Master Informatique, parcours MALIA

Carnets de note Python pour le cours de Network Analysis for Information Retrieval

Julien Velcin, laboratoire ERIC, Université Lyon 2

Visualisation (partie 3)

In [14]:
```
%matplotlib inline

from nltk.tokenize import RegexpTokenizer
from nltk.corpus import stopwords
from gensim import corpora, models
import gensim
import pandas as pd
import numpy as np
from sklearn.preprocessing import normalize
import matplotlib.pyplot as plt
from matplotlib import cm
import os
from gensim.models.ldamodel import LdaModel
from gensim.test.utils import datapath
from gensim.utils import simple_preprocess
import pandas
```

Dictionnaire et corpus

In [15]:
```
# lecture des données pour un fichier texte simple
with open(os.path.join("datasets", "Frank Herbert - Dune.txt")) as f:
    lines = [line.strip() for line in f.readlines()]
doc_set = lines

# fonction qui génère les listes de mots (token) à partir des textes
def sent_to_words(sentences):
    for sentence in sentences:
        yield(simple_preprocess(str(sentence), deacc=True))  # deacc=True removes punctuations

# on construit le corpus
data_words = list(sent_to_words(doc_set))
# nombre total de documents
ndocs = len(data_words)
```

In [16]:
```
from nltk.corpus import stopwords
stop_words = stopwords.words('english')

def remove_stopwords(texts):
    return [[word for word in simple_preprocess(str(doc)) if word not in stop_words] for doc in texts]

# on retire les mots-outils
data_words_nostops = remove_stopwords(data_words)

# création du dictionnaire
dico = corpora.Dictionary(data_words_nostops)

# ce qui permet par ex. de filtrer le vocabulaire
dico.filter_extremes(no_below=10)

# Create Corpus
texts = data_words_nostops

# matrice Term Document Frequency
corpus = [dico.doc2bow(text) for text in texts]
```

In [17]:
```
ntopics = 50
temp_file = "models/model_dataconf_" + str(ntopics)
```

```
ldamodel = LdaModel.load(temp_file)
```

In [18]:
```
pwz = ldamodel.get_topics()

print("On peut récupérer la matrice stockant p(w/z):", pwz.shape)

# on peut aussi utiliser ldamodel.get_topic_terms(topicid, topn=n) pour obtenir
# les top n mots via leur identifiant, accompagnés de la proba p(w/z)

#ldamodel.get_topic_terms(1,topn=len(dico))
```

```
On peut récupérer la matrice stockant p(w/z): (50, 1809)
```

In [19]:
```
# show_topics permet d'afficher les mots directement
ldamodel.show_topics(num_topics=ntopics,formatted=False)
```

Out[19]:
```
[(0,
  [('fremen', 0.39152008),
   ('use', 0.05933129),
   ('wish', 0.04528319),
   ('al', 0.0389602),
   ('desert', 0.03285488),
   ('live', 0.028301666),
   ('remember', 0.023284208),
   ('clear', 0.021177702),
   ('one', 0.020621184),
   ('different', 0.018773999)]),
 (1,
  [('duke', 0.20715399),
   ('leto', 0.087413795),
   ('atreides', 0.073362716),
   ('son', 0.07008426),
   ('said', 0.056732424),
   ('alia', 0.04982056),
   ('daughter', 0.03306871),
   ('long', 0.032335553),
   ('one', 0.02794743),
   ('thus', 0.02754769)]),
 (2,
  [('gurney', 0.18998945),
   ('halleck', 0.09835282),
   ('said', 0.06459459),
   ('go', 0.063076265),
   ('called', 0.05651979),
   ('paul', 0.05225922),
   ('made', 0.043874662),
   ('man', 0.030175516),
   ('instant', 0.023096582),
   ('darkness', 0.022464601)]),
 (3,
  [('know', 0.14806195),
   ('said', 0.10589882),
   ('old', 0.09825239),
   ('woman', 0.07482917),
   ('lord', 0.07100542),
   ('yes', 0.061191496),
   ('ah', 0.04418092),
   ('young', 0.034209512),
   ('perhaps', 0.03125972),
   ('one', 0.027654435)]),
 (4,
  [('saw', 0.17431653),
   ('another', 0.1079708),
   ('felt', 0.08253656),
   ('paul', 0.043813605),
   ('thing', 0.038380653),
   ('someone', 0.036623828),
   ('jessica', 0.03478227),
   ('killed', 0.034575544),
   ('watched', 0.03234983),
   ('guards', 0.024033742)]),
 (5,
  [('mother', 0.2499172),
   ('reverend', 0.098743),
   ('noted', 0.05245002),
   ('attention', 0.051718373),
   ('power', 0.050754163),
   ('often', 0.02744687),
   ('within', 0.025123289),
   ('food', 0.02458133),
   ('jessica', 0.02452121),
   ('produced', 0.022216829)]),
 (6,
  [('poison', 0.10300394),
   ('get', 0.06693086),
   ('hold', 0.052068826),
   ('message', 0.042719156),
   ('melange', 0.03934495),
   ('said', 0.035228346),
   ('reason', 0.03232111),
```

```
  ('orders', 0.02729153),
  ('produce', 0.02685917),
  ('assassins', 0.02628131)]),
 (7,
  [('asked', 0.26652384),
  ('idaho', 0.07488156),
  ('paul', 0.06701268),
  ('basin', 0.064606704),
  ('jessica', 0.0585036),
  ('green', 0.055228002),
  ('wild', 0.037388813),
  ('subtle', 0.033820696),
  ('duncan', 0.029701157),
  ('throat', 0.026203008)]),
 (8,
  [('rock', 0.05552254),
  ('light', 0.04187618),
  ('sand', 0.037450355),
  ('moved', 0.03314947),
  ('hulud', 0.027010739),
  ('shai', 0.027010739),
  ('across', 0.024885878),
  ('ahead', 0.023748005),
  ('white', 0.022351231),
  ('pack', 0.021845443)]),
 (9,
  [('guild', 0.13922215),
  ('yet', 0.09629041),
  ('imperium', 0.04694776),
  ('human', 0.046749588),
  ('slowly', 0.03803798),
  ('report', 0.03631439),
  ('mean', 0.035089944),
  ('ancient', 0.034857806),
  ('simple', 0.03068954),
  ('best', 0.029148176)]),
 (10,
  [('kynes', 0.1901735),
  ('man', 0.15439048),
  ('said', 0.06626435),
  ('used', 0.054833483),
  ('every', 0.05121806),
  ('kill', 0.03222243),
  ('liet', 0.023382148),
  ('way', 0.02231818),
  ('everything', 0.02079899),
  ('ritual', 0.020643305)]),
 (11,
  [('eyes', 0.15014988),
  ('stared', 0.070243515),
  ('paul', 0.061641477),
  ('night', 0.039855924),
  ('feet', 0.039196406),
  ('usul', 0.035300333),
  ('rocks', 0.035215285),
  ('cave', 0.031640723),
  ('weapon', 0.030102286),
  ('man', 0.028910171)]),
 (12,
  [('stilgar', 0.16005315),
  ('gesserit', 0.124152415),
  ('bene', 0.124152415),
  ('said', 0.055640686),
  ('count', 0.053947408),
  ('nodded', 0.041490674),
  ('dead', 0.03711738),
  ('fenring', 0.034534775),
  ('paul', 0.03212949),
  ('way', 0.031594284)]),
 (13,
  [('say', 0.13123083),
  ('want', 0.064935185),
  ('three', 0.058421325),
  ('haderach', 0.037611097),
```

```
  ('kwisatz', 0.037611097),
  ('figure', 0.035037443),
  ('suddenly', 0.034354415),
  ('ring', 0.032839257),
  ('per', 0.030763177),
  ('said', 0.03014509)]),
(14,
 [('turned', 0.15377325),
  ('away', 0.061913818),
  ('beneath', 0.04617186),
  ('paul', 0.043457527),
  ('back', 0.038851023),
  ('lay', 0.034332138),
  ('robe', 0.033794086),
  ('question', 0.030437224),
  ('answer', 0.028789742),
  ('factory', 0.027553469)]),
(15,
 [('within', 0.12079146),
  ('shield', 0.115792155),
  ('touched', 0.052811593),
  ('field', 0.045379505),
  ('force', 0.041331105),
  ('fighting', 0.04130821),
  ('slow', 0.031256907),
  ('paul', 0.028739836),
  ('shields', 0.028225858),
  ('convention', 0.02400368)]),
(16,
 [('shall', 0.07530684),
  ('arm', 0.06649621),
  ('well', 0.06122491),
  ('upon', 0.048249193),
  ('certain', 0.041403834),
  ('set', 0.03584441),
  ('control', 0.034840427),
  ('said', 0.028866824),
  ('better', 0.02823575),
  ('free', 0.026955582)]),
(17,
 [('room', 0.07471277),
  ('yueh', 0.056084093),
  ('around', 0.03934243),
  ('paul', 0.039013818),
  ('table', 0.03804828),
  ('stood', 0.03416303),
  ('along', 0.030667052),
  ('hall', 0.027272547),
  ('stopped', 0.025716942),
  ('passage', 0.024208717)]),
(18,
 [('could', 0.13298492),
  ('knew', 0.089784525),
  ('training', 0.06192537),
  ('sietch', 0.05764583),
  ('new', 0.052461527),
  ('trained', 0.040432587),
  ('gave', 0.03589431),
  ('school', 0.033736963),
  ('fact', 0.03327423),
  ('one', 0.026005318)]),
(19,
 [('time', 0.18876338),
  ('nothing', 0.07228166),
  ('fear', 0.05561686),
  ('past', 0.051214196),
  ('real', 0.032396667),
  ('spread', 0.030337205),
  ('common', 0.027748667),
  ('effect', 0.026459336),
  ('gone', 0.025573675),
  ('looking', 0.024868153)]),
(20,
 [('first', 0.14794263),
```

```
  ('need', 0.08422058),
  ('like', 0.07192972),
  ('returned', 0.04296963),
  ('entire', 0.042536426),
  ('arrakeen', 0.03754136),
  ('pattern', 0.026617154),
  ('old', 0.024872828),
  ('except', 0.024270514),
  ('time', 0.023078674)]),
(21,
 [('great', 0.109255135),
  ('many', 0.090155095),
  ('things', 0.065794155),
  ('one', 0.06487031),
  ('always', 0.046765085),
  ('already', 0.036172904),
  ('rule', 0.033393335),
  ('order', 0.03266201),
  ('said', 0.026155818),
  ('animal', 0.020401852)]),
(22,
 [('arrakis', 0.2361108),
  ('never', 0.10496085),
  ('told', 0.06406042),
  ('lady', 0.055734463),
  ('person', 0.048116848),
  ('filled', 0.041881867),
  ('known', 0.034880113),
  ('become', 0.027011689),
  ('houses', 0.025671722),
  ('accepted', 0.025291461)]),
(23,
 [('name', 0.10908154),
  ('god', 0.08313867),
  ('world', 0.07303766),
  ('missionaria', 0.03912499),
  ('protectiva', 0.03912499),
  ('legend', 0.038937718),
  ('came', 0.036601163),
  ('brought', 0.036292277),
  ('faced', 0.03527855),
  ('removed', 0.035121955)]),
(24,
 [('rabban', 0.07662282),
  ('child', 0.07149293),
  ('knife', 0.066393815),
  ('troop', 0.061824944),
  ('slave', 0.046868138),
  ('matter', 0.038629983),
  ('plant', 0.03557482),
  ('hangings', 0.028979111),
  ('eye', 0.02778295),
  ('metal', 0.02706986)]),
(25,
 [('emperor', 0.19092879),
  ('sardaukar', 0.11565087),
  ('floor', 0.05338152),
  ('enough', 0.05272819),
  ('battle', 0.04544667),
  ('understand', 0.04512182),
  ('caught', 0.04104359),
  ('imperial', 0.038640518),
  ('said', 0.028641896),
  ('majesty', 0.026929954)]),
(26,
 [('hawat', 0.2437169),
  ('give', 0.077500366),
  ('tell', 0.07032109),
  ('whispered', 0.069836),
  ('silence', 0.061967067),
  ('said', 0.060754567),
  ('thufir', 0.04743242),
  ('paul', 0.03740274),
  ('beginning', 0.028406613),
```

```
  ('anger', 0.0282704)]),
 (27,
  [('would', 0.15959162),
   ('come', 0.081279874),
   ('might', 0.059862405),
   ('still', 0.05386671),
   ('could', 0.049363762),
   ('beyond', 0.045764096),
   ('far', 0.044910934),
   ('silent', 0.03256906),
   ('ever', 0.028048769),
   ('kind', 0.025106387)]),
 (28,
  [('people', 0.1619574),
   ('thopter', 0.05377284),
   ('dune', 0.045623135),
   ('paul', 0.04025301),
   ('hidden', 0.0375455),
   ('strength', 0.03720771),
   ('five', 0.032819156),
   ('remembered', 0.028475665),
   ('wait', 0.026875647),
   ('dream', 0.025707114)]),
 (29,
  [('see', 0.3100301),
   ('feyd', 0.14444727),
   ('rautha', 0.12849264),
   ('maker', 0.057586282),
   ('boy', 0.040199015),
   ('part', 0.032277953),
   ('went', 0.028768046),
   ('found', 0.020856211),
   ('kept', 0.02038998),
   ('upward', 0.018641936)]),
 (30,
  [('water', 0.21714847),
   ('spice', 0.13389687),
   ('life', 0.11129377),
   ('little', 0.048089817),
   ('moisture', 0.034343205),
   ('said', 0.030983191),
   ('ten', 0.02099759),
   ('open', 0.02096758),
   ('inner', 0.01970159),
   ('meters', 0.016941449)]),
 (31,
  [('men', 0.18793824),
   ('door', 0.058052193),
   ('two', 0.053999927),
   ('air', 0.03820895),
   ('one', 0.037673946),
   ('guard', 0.029868923),
   ('pressed', 0.029328724),
   ('ledge', 0.026458904),
   ('friend', 0.025432592),
   ('leader', 0.02478685)]),
 (32,
  [('house', 0.10341825),
   ('almost', 0.089465365),
   ('call', 0.06752987),
   ('blue', 0.04593871),
   ('grew', 0.034978997),
   ('secundus', 0.034355097),
   ('salusa', 0.034355097),
   ('creature', 0.03243554),
   ('planet', 0.032094117),
   ('turn', 0.031954914)]),
 (33,
  [('face', 0.1614681),
   ('held', 0.08341671),
   ('keep', 0.059335694),
   ('sound', 0.05067602),
   ('mouth', 0.038340487),
   ('uncle', 0.03558729),
```

```
  ('help', 0.034966707),
  ('watch', 0.03223605),
  ('strange', 0.029975617),
  ('arms', 0.028220711)]),
 (34,
  [('paul', 0.106312685),
  ('glanced', 0.08900973),
  ('back', 0.08326162),
  ('good', 0.072035946),
  ('said', 0.07147465),
  ('day', 0.06884489),
  ('blood', 0.036237806),
  ('leave', 0.035924207),
  ('love', 0.032002695),
  ('group', 0.024414087)]),
 (35,
  [('word', 0.06841086),
  ('began', 0.058234576),
  ('jihad', 0.05318428),
  ('blade', 0.04210628),
  ('command', 0.03983185),
  ('also', 0.039496444),
  ('mapes', 0.03519532),
  ('butlerian', 0.033036098),
  ('dry', 0.032688536),
  ('crysknife', 0.03012572)]),
 (36,
  [('beside', 0.064983),
  ('step', 0.053780716),
  ('lips', 0.050606735),
  ('forced', 0.04596928),
  ('times', 0.04507487),
  ('hood', 0.036330454),
  ('knowledge', 0.032825857),
  ('secret', 0.030876935),
  ('cold', 0.029466107),
  ('said', 0.027508616)]),
 (37,
  [('hand', 0.12654708),
  ('looked', 0.095903546),
  ('right', 0.08561679),
  ('left', 0.073751815),
  ('paul', 0.06479004),
  ('without', 0.060826868),
  ('wondered', 0.033846065),
  ('though', 0.031194657),
  ('spoke', 0.029636558),
  ('away', 0.023874078)]),
 (38,
  [('desert', 0.09074755),
  ('took', 0.079913445),
  ('hands', 0.06374237),
  ('deep', 0.06330271),
  ('open', 0.04944181),
  ('worm', 0.046051513),
  ('second', 0.033141635),
  ('last', 0.030333806),
  ('forward', 0.0263377),
  ('breath', 0.023970338)]),
 (39,
  [('muad', 0.2001924),
  ('dib', 0.2001924),
  ('princess', 0.050506797),
  ('done', 0.048964284),
  ('future', 0.047125936),
  ('irulan', 0.039398663),
  ('arrakis', 0.031725932),
  ('words', 0.03125862),
  ('carry', 0.024958812),
  ('consider', 0.022794586)]),
 (40,
  [('sand', 0.13671437),
  ('dust', 0.055911824),
  ('surface', 0.050789136),
```

```
  ('across', 0.047704767),
  ('system', 0.04188492),
  ('dunes', 0.03828411),
  ('wind', 0.03617175),
  ('bible', 0.033219904),
  ('movement', 0.03137355),
  ('toward', 0.026327815)]),
 (41,
  [('must', 0.14934741),
  ('said', 0.060816288),
  ('may', 0.058116272),
  ('make', 0.05455242),
  ('among', 0.05367483),
  ('harkonnens', 0.048813354),
  ('course', 0.037656985),
  ('us', 0.037137102),
  ('thought', 0.03592772),
  ('religion', 0.035897903)]),
 (42,
  [('voice', 0.15205632),
  ('said', 0.08256458),
  ('heard', 0.076858655),
  ('cannot', 0.06855382),
  ('jessica', 0.057900995),
  ('something', 0.055622473),
  ('paul', 0.042244274),
  ('true', 0.037390944),
  ('tone', 0.032773767),
  ('hear', 0.026412904)]),
 (43,
  [('baron', 0.30991906),
  ('said', 0.0787332),
  ('take', 0.06491038),
  ('piter', 0.06337164),
  ('mentat', 0.043625418),
  ('change', 0.040966578),
  ('stop', 0.032754492),
  ('women', 0.02879308),
  ('could', 0.023631535),
  ('ship', 0.02096868)]),
 (44,
  [('side', 0.049000643),
  ('body', 0.03977899),
  ('stillsuit', 0.03936191),
  ('space', 0.036432575),
  ('awareness', 0.03549086),
  ('eyes', 0.028812343),
  ('lifted', 0.02636212),
  ('face', 0.020750437),
  ('studied', 0.020696018),
  ('felt', 0.019998286)]),
 (45,
  [('head', 0.110304356),
  ('years', 0.062038325),
  ('storm', 0.06142045),
  ('shook', 0.043833572),
  ('soon', 0.036718074),
  ('stare', 0.035834733),
  ('snapped', 0.03448375),
  ('bring', 0.032602303),
  ('talk', 0.03010532),
  ('appeared', 0.028082756)]),
 (46,
  [('thought', 0.17845885),
  ('think', 0.07309521),
  ('father', 0.061636817),
  ('paul', 0.05410372),
  ('let', 0.04916001),
  ('jessica', 0.04410644),
  ('look', 0.044026565),
  ('religious', 0.0415195),
  ('put', 0.03560068),
  ('seen', 0.034135517)]),
 (47,
```

```
 [('place', 0.1344542),
  ('planet', 0.12025114),
  ('even', 0.07199668),
  ('much', 0.055267043),
  ('caladan', 0.044730254),
  ('arrakis', 0.04418067),
  ('test', 0.03477705),
  ('full', 0.031960007),
  ('followed', 0.031850494),
  ('show', 0.030024385)]),
 (48,
  [('harkonnen', 0.14870979),
  ('us', 0.14105566),
  ('point', 0.048578393),
  ('behind', 0.04706231),
  ('death', 0.037594385),
  ('half', 0.037526365),
  ('sire', 0.03346051),
  ('one', 0.030724728),
  ('six', 0.02548095),
  ('said', 0.024854934)]),
 (49,
  [('high', 0.08901027),
  ('small', 0.0531295),
  ('presently', 0.040310267),
  ('family', 0.032458954),
  ('major', 0.03198854),
  ('obvious', 0.030293174),
  ('form', 0.029185886),
  ('onto', 0.028794333),
  ('higher', 0.028585242),
  ('permitted', 0.02844581)])]
```

