In [1]:
```
#Create a Word2Vec Model
#Import libraries
import os
import nltk
from nltk.corpus import stopwords
```

In [2]:
```
#Download stopwords
nltk.download('stopwords')
nltk.download('punkt')
```

```
[nltk_data] Downloading package stopwords to /home/adel/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to /home/adel/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

Out[2]:
```
True
```

In [3]:
```
#Read content and sentinize
all_sentences = []

for filename in os.listdir('files/holmes'):
    with open(f'files/holmes/{filename}') as f:
        content = f.read()
        all_sentences += nltk.sent_tokenize(content.lower())
```

In [ ]:
```

```

In [4]:
```
#Tokenize each sentence
all_words = [nltk.word_tokenize(sent) for sent in all_sentences]
```

In [5]:
```
all_words[0][:10]
```

Out[5]:
```
['the', '``', 'gloria', 'scott', "''", '``', 'i', 'have', 'some', 'papers']
```

In [6]:
```
#Remove all stop words
for i in range(len(all_words)):
    all_words[i] = [w for w in all_words[i] if w not in stopwords.words('english')]
```

In [7]:
```
#Remove special characters¶
for i in range(len(all_words)):
    all_words[i] = [w for w in all_words[i] if w.isalpha()]
```

In [8]:
```
#Install gensim and python-Levenshtein
!pip install gensim
```

```
Requirement already satisfied: gensim in /home/adel/miniconda3/envs/ML_DS/lib/python3.9/site-packages (4.1.2)
Requirement already satisfied: numpy>=1.17.0 in /home/adel/miniconda3/envs/ML_DS/lib/python3.9/site-packages (fro
m gensim) (1.22.3)
Requirement already satisfied: smart-open>=1.8.1 in /home/adel/miniconda3/envs/ML_DS/lib/python3.9/site-packages
(from gensim) (5.2.1)
Requirement already satisfied: scipy>=0.18.1 in /home/adel/miniconda3/envs/ML_DS/lib/python3.9/site-packages (fro
m gensim) (1.8.0)
```

In [9]:
```
!pip install python-Levenshtein
```

```
Requirement already satisfied: python-Levenshtein in /home/adel/miniconda3/envs/ML_DS/lib/python3.9/site-packages
(0.12.2)
Requirement already satisfied: setuptools in /home/adel/miniconda3/envs/ML_DS/lib/python3.9/site-packages (from p
ython-Levenshtein) (62.1.0)
```

In [10]:
```
# Import another library
from gensim.models import Word2Vec
```

Step 9: Create a model

Step 3: Create a model

- Use **Word2Vec** on **all_words**
  - Use **min_count=2** : Ignores all words with total frequency lower than this.

In [11]:
```
#Create a model
model = Word2Vec(all_words, min_count=2)
```

In [12]:
```
#Find distances
model.wv.distance('holmes', 'watson')
```

Out[12]:
```
0.0005608201026916504
```

In [13]:
```
model.wv.distance('holmes', 'water')
```

Out[13]:
```
0.0012046098709106445
```

In [14]:
```
#Find closests words
words = model.wv.index_to_key

def closets_words(word):
    distances = {w: model.wv.distance(word, w) for w in words}
    return sorted(distances, key=lambda w: distances[w])[:15]
```

In [15]:
```
closets_words('holmes')
```

Out[15]:
```
['holmes',
 'friend',
 'hand',
 'made',
 'without',
 'eyes',
 'turned',
 'first',
 'colonel',
 'yet',
 'must',
 'quite',
 'come',
 'little',
 'words']
```