BAG OF WORDs

Vamos a realizar una implementación sencilla del algoritmo BOW que se utiliza en procesado de lenguaje natural.

https://ongspxm.github.io/blog/2014/12/bag-of-words-natural-language-processing/ (https://ongspxm.github.io/blog/2014/12/bag-of-words-natural-language-processing/)

Lee el archivo bow.txt y elimina los signos de puntución

Elimina primero los signos de puntuacion

Crear una diccionario que cuente la aparicion de cada palabra

{'all': 14, 'forget': 2, 'vermin': 2, 'four': 4, 'mild': 2, 'sleep': 6,
'go': 8, 'better;': 2, 'seemed': 2, 'certainly': 4, 'to': 60, 'friendly':
2, 'brown': 2, 'sitting': 4, 'samsa': 4, 'accuse': 2, 'fall': 2, 'illustr
ated': 2, 'heaven': 2, 'stiff': 2, 'seven;': 2, 'try': 2, 'small': 2, 'ro
und': 2, 'thin': 2, 'past': 4, 'even': 5, 'what': 14, 'curse': 2, 'cloc
k': 4, 'deeply': 2, 'above': 2, 'ever': 1, 'told': 2, 'kicked': 2, 'neve
r': 6, 'drew': 2, 'understanding': 2, 'reported': 2, 'let': 2, 'fifteen':
2, 'change': 2, 'my': 10, 'leaves': 2, 'drawers': 2, 'boa': 2, 'makes':
2, 'thats': 4, 'from': 9, 'takes': 2, 'would': 17, 'oclock': 4, 'gregor':
7, 'next': 2, 'live': 2, 'tell': 2, 'more': 7, 'sort': 2, 'knows': 2, 'ma
d': 2, 'train': 10, 'accept': 2, 'effort': 2, 'must': 4, 'me': 4, 'room':
4, 'heard': 2, 'this': 5, 'getting': 2, 'arches': 2, 'can': 4, 'believe
d': 2, 'making': 2, 'drops': 2, 'didnt': 4, 'claim': 2, 'slid': 2, 'sales
men': 2, 'now': 2, 'pushed': 2, 'something': 2, 'divided': 2, 'six': 4,
'how': 2, 'assistant': 4, 'waved': 2, 'tried': 4, 'after': 1, 'spot': 2,
'collection': 4, 'wrong': 1, 'lay': 6, 'man': 2, 'a': 46, 'whenever': 2,
'maybe': 2, 'so': 7, 'dream': 2, 'began': 2, 'office': 4, 'over': 2, 'rat
tling': 2, 'soon': 2, 'years': 4, 'through': 2, 'pane': 2, 'hell': 2, 'co
ld': 2, 'still': 6, 'its': 4, 'before': 2, 'textile': 2, 'chosen': 2, 'sh
udder': 2, 'window': 2, 'covered': 4, 'travelling': 8, 'troubled': 2, 'ov
ercome': 2, 'happened': 2, 'then': 2, 'them': 2, 'food': 2, 'subordinate
s': 2, 'sleepiness': 1, 'walls': 2, 'half': 4, 'not': 14, 'one': 4, 'da
y': 4, 'headboard': 2, 'down': 2, 'always': 4, 'did': 5, 'slide': 2, 'sto
pped': 2, 'found': 4, 'went': 2, 'quarter': 2, 'heavy': 2, 'lifted': 2,
'doing': 4, 'house': 2, 'hard': 4, 'salesman': 2, 'insurance': 2, 'out':
14, 'god': 4, 'got': 6, 'shut': 2, 'housed': 2, 'given': 2, 'quite': 2,
'completely': 1, 'put': 2, 'could': 8, 'times': 2, 'thing': 2, 'place':
2, 'domed': 2, 'onto': 2, 'think': 4, 'first': 2, 'forwards': 2, 'feel':
11, 'hearing': 2, 'spots': 2, 'another': 2, 'size': 2, 'little': 8, 'gues
t': 2, 'service': 2, 'top': 2, 'slept': 2, 'anyone': 2, 'their': 2, 'to
o': 2, 'white': 2, 'hundred': 2, 'eyes': 2, 'that': 39, 'doctors': 2, 'co
py': 2, 'than': 5, 'legs': 6, 'showed': 2, 'was': 29, 'bed': 4, 'were':
5, 'and': 52, 'viewer': 2, 'suspicious': 2, 'turned': 2, 'sad': 2, 'talki
ng': 2, 'have': 23, 'anger': 2, 'saw': 2, 'any': 4, 'sat': 2, 'able': 2,
'instance': 2, 'chest': 2, 'luxury': 2, 'which': 4, 'towards': 4, 'pain':
2, 'though': 2, 'who': 4, 'extremely': 2, 'especially': 2, 'alarm': 4, 'l
ater': 2, 'cover': 2, 'helplessly': 2, 'looked': 4, 'definitely': 2, 'fit
ted': 2, 'particularly': 2, 'gilded': 2, 'spineless': 2, 'true': 2, 'stat
