In [ ]:

```
python --version
```

In [2]:

```
import sys
sys.version
```

Out[2]:

```
'3.7.4 (default, Aug  9 2019, 18:34:13) [MSC v.1915 64 bit (AMD64)]'
```

In [3]:

```
from platform import python_version
print(python_version())
```

```
3.7.4
```

In [4]:

```
import numpy
numpy.version.version
```

Out[4]:

```
'1.16.5'
```

In [2]:

```
pip install  genism
```

```
Collecting genism
Note: you may need to restart the kernel to use updated packages.

  ERROR: Could not find a version that satisfies the requirement genism (f
rom versions: none)
ERROR: No matching distribution found for genism
```

In [ ]:

```
pip install --upgrade genism
```

In [4]:

```
pip install gensim
```

```
Requirement already satisfied: gensim in c:\users\ahmad\anaconda3\lib\site
-packages (3.8.3)
Requirement already satisfied: six>=1.5.0 in c:\users\ahmad\anaconda3\lib
\site-packages (from gensim) (1.12.0)
Requirement already satisfied: Cython==0.29.14 in c:\users\ahmad\anaconda3
\lib\site-packages (from gensim) (0.29.14)
Requirement already satisfied: numpy>=1.11.3 in c:\users\ahmad\anaconda3\l
ib\site-packages (from gensim) (1.16.5)
Requirement already satisfied: scipy>=0.18.1 in c:\users\ahmad\anaconda3\l
ib\site-packages (from gensim) (1.3.1)
Requirement already satisfied: smart-open>=1.8.1 in c:\users\ahmad\anacond
a3\lib\site-packages (from gensim) (2.1.1)
Requirement already satisfied: requests in c:\users\ahmad\anaconda3\lib\si
te-packages (from smart-open>=1.8.1->gensim) (2.22.0)
Requirement already satisfied: boto in c:\users\ahmad\anaconda3\lib\site-p
ackages (from smart-open>=1.8.1->gensim) (2.49.0)
Requirement already satisfied: boto3 in c:\users\ahmad\anaconda3\lib\site-
packages (from smart-open>=1.8.1->gensim) (1.14.56)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in
c:\users\ahmad\anaconda3\lib\site-packages (from requests->smart-open>=1.
8.1->gensim) (1.24.2)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\ahmad\anacon
da3\lib\site-packages (from requests->smart-open>=1.8.1->gensim) (2019.9.1
1)
Requirement already satisfied: chardet<3.1.0,>=3.0.2 in c:\users\ahmad\ana
conda3\lib\site-packages (from requests->smart-open>=1.8.1->gensim) (3.0.
4)
Requirement already satisfied: idna<2.9,>=2.5 in c:\users\ahmad\anaconda3
\lib\site-packages (from requests->smart-open>=1.8.1->gensim) (2.8)
Requirement already satisfied: botocore<1.18.0,>=1.17.56 in c:\users\ahmad
\anaconda3\lib\site-packages (from boto3->smart-open>=1.8.1->gensim) (1.1
7.56)
Requirement already satisfied: jmespath<1.0.0,>=0.7.1 in c:\users\ahmad\an
aconda3\lib\site-packages (from boto3->smart-open>=1.8.1->gensim) (0.10.0)
Requirement already satisfied: s3transfer<0.4.0,>=0.3.0 in c:\users\ahmad
\anaconda3\lib\site-packages (from boto3->smart-open>=1.8.1->gensim) (0.3.
3)
Requirement already satisfied: python-dateutil<3.0.0,>=2.1 in c:\users\ahm
ad\anaconda3\lib\site-packages (from botocore<1.18.0,>=1.17.56->boto3->sma
rt-open>=1.8.1->gensim) (2.8.0)
Requirement already satisfied: docutils<0.16,>=0.10 in c:\users\ahmad\anac
onda3\lib\site-packages (from botocore<1.18.0,>=1.17.56->boto3->smart-open
>=1.8.1->gensim) (0.15.2)
Note: you may need to restart the kernel to use updated packages.
```

In [6]:

```
Sentence= “Tokenization is the process of breaking down text document
apart into those pieces”
```

```
  File "<ipython-input-6-9f1be0c3c6b5>", line 1
    Sentence= “Tokenization is the process of breaking down text document
                            ^
SyntaxError: invalid character in identifier
```

In [8]:

```
import gensim as gs
```

In [11]:

```
Sentence = 'Tokenization is the process of breaking down text documentapart into those
 pieces'
```

In [12]:

```
tokenizedWord= list(gs.utils.tokenize(Sentence))
```

In [13]:

```
print(tokenizedWord)
```

```
['Tokenization', 'is', 'the', 'process', 'of', 'breaking', 'down', 'text',
'documentapart', 'into', 'those', 'pieces']
```

In [14]:

```
gs.utils.tokenize
```

Out[14]:

```
<function gensim.utils.tokenize(text, lowercase=False, deacc=False, encodi
ng='utf8', errors='strict', to_lower=False, lower=False)>
```

