:[5] In

```
import sys
sys.version
```

Out[5]:

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

:[6] In

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

Out[6]:

```
'1.16.5'
```

:[7] In

```
pip install gensim
```

```
Requirement already satisfied: gensim in d:\users\d7me_\anaconda3\lib\site-p
(ackages (3.8.3
Requirement already satisfied: numpy>=1.11.3 in d:\users\d7me_\anaconda3\lib
(\site-packages (from gensim) (1.16.5
Requirement already satisfied: Cython==0.29.14 in d:\users\d7me_\anaconda3\l
(ib\site-packages (from gensim) (0.29.14
Requirement already satisfied: six>=1.5.0 in d:\users\d7me_\anaconda3\lib\si
(te-packages (from gensim) (1.12.0
Requirement already satisfied: scipy>=0.18.1 in d:\users\d7me_\anaconda3\lib
(\site-packages (from gensim) (1.3.1
Requirement already satisfied: smart-open>=1.8.1 in d:\users\d7me_\anaconda3
(\lib\site-packages (from gensim) (2.1.1
Requirement already satisfied: boto in d:\users\d7me_\anaconda3\lib\site-pac
(kages (from smart-open>=1.8.1->gensim) (2.49.0
Requirement already satisfied: boto3 in d:\users\d7me_\anaconda3\lib\site-pa
(ckages (from smart-open>=1.8.1->gensim) (1.14.56
Requirement already satisfied: requests in d:\users\d7me_\anaconda3\lib\site
(-packages (from smart-open>=1.8.1->gensim) (2.22.0
Requirement already satisfied: jmespath<1.0.0,>=0.7.1 in d:\users\d7me_\anac
(onda3\lib\site-packages (from boto3->smart-open>=1.8.1->gensim) (0.10.0
Requirement already satisfied: botocore<1.18.0,>=1.17.56 in d:\users\d7me_\a
(naconda3\lib\site-packages (from boto3->smart-open>=1.8.1->gensim) (1.17.56
Requirement already satisfied: s3transfer<0.4.0,>=0.3.0 in d:\users\d7me_\an
(aconda3\lib\site-packages (from boto3->smart-open>=1.8.1->gensim) (0.3.3
Requirement already satisfied: chardet<3.1.0,>=3.0.2 in d:\users\d7me_\anaco
(nda3\lib\site-packages (from requests->smart-open>=1.8.1->gensim) (3.0.4
 Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in
d:\users\d7me_\anaconda3\lib\site-packages (from requests->smart-open>=1.8.1
(->gensim) (1.24.2
Requirement already satisfied: certifi>=2017.4.17 in d:\users\d7me_\anaconda
(3\lib\site-packages (from requests->smart-open>=1.8.1->gensim) (2019.9.11
Requirement already satisfied: idna<2.9,>=2.5 in d:\users\d7me_\anaconda3\li
(b\site-packages (from requests->smart-open>=1.8.1->gensim) (2.8
Requirement already satisfied: python-dateutil<3.0.0,>=2.1 in d:\users\d7me_
\anaconda3\lib\site-packages (from botocore<1.18.0,>=1.17.56->boto3->smart-o
(pen>=1.8.1->gensim) (2.8.0
Requirement already satisfied: docutils<0.16,>=0.10 in d:\users\d7me_\anacon
da3\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
```

:[8] In

```
import gensim as gs
print(gs.__version__)
```

```
3.8.3
```

:[9] In

```
Sentence= 'Tokenization is the process of breaking down text document apart into thosepiece
print(Sentence)
```

```
Tokenization is the process of breaking down text document apart into thosep
ieces
```

:[11] In

```
import gensim as gs
tokenizedWord = list(gs.utils.tokenize(Sentence))
```

:[12] In

```
tokenizedWord
```

Out[12]:

```
            ,'Tokenization']
            ,'is'
            ,'the'
            ,'process'
            ,'of'
            ,'breaking'
            ,'down'
            ,'text'
            ,'document'
            ,'apart'
            ,'into'
            ['thosepieces'
```

:[13] In

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

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

tokenize(text, lowercase=False, deacc=False, encoding='utf8', errors='stric
(t', to_lower=False, lower=False
 Iteratively yield tokens as unicode strings, optionally removing accent
.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.util
.`s.to_unicode
errors : str, optional
Error handling behaviour, used as parameter for :func:`~gensim.util
.`s.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 :f
`unc:`~gensim.utils.simple_tokenize