e': 2, 'should': 4, 'only': 2, 'touched': 2, 'do': 6, 'his': 40, 'get': 1
0, 'workshy': 1, 'familiar': 2, 'horrible': 2, 'report': 2, 'during': 2,
'him': 12, 'packed': 2, 'morning': 4, 'bad': 2, 'stupid': 2, 'rung': 4,
'worries': 2, 'where': 4, 'set': 2, 'frame': 2, 'lazy': 2, 'connections':
2, 'see': 6, 'are': 2, 'close': 2, 'arm': 2, 'best': 2, 'lots': 2, 'enoug
h': 2, 'unable': 2, 'between': 2, 'probably': 2, 'notice': 2, 'recently':
2, 'rolled': 2, 'however': 2, 'boss': 6, 'floundering': 2, 'debt': 2, 'co
me': 2, 'irregular': 2, 'career': 2, 'many': 3, 'ill': 7, 'contract': 2,

```
'whole': 2, 'breakfasts': 2, 'table': 2, 'belly;': 2, 'been': 7, 'quickl
y': 2, 'hitting': 2, 'wasnt': 2, 'much': 3, 'slowly': 2, 'parents': 6, 'h
ardly': 2, 'woke': 2, 'threw': 2, 'life': 2, 'lift': 2, 'catch': 4, 'form
er': 2, 'present': 2, 'case': 1, 'fur': 6, 'look': 4, 'these': 2, 'suppos
e': 2, 'theres': 4, 'ive': 6, 'seven': 2, 'is': 4, 'it': 34, 'sleeping':
3, 'in': 18, 'ready': 2, 'id': 6, 'if': 12, 'funny': 2, 'different': 2,
'doctor': 4, 'pay': 2, 'make': 6, 'gentlemen': 2, 'belly': 2, 'of;': 2,
'been;': 2, 'used': 2, 'itch': 4, 'rain': 2, 'moment': 2, 'moving': 2, 'l
ower': 2, 'off': 6, 'i': 16, 'well': 3, 'thought': 10, 'contact': 2, 'pos
ition': 4, 'the': 84, 'peacefully': 4, 'excessive': 1, 'just': 6, 'bein
g': 2, 'money': 2, 'armour': 2, 'hands': 2, 'rest': 2, 'human': 2, 'yes':
2, 'yet': 2, 'cut': 2, 'had': 12, 'spread': 2, 'transformed': 2, 'sample
s': 4, 'apart': 1, 'hat': 2, 'ought': 2, 'big': 2, 'possible': 2, 'earl
y': 2, 'five': 6, 'know': 6, 'desk': 4, 'bit': 2, 'lady': 2, 'dreams': 2,
'furniture': 2, 'like': 8, 'hung': 2, 'become': 2, 'because': 8, 'peopl
e': 2, 'some': 2, 'back': 12, 'recommendation': 2, 'proper': 2, 'home':
2, 'for': 7, 'bedding': 2, 'muff': 2, 'everything': 2, 'ticking': 2, 'b
e': 8, 'noise': 2, 'eating': 2, 'business': 6, 'slight': 2, 'strained':
2, 'although': 2, 'by': 4, 'on': 16, 'about': 12, 'wouldnt': 2, 'oh': 2,
'of': 38, 'gregors': 2, 'magazine': 2, 'youve': 2, 'slightly': 2, 'or':
4, 'quietly': 4, 'own': 2, 'whats': 3, 'hungrier': 1, 'into': 8, 'nonsens
e': 2, 'son': 2, 'hope;': 2, 'weather': 2, 'lively': 2, 'couldnt': 2, 'yo
ur': 6, 'boss;': 2, 'her': 2, 'there': 14, 'long': 5, 'hed': 2, 'section
s': 2, 'head': 4, 'himself': 6, 'but': 11, 'strenuous': 2, 'bosss': 4, 'u
pright': 2, 'with': 18, 'dull': 4, 'raising': 2, 'he': 79, 'rush': 2, 'ma
de': 2, 'up': 16, 'medical': 2, 'gone': 2, 'an': 2, 'as': 14, 'right': 8,
'at': 18, 'compared': 2, 'entirely': 1, 'no': 3, 'usualone': 1, 'when': 1
0, 'other': 2, 'sick': 2, 'you': 6, 'company': 2, 'nice': 2, 'picture':
2, 'felt': 5, 'ago': 4, 'pitifully': 2, 'longer': 2, 'together': 2, 'havi
ng': 2, 'fact': 1, 'time': 8, 'fresh': 2, 'avoid': 2, 'once': 4}
```

Transforma el diccionario en una lista de palabras y ordénalo de forma descendente

Imprime los 20 valores mas usados

```
[('the', 84), ('he', 79), ('to', 60), ('and', 52), ('a', 46), ('his', 4
0), ('that', 39), ('of', 38), ('it', 34), ('was', 29), ('have', 23), ('a
t', 18), ('with', 18), ('in', 18), ('would', 17), ('up', 16), ('on', 16),
('i', 16), ('as', 14), ('there', 14)]
```