On peut vérifier que les mots ont une probabilité d'appartenir à *plusieurs* thématiques

jolie visualisation avec pyLDAvis

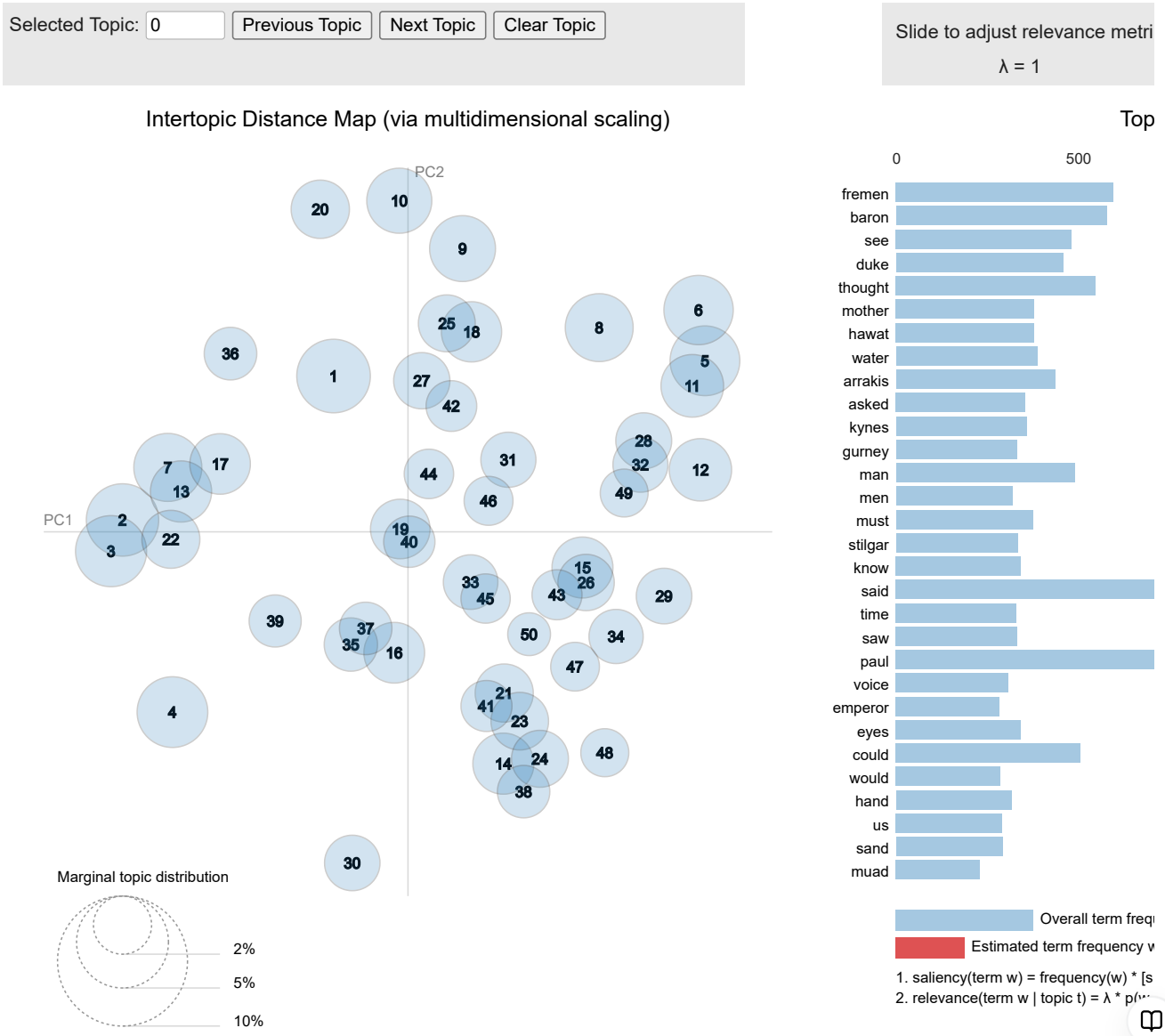
Attention, les thématiques sont remunérotées.

In [20]:
```
import pyLDAvis
import pyLDAvis.gensim_models as gensimvis
pyLDAvis.enable_notebook()
lda_viz = gensimvis.prepare(ldamodel, gensim.matutils.corpus2csc(corpus), dictionary=ldamodel.id2word)
pyLDAvis.display(lda_viz)

#import pyLDAvis.gensim
#pyLDAvis.enable_notebook()
#vis = pyLDAvis.gensim.prepare(ldamodel, gensim.matutils.corpus2csc(corpus), dictionary=ldamodel.id2word)
#vis
#pyLDAvis.display(vis)
```

```
/Users/jvelcin/miniforge3/envs/tf2022/lib/python3.10/site-packages/pyLDAvis/_prepare.py:247: FutureWarning: In a f
ure version of pandas all arguments of DataFrame.drop except for the argument 'labels' will be keyword-only.
  by='saliency', ascending=False).head(R).drop('saliency', 1)
```

Out[20]:



Si on souhaite obtenir p(z|d), il faut réexécuter le modèle sur les données (par ex., le corpus).