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

from gensim.utils import tokenize <<<
list(tokenize('Nic nemůže letět rychlostí vyšší, než 300 tisíc k <<<
((ilometrů za sekundu!', deacc=True
u'Nic', u'nemuze', u'letet', u'rychlosti', u'vyssi', u'nez', u'tisi]
['c', u'kilometru', u'za', u'sekundu
```

:[13] In

```
import gensim
from gensim import corpora
from pprint import pprint
text = ["""In computer science, artificial intelligence (AI), sometimes called machine inte

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)
```

```
,(1 ,8) ,(1 ,7) ,(2 ,6) ,(1 ,5) ,(1 ,4) ,(1 ,3) ,(1 ,2) ,(1 ,1) ,(1 ,0)]]
,17) ,(1 ,16) ,(1 ,15) ,(1 ,14) ,(1 ,13) ,(1 ,12) ,(2 ,11) ,(1 ,10) ,(1 ,9)
,(1 ,25) ,(1 ,24) ,(3 ,23) ,(1 ,22) ,(1 ,21) ,(1 ,20) ,(1 ,19) ,(1 ,18) ,(1
,34) ,(1 ,33) ,(2 ,32) ,(1 ,31) ,(1 ,30) ,(1 ,29) ,(1 ,28) ,(3 ,27) ,(1 ,26)
,(2 ,42) ,(2 ,41) ,(1 ,40) ,(1 ,39) ,(1 ,38) ,(1 ,37) ,(1 ,36) ,(1 ,35) ,(1
[[(1 ,43)
```

:[14] In

```
print(gensim_dictionary)
```

```
Dictionary(44 unique tokens: ['(AI),', 'AI', 'Computer', 'In', 'action
(...['s
```

:[15] In

```
word_frequencies = [[(gensim_dictionary[id], frequence) for id, frequence in couple] for cc
print(word_frequencies)
```

```
AI),', 1), ('AI', 1), ('Computer', 1), ('In', 1), ('actions', 1), ('age)')]]
nts:', 1), ('and', 2), ('animals.', 1), ('any', 1), ('artificial', 1), ('a
s', 1), ('by', 2), ('called', 1), ('chance', 1), ('computer', 1), ('define
s', 1), ('demonstrated', 1), ('device', 1), ('displayed', 1), ('environmen
 t', 1), ('goals.', 1), ('humans', 1), ('incontrast', 1), ('intelligence',
3), ('intelligence,', 1), ('intelligent', 1), ('is', 1), ('its', 3), ('machi
ne', 1), ('machines,', 1), ('maximize', 1), ('natural', 1), ('of', 2), ('per
 ceives', 1), ('research', 1), ('science', 1), ('science,', 1), ('sometimes',
 1), ('study', 1), ('successfullyachieving', 1), ('takes', 1), ('that', 2),
[[(('the', 2), ('to', 1
```

:HW1

:[18] In

```
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'D:\Users\D7me_\Ana
gensim_dictionary = corpora.Dictionary()
gensim_corpus = [gensim_dictionary.doc2bow(token, allow_update=True) for token in tokens]
word_frequencies = [[(gensim_dictionary[id], frequence) for id, frequence in couple] for cc
print(word_frequencies)
```

```
actions', 1), ('agents', 1), ('ai', 2), ('and', 2), ('animals', 1), ('an')]]
 y', 1), ('artificial', 1), ('as', 1), ('by', 2), ('called', 1), ('chance',
1), ('computer', 2), ('defines', 1), ('demonstrated', 1), ('device', 1), ('d
 isplayed', 1), ('environment', 1), ('goals', 1), ('humans', 1), ('in', 1),
('incontrast', 1), ('intelligence', 4), ('intelligent', 1), ('is', 1), ('it
 s', 3), ('machine', 1), ('machines', 1), ('maximize', 1), ('natural', 1),
 ('of', 2), ('perceives', 1), ('research', 1), ('science', 2), ('sometimes',
[[(1), ('study', 1), ('takes', 1), ('that', 2), ('the', 2), ('to', 1
```