In [15]:

```
help(gs.utils.tokenize)
```

```
Help on function tokenize in module gensim.utils:

tokenize(text, lowercase=False, deacc=False, encoding='utf8', errors='stri
ct', to_lower=False, lower=False)
    Iteratively yield tokens as unicode strings, optionally removing accen
t marks and lowercasing it.

    Parameters
    ----------
    text : str or bytes
        Input string.
    deacc : bool, optional
        Remove accentuation using :func:`~gensim.utils.deaccent`?
    encoding : str, optional
        Encoding of input string, used as parameter for :func:`~gensim.uti
ls.to_unicode`.
    errors : str, optional
        Error handling behaviour, used as parameter for :func:`~gensim.uti
ls.to_unicode`.
    lowercase : bool, optional
        Lowercase the input string?
    to_lower : bool, optional
        Same as `lowercase`. Convenience alias.
    lower : bool, optional
        Same as `lowercase`. Convenience alias.

    Yields
    ------
    str
        Contiguous sequences of alphabetic characters (no digits!), using
:func:`~gensim.utils.simple_tokenize`

    Examples
    --------
    .. sourcecode:: pycon

        >>> from gensim.utils import tokenize
        >>> list(tokenize('Nic nemůže letět rychlostí vyšší, než 300 tisíc
kilometrů za sekundu!', deacc=True))
        [u'Nic', u'nemuze', u'letet', u'rychlosti', u'vyssi', u'nez', u'ti
sic', u'kilometru', u'za', u'sekundu']
```

In [17]:

```
Sentence= "In computer science, artificial intelligence (AI), sometimes called machine
 intelligence, is intelligence demonstrated by machines, in contrast to the natural int
elligence displayed by humans and animals. Computer science defines AI research as the
 study of intelligent agents: any device that perceives its environment and takes actio
ns that maximize its chance of successfully achieving its goals."
```

In [23]:

```
import gensim
from gensim import corpora
from pprint import pprint
text = ["""In computer science, artificial intelligence (AI), sometimes called machine
 intelligence, is intelligence demonstrated by machines, in contrast to the natural int
elligence displayed by humans and animals. Computer science defines AI research as the
 study of intelligent agents: any device that perceives its environment and takes actio
ns that maximize its chance of successfully achieving its goals."""]
tokens = [[token for token in sentence.split()] for sentence
in text]
gensim_dictionary = corpora.Dictionary()
gensim_corpus = [gensim_dictionary.doc2bow(token,
allow_update=True) for token in tokens]
print(gensim_corpus)
```

```
[[(0, 1), (1, 1), (2, 1), (3, 1), (4, 1), (5, 1), (6, 1), (7, 2), (8, 1),
(9, 1), (10, 1), (11, 1), (12, 2), (13, 1), (14, 1), (15, 1), (16, 1), (1
7, 1), (18, 1), (19, 1), (20, 1), (21, 1), (22, 1), (23, 1), (24, 1), (25,
3), (26, 1), (27, 1), (28, 1), (29, 3), (30, 1), (31, 1), (32, 1), (33,
1), (34, 2), (35, 1), (36, 1), (37, 1), (38, 1), (39, 1), (40, 1), (41,
1), (42, 1), (43, 2), (44, 2), (45, 1)]]
```

In [24]:

```
word_frequencies = [[(gensim_dictionary[id], frequence) for id, frequence in couple] fo
r couple in gensim_corpus]
print(word_frequencies)
```

```
[[('(AI),', 1), ('AI', 1), ('Computer', 1), ('In', 1), ('achieving', 1),
('actions', 1), ('agents:', 1), ('and', 2), ('animals.', 1), ('any', 1),
('artificial', 1), ('as', 1), ('by', 2), ('called', 1), ('chance', 1), ('c
omputer', 1), ('contrast', 1), ('defines', 1), ('demonstrated', 1), ('devi
ce', 1), ('displayed', 1), ('environment', 1), ('goals.', 1), ('humans',
1), ('in', 1), ('intelligence', 3), ('intelligence,', 1), ('intelligent',
1), ('is', 1), ('its', 3), ('machine', 1), ('machines,', 1), ('maximize',
1), ('natural', 1), ('of', 2), ('perceives', 1), ('research', 1), ('scienc
e', 1), ('science,', 1), ('sometimes', 1), ('study', 1), ('successfully',
1), ('takes', 1), ('that', 2), ('the', 2), ('to', 1)]]
```

In [27]:

```
from gensim.utils import simple_preprocess
from smart_open import smart_open
import os

tokens = [simple_preprocess(sentence, deacc=True) for sentence in open(r'E:\filetext.tx
t', encoding='utf-8')]

gensim_dictionary = corpora.Dictionary()
gensim_corpus = [gensim_dictionary.doc2bow(token, allow_update=True) for token in token
s]
word_frequencies = [[(gensim_dictionary[id], frequence) for id, frequence in couple] fo
r couple in gensim_corpus]

print(word_frequencies)
```

```
---------------------------------------------------------------------------
-
FileNotFoundError                         Traceback (most recent call las
t)
<ipython-input-27-cd9e041df49a> in <module>
      3 import os
      4
----> 5 tokens = [simple_preprocess(sentence, deacc=True) for sentence in
 open(r'E:\filetext.txt', encoding='utf-8')]
      6
      7 gensim_dictionary = corpora.Dictionary()

FileNotFoundError: [Errno 2] No such file or directory: 'E:\\filetext.txt'
```

In [ ]:

```

```