In [21]:
```
ldc = ldamodel[corpus]
```

Les procédures suivantes fournissent plusieurs "vues" intéressantes sur le modèle. Elles viennent du site *machinelearningplus.com* :

https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/

Tout d'abord, on souhaite un tableau qui liste la thématique majoritaire pour chaque document, accompagnée par ses mots les plus probables.

In [22]:
```
def format_topics_sentences(ldamodel, corpus, texts):
    # Init output
    sent_topics_df = pd.DataFrame()

    # Get main topic in each document
    i=0
    for i, row in enumerate(ldamodel[corpus]):
        row = sorted(row[0], key=lambda x: (x[1]), reverse=True)
        #print(row)
        # Get the Dominant topic, Perc Contribution and Keywords for each document
```

```
        for j, (topic_num, prop_topic) in enumerate(row):
            if j == 0:  # => dominant topic
                wp = ldamodel.show_topic(topic_num)
                topic_keywords = ", ".join([word for word, prop in wp])
                sent_topics_df = sent_topics_df.append(pd.Series([int(topic_num), round(prop_topic,4), topic_keywords
            else:
                break
    sent_topics_df.columns = ['Dominant_Topic', 'Perc_Contribution', 'Topic_Keywords']

    # Add original text to the end of the output
    contents = pd.Series(texts)
    sent_topics_df = pd.concat([sent_topics_df, contents], axis=1)
    return(sent_topics_df)


df_topic_sents_keywords = format_topics_sentences(ldamodel=ldamodel, corpus=corpus, texts=doc_set)

# Format
df_dominant_topic = df_topic_sents_keywords.reset_index()
df_dominant_topic.columns = ['Document_No', 'Dominant_Topic', 'Topic_Perc_Contrib', 'Keywords', 'Text']

# Show
df_dominant_topic.head(10)
```

```
/var/folders/44/_q8kssp12vb3ks1rlb59jm6m0000gp/T/ipykernel_22966/3836988281.py:15: FutureWarning: The frame.append m
ethod is deprecated and will be removed from pandas in a future version. Use pandas.concat instead.
  sent_topics_df = sent_topics_df.append(pd.Series([int(topic_num), round(prop_topic,4), topic_keywords]), ignore_in
dex=True)
```

Out[22]:

	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text
0	0	28	0.51	people, thopter, dune, paul, hidden, strength,...	Dune
1	1	0	0.02	fremen, use, wish, al, desert, live, remember,...	Frank Herbert
2	2	0	0.02	fremen, use, wish, al, desert, live, remember,...	
3	3	0	0.02	fremen, use, wish, al, desert, live, remember,...	Copyright 1965
4	4	0	0.02	fremen, use, wish, al, desert, live, remember,...	
5	5	49	0.51	high, small, presently, family, major, obvious...	Book 1
6	6	28	0.51	people, thopter, dune, paul, hidden, strength,...	DUNE
7	7	0	0.02	fremen, use, wish, al, desert, live, remember,...	
8	8	0	0.02	fremen, use, wish, al, desert, live, remember,...	= = = = = =
9	9	0	0.02	fremen, use, wish, al, desert, live, remember,...	

On peut vouloir obtenir les documents les plus "représentatifs" de chaque thématique (attention, au sens de p(z|d)).

In [23]:

```
sent_topics_sorteddf = pd.DataFrame()

sent_topics_outdf_grpd = df_topic_sents_keywords.groupby('Dominant_Topic')

for i, grp in sent_topics_outdf_grpd:
    sent_topics_sorteddf = pd.concat([sent_topics_sorteddf,
                                      grp.sort_values(['Perc_Contribution'], ascending=[0]).head(1)],
                                     axis=0)

# Reset Index
sent_topics_sorteddf.reset_index(drop=True, inplace=True)

# Format
sent_topics_sorteddf.columns = ['Topic_Num', "Topic_Perc_Contrib", "Keywords", "Text"]

# Show
sent_topics_sorteddf
```

Out[23]:

