In [1]:
```
import gzip
import gensim
```

```
C:\Users\Admin\Anaconda3\lib\site-packages\gensim\utils.py:1197: UserWarning: dete
cted Windows; aliasing chunkize to chunkize_serial
  warnings.warn("detected Windows; aliasing chunkize to chunkize_serial")
```

In [2]:
```
data_file="D:/reviews_data.txt.gz"

with gzip.open (data_file, 'rb') as f:
    for i,line in enumerate (f):
        print(line)
        break
```

```
b"Oct 12 2009 \tNice trendy hotel location not too bad.\tI stayed in this hotel fo
r one night. As this is a fairly new place some of the taxi drivers did not know w
here it was and/or did not want to drive there. Once I have eventually arrived at
the hotel, I was very pleasantly surprised with the decor of the lobby/ground floo
r area. It was very stylish and modern. I found the reception's staff geeting me w
ith 'Aloha' a bit out of place, but I guess they are briefed to say that to keep u
p the coroporate image.As I have a Starwood Preferred Guest member, I was given a
small gift upon-check in. It was only a couple of fridge magnets in a gift box, bu
t nevertheless a nice gesture.My room was nice and roomy, there are tea and coffee
facilities in each room and you get two complimentary bottles of water plus some t
oiletries by 'bliss'.The location is not great. It is at the last metro stop and y
ou then need to take a taxi, but if you are not planning on going to see the histo
ric sites in Beijing, then you will be ok.I chose to have some breakfast in the ho
tel, which was really tasty and there was a good selection of dishes. There are a
couple of computers to use in the communal area, as well as a pool table. There is
also a small swimming pool and a gym area.I would definitely stay in this hotel ag
ain, but only if I did not plan to travel to central Beijing, as it can take a lon
g time. The location is ok if you plan to do a lot of shopping, as there is a big
shopping centre just few minutes away from the hotel and there are plenty of eatin
g options around, including restaurants that serve a dog meat!\t\r\n"
```

In [3]:
```
def read_input(input_file):
    with gzip.open (input_file, 'rb') as f:
        for i, line in enumerate (f):
           # do some pre-processing and return a list of words for each review text
            yield gensim.utils.simple_preprocess (line)

# read the tokenized reviews into a list
# each review item becomes a serries of words
# so this becomes a list of lists
documents = list (read_input (data_file))
print(documents[0])
```

```
['oct', 'nice', 'trendy', 'hotel', 'location', 'not', 'too', 'bad', 'stayed', 'i
n', 'this', 'hotel', 'for', 'one', 'night', 'as', 'this', 'is', 'fairly', 'new',
'place', 'some', 'of', 'the', 'taxi', 'drivers', 'did', 'not', 'know', 'where', 'i
t', 'was', 'and', 'or', 'did', 'not', 'want', 'to', 'drive', 'there', 'once', 'hav
e', 'eventually', 'arrived', 'at', 'the', 'hotel', 'was', 'very', 'pleasantly', 's
urprised', 'with', 'the', 'decor', 'of', 'the', 'lobby', 'ground', 'floor', 'are
a', 'it', 'was', 'very', 'stylish', 'and', 'modern', 'found', 'the', 'reception',
'staff', 'geeting', 'me', 'with', 'aloha', 'bit', 'out', 'of', 'place', 'but', 'gu
ess', 'they', 'are', 'briefed', 'to', 'say', 'that', 'to', 'keep', 'up', 'the', 'c
oroporate', 'image', 'as', 'have', 'starwood', 'preferred', 'guest', 'member', 'wa
s', 'given', 'small', 'gift', 'upon', 'check', 'in', 'it', 'was', 'only', 'coupl
e', 'of', 'fridge', 'magnets', 'in', 'gift', 'box', 'but', 'nevertheless', 'nice',
'gesture', 'my', 'room', 'was', 'nice', 'and', 'roomy', 'there', 'are', 'tea', 'an
d', 'coffee', 'facilities', 'in', 'each', 'room', 'and', 'you', 'get', 'two', 'com
plimentary', 'bottles', 'of', 'water', 'plus', 'some', 'toiletries', 'by', 'blis
s', 'the', 'location', 'is', 'not', 'great', 'it', 'is', 'at', 'the', 'last', 'met
ro', 'stop', 'and', 'you', 'then', 'need', 'to', 'take', 'taxi', 'but', 'if', 'yo
u', 'are', 'not', 'planning', 'on', 'going', 'to', 'see', 'the', 'historic', 'site
s', 'in', 'beijing', 'then', 'you', 'will', 'be', 'ok', 'chose', 'to', 'have', 'so
me', 'breakfast', 'in', 'the', 'hotel', 'which', 'was', 'really', 'tasty', 'and',
'there', 'was', 'good', 'selection', 'of', 'dishes', 'there', 'are', 'couple', 'o
f', 'computers', 'to', 'use', 'in', 'the', 'communal', 'area', 'as', 'well', 'as',
'pool', 'table', 'there', 'is', 'also', 'small', 'swimming', 'pool', 'and', 'gym',
'area', 'would', 'definitely', 'stay', 'in', 'this', 'hotel', 'again', 'but', 'onl
y', 'if', 'did', 'not', 'plan', 'to', 'travel', 'to', 'central', 'beijing', 'as',
'it', 'can', 'take', 'long', 'time', 'the', 'location', 'is', 'ok', 'if', 'you',
'plan', 'to', 'do', 'lot', 'of', 'shopping', 'as', 'there', 'is', 'big', 'shoppin
g', 'centre', 'just', 'few', 'minutes', 'away', 'from', 'the', 'hotel', 'and', 'th
ere', 'are', 'plenty', 'of', 'eating', 'options', 'around', 'including', 'restaura
nts', 'that', 'serve', 'dog', 'meat']
```

In [4]:
```
model = gensim.models.Word2Vec (documents, size=150, window=10, min_count=2, worker
model.train(documents,total_examples=len(documents),epochs=10)
print("done")
```

```
done
```

In [7]:
```
w1 = "happy"
model.wv.most_similar (positive=w1)
```

Out[7]:
```
[('pleased', 0.8090066909790039),
 ('satisfied', 0.7422274351119995),
 ('delighted', 0.6532208919525146),
 ('impressed', 0.6428433656692505),
 ('thrilled', 0.639594316482544),
 ('disappointed', 0.5893115401268005),
 ('dissapointed', 0.5595178604125977),
 ('grateful', 0.5474934577941895),
 ('willing', 0.5382423400878906),
 ('dissappointed', 0.5308459401130676)]
```

In [6]:
```
print(model.wv.most_similar (positive=w1))
```

```
[('filthy', 0.8616929054260254), ('stained', 0.7706809043884277), ('dusty', 0.7702
877521514893), ('unclean', 0.7678651213645935), ('grubby', 0.7658600807189941),
('smelly', 0.7634941935539246), ('dingy', 0.744700014591217), ('grimy', 0.73308122
15805054), ('mouldy', 0.7158473134040833), ('gross', 0.7145285606384277)]
```

In [8]:
```
w1 = "dirty"
model.wv.most_similar (positive=w1)
```

Out[8]:
```
[('filthy', 0.8616929054260254),
 ('stained', 0.7706809043884277),
 ('dusty', 0.7702877521514893),
 ('unclean', 0.7678651213645935),
 ('grubby', 0.7658600807189941),
 ('smelly', 0.7634941935539246),
 ('dingy', 0.744700014591217),
 ('grimy', 0.7330812215805054),
 ('mouldy', 0.7158473134040833),
 ('gross', 0.7145285606384277)]
```

In [ ]: