>>> from gensim import corpora, models, similarities

>>> import logging

>>> documents = ["total death toll in Kunming terror attack rises to 29", "China ready for annual two sessions", "China no one is above the law", "manhunt after terror attack", "Kunming restores order after deadly terror attack"]

>>> texts = [[word for word in document.lower().split()] for document in documents]

>>> print texts

[['total', 'death', 'toll', 'in', 'kunming', 'terror', 'attack', 'rises', 'to', '29'], ['china', 'ready', 'for', 'annual', 'two', 'sessions'], ['china', 'no', 'one', 'is', 'above', 'the', 'law'], ['manhunt', 'after', 'terror', 'attack'], ['kunming', 'restores', 'order', 'after', 'deadly', 'terror', 'attack']]

>>> dictionary = corpora.Dictionary(texts)

>>> print dictionary

Dictionary(27 unique tokens)

>>> print dictionary.token2id

{'is': 17, 'one': 20, 'in': 3, 'ready': 13, 'total': 9, 'death': 2, 'for': 12, 'no': 19, 'deadly': 24, '29': 0, 'to': 7, 'attack': 1, 'manhunt': 23, 'above': 16, 'toll': 8, 'terror': 6, 'law': 18, 'sessions': 14, 'after': 22, 'kunming': 4, 'china': 11, 'two': 15, 'rises': 5, 'restores': 26, 'annual': 10, 'the': 21, 'order': 25}

>>> corpus = [dictionary.doc2bow(text) for text in texts]

>>> print corpus

[[(0, 1), (1, 1), (2, 1), (3, 1), (4, 1), (5, 1), (6, 1), (7, 1), (8, 1), (9, 1)], [(10, 1), (11, 1), (12, 1), (13, 1), (14, 1), (15, 1)], [(11, 1), (16, 1), (17, 1), (18, 1), (19, 1), (20, 1), (21, 1)], [(1, 1), (6, 1), (22, 1), (23, 1)], [(1, 1), (4, 1), (6, 1), (22, 1), (24, 1), (25, 1), (26, 1)]]

>>> tfidf = models.TfidfModel(corpus)

>>> corpus\_tfidf = tfidf[corpus]

>>> for doc in corpus\_tfidf:

print doc

[(0, 0.36452661036094103), (1, 0.11569848807358354), (2, 0.36452661036094103), (3, 0.36452661036094103), (4, 0.2075335444845301), (5, 0.36452661036094103), (6, 0.11569848807358354), (7, 0.36452661036094103), (8, 0.36452661036094103), (9, 0.36452661036094103)]

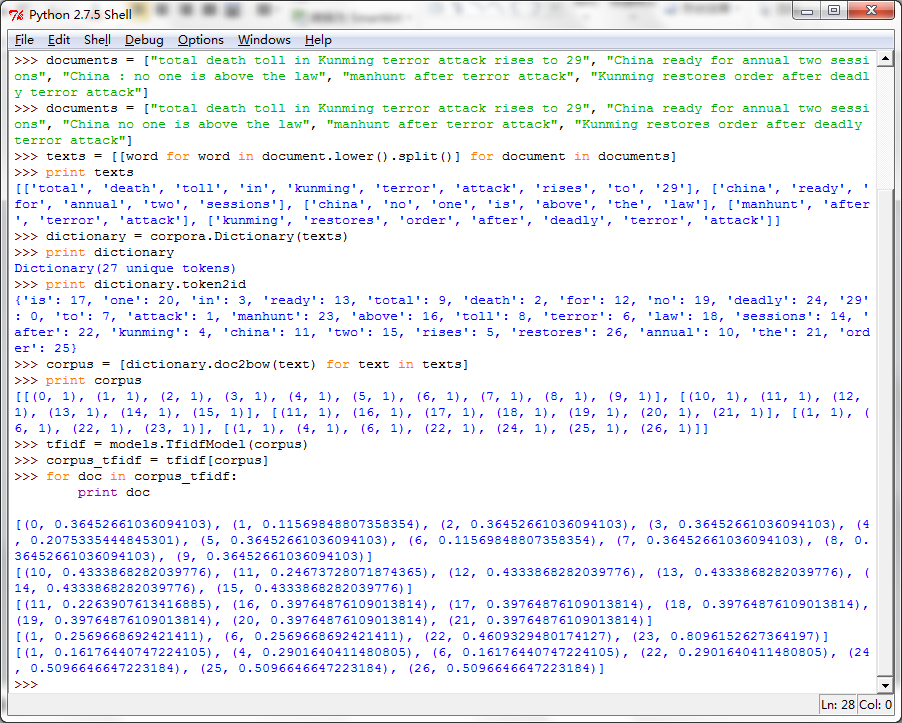
[(10, 0.4333868282039776), (11, 0.24673728071874365), (12, 0.4333868282039776), (13, 0.4333868282039776), (14, 0.4333868282039776), (15, 0.4333868282039776)]

[(11, 0.2263907613416885), (16, 0.39764876109013814), (17, 0.39764876109013814), (18, 0.39764876109013814), (19, 0.39764876109013814), (20, 0.39764876109013814), (21, 0.39764876109013814)]

[(1, 0.2569668692421411), (6, 0.2569668692421411), (22, 0.4609329480174127), (23, 0.8096152627364197)]

[(1, 0.16176440747224105), (4, 0.2901640411480805), (6, 0.16176440747224105), (22, 0.2901640411480805), (24, 0.5096646647223184), (25, 0.5096646647223184), (26, 0.5096646647223184)]

>>>



>>> lda = models.LdaModel(corpus\_tfidf, id2word=dictionary, num\_topics=2)

>>> lda.show\_topics(2)

['0.042\*two + 0.042\*after + 0.042\*restores + 0.041\*annual + 0.041\*china + 0.041\*order + 0.041\*ready + 0.041\*deadly + 0.041\*for + 0.041\*sessions', '0.052\*manhunt + 0.047\*after + 0.041\*attack + 0.041\*terror + 0.040\*toll + 0.040\*kunming + 0.040\*rises + 0.040\*29 + 0.040\*total + 0.040\*to']