	Topic_Num	Topic_Perc_Contrib	Keywords	Text
0	0	0.7550	fremen, use, wish, al, desert, live, remember,...	"Lisan al-Gaib!"
1	1	0.7800	duke, leto, atreides, son, said, alia, daughte...	Jessica spoke bitterly: "Chips in the path of ...
2	2	0.8367	gurney, halleck, said, go, called, paul, made,...	And Paul: "Gurney, man! Gurney, man!"
3	3	0.7550	know, said, old, woman, lord, yes, ah, young, ...	"When we've rested," Jessica said, "we should ...
4	4	0.6733	saw, another, felt, paul, thing, someone, jess...	Paul collapsing their tent, recovering it up t...
5	5	0.8040	mother, reverend, noted, attention, power, oft...	"Reverend Mother!" Chani said. "What is wrong?"
6	6	0.7550	poison, get, hold, message, melange, said, rea...	"I'm perfectly safe here," Paul said.
7	7	0.7550	asked, idaho, paul, basin, jessica, green, wil...	"Are there any plants down there?" Paul asked.
8	8	0.9423	rock, light, sand, moved, hulud, shai, across,...	Slowly, the filtered sun buried itself beneath...
9	9	0.7550	guild, yet, imperium, human, slowly, report, m...	"I'm a soldier of the Imperium," Paul said, "t...
10	10	0.7550	kynes, man, said, used, every, kill, liet, way...	"At once, Liet," the man said.
11	11	0.7550	eyes, stared, paul, night, feet, usul, rocks, ...	Paul's eyes closed.
12	12	0.7550	stilgar, gesserit, bene, said, count, nodded, ...	"Bene Gesserit ain't all highborn," the pilot ...
13	13	0.8040	say, want, three, haderach, kwisatz, figure, s...	"If you're not the Kwisatz Haderach," Jessica ...
14	14	0.8040	turned, away, beneath, paul, back, lay, robe, ...	Lump-lump-lump-lump!
15	15	0.6733	within, shield, touched, field, force, fightin...	All the while his mind was adding sense impres...
16	16	0.7550	shall, arm, well, upon, certain, set, control,...	"Try the communinet receiver again," Paul said.
17	17	0.7550	room, yueh, around, paul, table, stood, along,...	Paul remained bent over his studies.
18	18	0.5100	could, knew, training, sietch, new, trained, g...	She stiffened.
19	19	0.7901	time, nothing, fear, past, real, spread, commo...	"Fear is the mind-killer. Fear is the little d...
20	20	0.6733	first, need, like, returned, entire, arrakeen,...	They glided lower . . . lower . . .
21	21	0.5342	great, many, things, one, always, already, rul...	One said: "A great-great-great grandmother of ...
22	22	0.8040	arrakis, never, told, lady, person, filled, kn...	"I vowed never to regret my decision," Jessica...
23	23	0.6733	name, god, world, missionaria, protectiva, leg...	"Who said it?" Harah repeated.
24	24	0.6733	rabban, child, knife, troop, slave, matter, pl...	"Such was my suspicion," he said.
25	25	0.6733	emperor, sardaukar, floor, enough, battle, und...	"I am your ruler," the Emperor said.
26	26	0.7550	hawat, give, tell, whispered, silence, said, t...	"Like a fairyland," Paul whispered.
27	27	0.6733	would, come, might, still, could, beyond, far,...	"They're still examining the dead."
28	28	0.5536	people, thopter, dune, paul, hidden, strength,...	Silence fell like a blanket on the cavern.
29	29	0.6733	see, feyd, rautha, maker, boy, part, went, fou...	Maker? Maker.
30	30	0.7550	water, spice, life, little, moisture, said, te...	"Indeed," Paul said.
31	31	0.7550	men, door, two, air, one, guard, pressed, ledg...	"That shouldn't have happened," Paul said. "I ...
32	32	0.6733	house, almost, call, blue, grew, secundus, sal...	"For the funeral plain," he said.
33	33	0.7550	face, held, keep, sound, mouth, uncle, help, w...	"They fit the description," Paul said.
34	34	0.8040	paul, glanced, back, good, said, day, blood, l...	"Run!" Jessica screamed. "Paul, run!"
35	35	0.7377	word, began, jihad, blade, command, also, mape...	Jessica tried to swallow in a dry throat, said...
36	36	0.7550	beside, step, lips, forced, times, hood, knowl...	"It was Otheym," Paul said. "He was listening."
37	37	0.7550	hand, looked, right, left, paul, without, wond...	"It . . . seemed the right way."
38	38	0.7550	desert, took, hands, deep, open, worm, second,...	"Worm," Paul said.
39	39	0.8911	muad, dib, princess, done, future, irulan, arr...	"Muad'Dib! Muad'Dib! Muad'Dib! Muad'Dib!"
40	40	0.6275	sand, dust, surface, across, system, dunes, wi...	DRUM SAND: impaction of sand in such away that...
41	41	0.7550	must, said, may, make, among, harkonnens, cour...	"This way, sir," Nefud said.
42	42	0.8367	voice, said, heard, cannot, jessica, something...	"You cannot do this thing," Jessica said. "Pau...
43	43	0.7550	baron, said, take, piter, mentat, change, stop...	"I will take the duchy," Piter said.

	Topic_Num	Topic_Perc_Contrib	Keywords	Text
44	44	0.6733	side, body, stillsuit, space, awareness, eyes,...	Jessica crossed to him.
45	45	0.8367	head, years, storm, shook, soon, stare, snappe...	Immediately, their nostrils were assailed by t...
46	46	0.8600	thought, think, father, paul, let, jessica, lo...	They think Paul's toying with Jamis, Jessica t...
47	47	0.6733	place, planet, even, much, caladan, arrakis, t...	"You should conserve your energies for the tes...
48	48	0.8040	harkonnen, us, point, behind, death, half, sir...	"Jetflares behind us!" Jessica said.
49	49	0.6733	high, small, presently, family, major, obvious...	"He's not our only hope," she said.

