**# of raw words: collected 501567-word types from a corpus of 47919336 raw words and 512629 sentences**

**sentences**

**Word embedding test 1:**

Training word to vec with simple normalization.

No fix size and count, workers = 4

effective\_min\_count=5 retains 123907 unique words (24% of original 501567, drops 377660)

training on 47919336 raw words (45658252 effective words) took 317.6s, 143754 effective

words

train\_time 764.5972957611084

model name: word\_2\_vec\_normalized

size on disk: 138 MB

124321 entries, 123907 vectors

train for 5 epochs

**Word Embedding test 2:**

Normalized data

Workers=4, min\_count = 10

effective\_min\_count=10 retains 82356 unique words (16% of original 501567, drops 419211)

training on 47919336 raw words (45372509 effective words) took 289.3s, 156855 effective words/s

model name: word\_2\_vec\_normalized\_count\_10

size on disk: 91.1 MB

train\_time: 810.3877029418945

82770 entries, 82356 vectors

train for 5 epochs

**Word embedding test 3:**

Normalized data

Workers=4, size=200

effective\_min\_count=5 retains 123907 unique words (24% of original 501567, drops 377660)

training model with 4 workers on 123907 vocabulary and 200 features, using sg=0 hs=0 sample=0.001 negative=5 window=5

train\_time 784.940904378891

model name: word\_2\_vec\_normalized\_size\_200

size on disk: 280 MB

124321 entries, 123907 vectors

train for 5 epochs

**Word embedding test 4:**

Normalized data

Workers=4, size=200, count= 10

effective\_min\_count=10 retains 82356 unique words (16% of original 501567, drops 419211)

train\_time 773.557516336441

size on disk: 183 MB

model Name: word\_2\_vec\_normalized\_size\_200\_count\_10.txt

82770 entries, 82356 vectors

Window size = 2:

Vocab size: 123907

Time to train: 760.433456659317

Size on disk: 140 MB

Number vectors: 123907