|  |  |
| --- | --- |
| In[1] | import gensim  import codecs  import numpy as np  import pandas as pd  import jieba  import jieba.analyse  from scipy import stats |

|  |  |
| --- | --- |
| In[2] | def file2doc(file\_name):  """ tokenize file to doc """  jieba.load\_userdict("./data/tcm\_all\_dict\_2.txt")  doc = [w for x in codecs.open(file\_name, 'r', 'utf-8').readlines() for w in jieba.cut(x.strip())]  return doc |

|  |  |
| --- | --- |
| In[3] | def stop\_word(filename):  stopwords = [line.strip() for line in open(filename, 'r', encoding='utf-8').readlines()]  return stopwords |

|  |  |
| --- | --- |
| In[4] | def file2doc\_stopword(doc,file\_stopword):  """ delete stopword from doc """  stopword\_list=stop\_word(file\_stopword)  doc\_stopword=[y for y in doc if y not in stopword\_list]  return doc\_stopword |

|  |  |
| --- | --- |
| In[5] | def doc2vec(doc, model):  """ :param model: pre-train sentence vectors model """  #start\_alpha = 0.01  #infer\_epoch = 1000  # text convert to sentence vector  #doc\_vec = model.infer\_vector(doc, alpha=start\_alpha, steps=infer\_epoch)  doc\_vec = model.infer\_vector(doc)  return doc\_vec |

|  |  |
| --- | --- |
| In[6] | def get\_vec(file,model):  file\_stopword="./data/stopwords.txt"  doc1=file2doc(file)  doc1stop=file2doc\_stopword(doc1,file\_stopword)  doc1\_vec1=doc2vec(doc1,model)  return doc1\_vec1 |

|  |  |
| --- | --- |
| In[7] | def sim\_cal(vec1, vec2):  """ :return: cosin similarity rate """  vec1mod = np.sqrt(vec1.dot(vec1))  vec2mod = np.sqrt(vec2.dot(vec2))  if vec2mod != 0 and vec1mod != 0:  sim\_prob = (vec1.dot(vec2)) / (vec1mod \* vec2mod)  else:  sim\_prob = 0  return sim\_prob |

|  |  |
| --- | --- |
| In[8] | model\_path = './model/zhiwiki\_news.doc2vec'  model = gensim.models.Doc2Vec.load(model\_path) |

|  |  |
| --- | --- |
| In[9] | file1 = './data/test1.txt'  file2 = './data/test2.txt'  vec1=get\_vec(file1,model)  vec2=get\_vec(file2,model)  sim\_2v1=sim\_cal(vec1, vec2)  print(sim\_2v1) |
| Out[9] | 0.102343425 |

|  |  |
| --- | --- |
| In[10] | file1 = './data/tcm\_hui\_feng.txt'  file2 = './data/tcm\_han\_feng.txt'  vec1=get\_vec(file1,model)  vec2=get\_vec(file2,model)  sim\_2v1=sim\_cal(vec1, vec2)  print(sim\_2v1) |
| Out[10] | 0.7527455 |

|  |  |
| --- | --- |
| In[11] | file1 = './data/tcm\_hui\_outline.txt'  file2 = './data/tcm\_han\_outline.txt'  vec1=get\_vec(file1,model)  vec2=get\_vec(file2,model)  sim\_2v1=sim\_cal(vec1, vec2)  print(sim\_2v1) |
| Out[11] | 0.8775975 |