Pour finir, le "volume" estimé de documents (en réalité, de mots) couverts par les différentes thématiques.

In [24]:
```
# Number of Documents for Each Topic
topic_counts = df_topic_sents_keywords['Dominant_Topic'].value_counts()
```

In [25]:
```
dim_space = ntopics

doc_vec = np.zeros(shape=(ndocs,dim_space))
#id_docs_nonvides = []

for i, d in enumerate(ldamodel[corpus]):
    for j, (topic, poids) in enumerate(d[0]):
        doc_vec[i, topic] = poids
```

In [32]:
```
import umap
import umap.plot

mapper_lda = umap.UMAP(metric='cosine').fit(doc_vec)
umap.plot.points(mapper_lda, labels=np.array(df_dominant_topic.Dominant_Topic))

hover_data = pd.DataFrame({'index': np.arange(1, ndocs+1),
                           'label': [doc_set[i] for i in range(ndocs)],
                           'topic': df_dominant_topic.Dominant_Topic})
#p = umap.plot.interactive(mapper, labels=id_docs_nonvides, hover_data=hover_data, point_size=2)
#p = umap.plot.interactive(mapper_lda, hover_data=hover_data, point_size=2)
p = umap.plot.interactive(mapper_lda, labels=df_dominant_topic.Dominant_Topic, hover_data=hover_data, point_size=2)
umap.plot.show(p)
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
Cell In [32], line 4
      1 import umap
      2 import umap.plot
----> 4 mapper_lda = umap.UMAP(metric='cosine').fit(doc_vec)
      5 umap.plot.points(mapper_lda, labels=np.array(df_dominant_topic.Dominant_Topic))
      7 hover_data = pd.DataFrame({'index': np.arange(1, ndocs+1),
      8                            'label': [doc_set[i] for i in range(ndocs)],
      9                            'topic': df_dominant_topic.Dominant_Topic})

AttributeError: module 'umap' has no attribute 'UMAP'
```

Post-traitement

In [33]:
```
#topic_tohide = [1, 5, 6, 13, 39]
topic_tohide = [0, 3]

doc_filtered = [id_doc for id_doc in range(ndocs) if len(corpus[id_doc])>2 and df_dominant_topic.Dominant_Topic[id_do

len(doc_filtered)
```

Out[33]:
```
6671
```

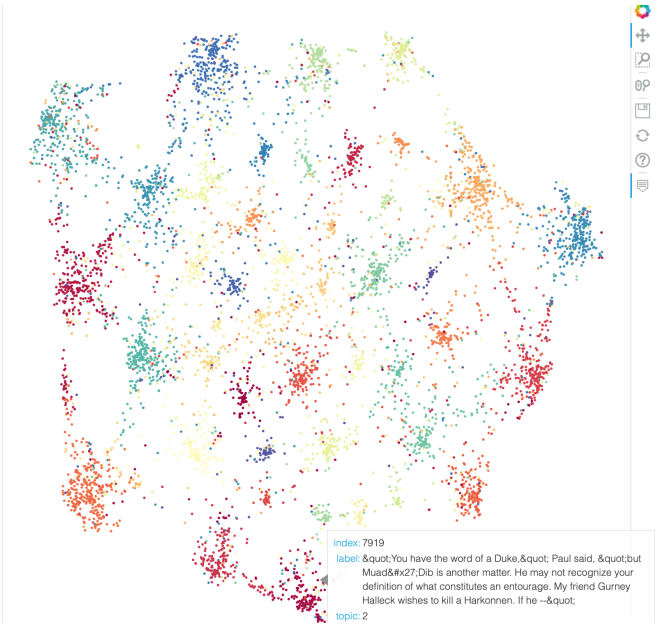
In [34]:
```
mapper_lda_small = umap.UMAP(metric='cosine').fit(doc_vec[doc_filtered])

hover_data = pd.DataFrame({'index': [i+1 for i in doc_filtered],
                           'label': [doc_set[i] for i in doc_filtered],
                           'topic': [df_dominant_topic.Dominant_Topic[i] for i in doc_filtered]})
p = umap.plot.interactive(mapper_lda_small, labels=[df_dominant_topic.Dominant_Topic[i] for i in doc_filtered], hover
```

```
umap.plot.show(p)
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
Cell In [34], line 1
----> 1 mapper_lda_small = umap.UMAP(metric='cosine').fit(doc_vec[doc_filtered])
      3 hover_data = pd.DataFrame({'index': [i+1 for i in doc_filtered],
      4                            'label': [doc_set[i] for i in doc_filtered],
      5                            'topic': [df_dominant_topic.Dominant_Topic[i] for i in doc_filtered]})
      6 p = umap.plot.interactive(mapper_lda_small, labels=[df_dominant_topic.Dominant_Topic[i] for i in doc_filtere
d], hover_data=hover_data, point_size=2)

AttributeError: module 'umap' has no attribute 'UMAP'
```



In [ ]: