 | 201.877 |
| cc | 2.000.000 | 1.665.247 |
| base | 1.052.405 | 984.226 |
| simples | 1.461.935 | 1.236.594 |
| maiusculas | 1.581.664 | 1.509.466 |
| mwe | 1.282.389 | 1.213.117 |
| lemas | 1.613.937 | 1.374.196 |
| lemasmwe | 1.626.086 | 1.385.548 |
| base-phrase | 1.467.915 | 1.427.370 |

Table 1: Número de palavras partilhadas entre as diversas pulverizações, em valores absolutos, contando todas as palavras

| | nilc | nlx | pt-lkb | cc | todosbase | todoslemas |
|---|---|---|---|---|---|---|
| nilc | - | 296.310 | 75.286 | 380.585 | **537.121** | 159.044 |
| nlx | 296310 | - | 58.726 | **637.484** | 233.795 | 337.107 |
| pt-lkb | 75.286 | 58.726 | - | 70.231 | **75.311** | 65.900 |
| cc | 380.585 | **637.484** | 70.231 | - | 369.899 | 456.284 |
| base | **537.121** | 233.795 | 75.311 | 369.899 | - | 314.314 |
| lemas | 159.044 | 337.107 | 65.900 | **456.284** | 314.314 | - |

# 4 Avaliação comparativa

O Alberto Simões desenvolveu um programa que testa o desempenho de palavras pulverizadas com base num conjunto de analogias, o analogiador.

A forma de o invocar é:

```
./analogiador.py FastText/todossimples300.bin Analogias/LX-4WAnalogies_v2.txt
 --tipo fasttext --binario
```

Table 2: A diferença entre as pulverizações com valores relativos: qual a percentagem de palavras também na outra pulverização, contando todas as palavras.

| | nilc | nlx | pt-lkb | cc | todosbase | todoslemas |
|---|---|---|---|---|---|---|
| nilc | - | 31,87% | 8,10% | 40,94% | 57,78% | 17,11% |
| nlx | 33,91% | - | 6,72% | 72,95% | 26,75% | 38,57% |
| pt-lkb | 37,27% | 29,07% | - | 34,76 | 37,28% | 32,62% |
| cc | 19,03% | 31,87% | 3,51% | - | 18,49% | 22,81% |
| base | 51,04& | 22,22% | 7,16% | 35,14% | - | 29,87% |
| lemas | 9,85% | 20,89% | 4,08% | 28,27% | 19,47% | - |

Table 3: Número de palavras partilhadas entre as diversas pulverizações, em valores absolutos, contando apenas as palavras sem algarismos

| | nilc | nlx | pt-lkb | cc | todosbase | todoslemas |
|---|---|---|---|---|---|---|
| nilc | - | 296.157 | 75.285 | 380.252 | **536.720** | 158.813 |
| nlx | 296.157 | - | 58.716 | **596.091** | 231.931 | 304.249 |
| pt-lkb | 75.286 | 58.726 | - | 70.217 | **75.311** | 65.900 |
| cc | 380.252 | **596.091** | 70.217 | - | 365.048 | 456.284 |
| base | **536.720** | 233.795 | 75.301 | 281.097 | - | 314.314 |
| lemas | 158.813 | 304.249 | 65.900 | **390.415** | 281.097 | - |

## 4.1 Analogias

Avaliação com as analogias originais NLX, e com as analogias cortadas, usando o analogiador. A percentagem é das palavras encontradas. Isto foi feito para models com dimensão 200 e 300.

As analogias cortadas são o resultado de retirar as analogias morfológicas, e alguns erros crassos, das analogias originais NLX.

Avaliação dos outros atores (todos w2v exceto cc, que e fasttext):

Avaliação usando antónimos e finalidade inversa do TALES:

Avaliação usando antónimos e finalidade inversa do TALES 1.5:

## 4.2 Distâncias

Para calcular distâncias, ou palavras mais próximas, basta fazer, no gensim,

```
model=fasttext.load_facebook_vectors("FastText/todossimples300.bin")
model=KeyedVectors.load_word2vec_format("Glove/literatecasimples_vectors.txt",
 binary=False, no_header=True)
model=KeyedVectors.load_word2vec_format("Word2vec/todossimples.bin",
```

Table 4: A diferença entre as pulverizações com valores relativos: qual a percentagem de palavras também na outra pulverização, contando apenas as palavras sem algarismos.

| | nilc | nlx | pt-lkb | cc | todosbase | todoslemas |
|---|---|---|---|---|---|---|
| nilc | - | 32,53% | 8,27% | 41,77% | 54,53% | 17,45% |
| nlx | 39,38% | - | 7,81% | 79,26% | 30,84% | 40,46% |
| pt-lkb | 37,29% | 29,08% | - | 34,78 | 37,28% | 32,62% |
| cc | 22,83% | 35,79% | 4,22% | - | 21,92% | 23,45% |
| base | 58,96& | 23,56% | 7,65% | 37,09% | - | 28,56% |
| lemas | 11,55% | 22,14% | 4,08% | 28,41% | 20,46% | - |

```
 binary=True)
result=model.most_similar("inteligência")
result=model.most_similar(positive=['inteligência','engenho'],
 negative=['imaginação'])
print (result)
```

## 4.3 Agrupamento

Table 5: Avaliação com as analogias originais NLX, e com as analogias cortadas, usando o analogiador. As duas primeiras são com 200 eixos, a terceira e quarta colunas são com 300 eixos.

| modelo | anNLX | cort | anNLX | cort |
|---|---|---|---|---|
| todosbasecbow.bin@w2v | 70.2997% | 48.8777% | 72.4486% | 50.6775% |
| todosmaiusculas.bin@w2v | 65.7051% | 59.2129% | 67.9030% | 61.5279% |
| todossimplescbow.bin@w2v | 63.7243% | 59.3391% | 66.1967% | 61.2162% |
| todossimplesskg.bin@w2v | 64.0824% | 59.3391% | 66.6184% | 62.0195% |
| todosmwecbow.bin@w2v | 65.3327% | 60.6406% | 67.5960% | 63.1483% |
| todosmweskg.bin@w2v | 64.8552% | 60.5287% | 67.5083% | 62.7313% |
| todoslemascbow.bin@w2v | 51.4904% | 61.9700% | 52.9313% | **62.4422%** |
| todoslemasskg.bin@w2v | -% | -% | 53.1703% | 62.7063% |
| todoslemasmwecbow.bin@w2v | 51.7878% | 61.0668% | 52.9234% | 62.5413% |
| todossimples-phrase.bin@w2v | 75.1329% | 53.0778% | | |
| todossimples-phraseskg.bin@w2v | -% | -% | | |
| todosbase-phrase.bin@w2v | **75.2555%** | 53.0573% | | |
| todosmaiusculas-phrase.bin@w2v | 67.8893% | 60.7598% | | |
| todosmwe-phrase.bin@w2v | 63.7060% | 60.9705% | | |
| todoslemas-phrase.bin@w2v | 60.3868% | 46.5447% | | |
| todoslemasmwe-phrase.bin@w2v | 46.5447% | 57.5557% | | |

Table 6: Avaliação dos modelos Glove com as analogias originais NLX, e com as analogias cortadas, usando o analogiador. As duas primeiras são com 50 (!) eixos, a terceira e quarta colunas são com 300 eixos.

| modelo | anNLX | cortadas | anNLX | cortadas |
|---|---|---|---|---|
| base@glove | 26.7302% | 14.3963% | **66.0989**% | 38.6760% |
| simples@glove | 22.1251%/19.2889% | 14.6213% | 57.5152% | 53.3049% |
| maiusculas@glove | 22.5484%/20.7624% | 15.1199% | 19.7472% | 20.8460% |
| mwe@glove | 26.5879%/19.9245% | 14.3467% | 30.2425% | 22.6459% |
| lemas@glove | 38.3720%/12.6086% | 15.6821% | 43.7789% | **57.5798**% |
| lemasmwe@glove | 37.7634%/12.6010% | 15.8734% | | |

Table 7: Avaliação dos modelos FastText com as analogias originais NLX, e com as analogias cortadas, usando o analogiador. As duas primeiras são com 100 eixos, a terceira e quarta colunas são com 300 eixos.

| modelo | anNLX | cortadas | anNLX | cortadas |
|---|---|---|---|---|
| todosbasecbow.bin@ft | 14.3242% | 14.6569% | 32.7453% | 16.4856% |
| todosbaseskp.bin@ft | | | 34.9104% | 16.6214% |
| todossimplescbow.bin@ft | 18.3963% | 10.6635% | 45.6040% | 30.6084% |
| todossimplesskg.bin@ft | | | 63.1479 % | **51.9555**% |
| todosmaiusculascbow.bin@ft | 36.3537% | 23.9906% | 42.5626% | 28.8340% |
| todosmwecbow.bin@ft | 18.5109% | 10.6635% | 46.4975% | 31.4775% |
| todosmweskg.bin@ft | | | **63.2969**% | 51.8830% |
| todoslemascbow.bin@ft | 38.6563% | 41.5535% | 35.3629% | 31.1063% |
| todoslemasskg.bin@ft | | | 43.9430% | 45.6907% |
| todoslemasmwecbow.bin@ft | 32.2241% | 27.9920% | 34.6182% | 26.1090% |
| todoslemasmweskg.bin@ft | 34.2497% | 33.2307% | 10.3721% | |

Table 8: Avaliação dos modelos dos outros autores: w2v: word2vec, gl: glove, ft: fasttext

| modelo | anNLX | cortadas | novas |
|---|---|---|---|
| NILC-300-cbow w2v | 70.3678% | 49.3421% | 44.9541% (49 em 109) |
| NILC-300 gl | 66.194% | 41.8118% | 63.208 (69 em 109) |
| NILC-300-cbow ft | 76.427% | 69.067% | 24.7706% (27 em 109) |
| NLX w2v | 49.6869% | 43.9764% | 39.6825% (50 em 126) |
| PT-LKB-64 w2v | 26.4438% | 28.5714% | 3.1250% (3 em 96) |
| PT-LKB-128 w2v | 26.2538% | 27.6732% | 5.2083% (5 em 96) |
| CC-300-cbow ft | 79.5406% | 73.7009% | 61.4754% (75 em 122) |
| Sense2vec | | | |

Table 9: Avaliação usando antónimos e finalidade inversa do TALES

| modelo | anton | | fin inv | |
|---|---|---|---|---|
| nilc@w2v | 30.8085% | 362 em 1175 | 11.9281% | 146 em 1224 |
| nilc@glove | 17.5654% | 215 em 1224 | 29.1915% | 343 em 1175 |
| nlx | 24.5106% | 288 em 1175 | | |
| pt-lkb-128 | 35.7843% | 438 em 1224 | 30.7190% | 376 em 1224 |
| lemascbow.bin@fasttext | 27.7778% | 340 em 1224 | 14.1277% | 166 em 1175 |
| mwe@glove | 9.1503% | 112 em 1224 | 12.8511% | 151 em 1175 |
| simples@glove | 10.0490% | 123 em 1224 | 14.2128% | 167 em 1175 |
| maiusculas@glove | 10.2941% | 126 em 1224 | 14.0426% | 165 em 1175 |
| lemas@glove | 11.6830% | 143 em 1224 | 13.2766% | 156 em 1175 |
| lemasmwe@glove | 11.5196% | 141 em 1224 | 13.2766% | 156 em 1175 |
| base.bin@w2v | 32.5957% | 383 em 1175 | 23.8562% | 292 em 1224 |
| simplescbow@w2v | 33.7872% | 397 em 1175 | 23.3660% | 286 em 1224 |
| maiusculas@w2v | 33.3617% | 392 em 1175 | 25.0817% | 307 em 1224 |
| mwecbow@w2v | 34.9787% | 411 em 1175 | 23.5294% | 288 em 1224 |
| mweskg@w2v | 34.9787% | 411 em 1175 | 23.2026% | 284 em 1224 |
| lemascbow@w2v | | | 24.6732% | 302 em 1224 |
| lemasmwecbow@w2v | | | 26.3072% | 322 em 1224 |
| lemasmweskg@w2v | | | 26.2255% | 321 em 1224 |
| simplescbow@w2v | | | 23.2026% | 284 em 1224 |
| simplesskg@w2v | | | 22.6307% | 277 em 1224 |
| base-phrase.bin@w2v | 30.1277% | 354 em 1175 | 24.3464% | 298 em 1224 |
| simples-phrase.bin@w2v | 29.1064% | 342 em 1175 | | |
| mwe-phrase.bin@w2v | 26.1277% | 307 em 1175 | 22.7124% | 278 em 1224 |
| maiusculas-phrase.bin@w2v | | | 23.2026% | 284 em 1224 |
| lemas-phrase.bin@w2v | | | 25.0000% | 306 em 1224 |
| lemasskg-phrase.bin@w2v | | | 24.5098% | 300 em 1224 |
| lemasmwe-phrase.bin@w2v | | | 19.5261% | 239 em 1224 |

Table 10: Avaliação usando antónimos e finalidade inversa do TALES 1.5

| modelo | antónimos | em quantas | finalidade inversa | em quantas |
|---|---|---|---|---|
| nilccbow | 31.1275% | em 1224 | 28.7582% | em 1224 |
| nlx | 24.5098% | em 1224 | 9.5588% | em 1224 |
| basecbow.phrase@w2v | 29.9837% | em 1224 | 27.9412% | em 1224 |
| simplescbow@w2v | 34.8039% | em 1224 | 28.9216% | em 1224 |
| lemascbow@w2v | 36.1928% | em 1224 | 13.2353% | em 1224 |
| lemasmwecbow@w2v | 24.6732% | em 1224 | | |