In [1]:
```python
%matplotlib inline

import gensim
import json
import logging
import math
import os
import random
import re
import spacy
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import pyLDAvis.gensim_models
import warnings

from gensim import corpora, models
from gensim.utils import simple_preprocess
from gensim.models.ldamodel import LdaModel
from gensim.models import CoherenceModel
from gensim.models import Word2Vec
from gensim.test.utils import datapath

from nltk.corpus import stopwords
from nltk import pos_tag, ne_chunk
from nltk.tokenize import word_tokenize
from nltk.collocations import BigramCollocationFinder
from nltk.metrics import BigramAssocMeasures
from nltk.collocations import TrigramCollocationFinder
from nltk.metrics import TrigramAssocMeasures
from nltk.tokenize import RegexpTokenizer

from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction.text import ENGLISH_STOP_WORDS
from sklearn.preprocessing import normalize

from stop_words import get_stop_words
from wordcloud import WordCloud
```

```
C:\Users\brune\anaconda3\envs\CoursHN\lib\site-packages\tqdm\auto.py:22: TqdmWarning: IProgress not found. Please update jupyter and ipywidgets. See https://ipywidgets.readthedocs.io/en/stable/user_install.html
  from .autonotebook import tqdm as notebook_tqdm
```

In [2]:
```python
with open(os.path.join('posts_maior_propre.json'), 'r', encoding = 'UTF-8') as fin_M2 :
    data = json.load(fin_M2)
```

In [3]:
```python
index = ''
a = ''
list_clean_content = []

for j in range(len(data)):
    index = data[j]
    a = index.replace("\n", "")
    list_clean_content.append(a)
```

In [4]:
```python
sp_stop_words = get_stop_words('spanish')
fr_stop_words = get_stop_words('french')
en_stop_words = get_stop_words('english')
```

In [5]:
```python
words = [ 'plus', 'rien', 'tres', 'ca', 'fr', 'com', 'www', 'http',
          'etais', 'etait', 'etaient', 'meme', 'non', 'bien', 'oui', 'quoi',
          'chez', 'va', 'apres']
```

In [6]:
```python
stop_words = sp_stop_words + fr_stop_words + en_stop_words + words
```

In [7]:
```python
# 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(list_clean_content))
```

In [8]:
```python
doc_set=list_clean_content
```

In [9]:
```python
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)
```

In [10]:
```python
# création du dictionnaire
dico = corpora.Dictionary(data_words_nostops)

### Filtering rare and extreme vocabulary ###
dico.filter_extremes( no_below = 5, no_above = 0.5)

# Create Corpus
texts = data_words_nostops

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

In [11]:
```python
ldamodel = LdaModel(corpus,
                    id2word = dico,
                    num_topics=40,
                    passes=2,
                    random_state=100,
                    per_word_topics=True)
```

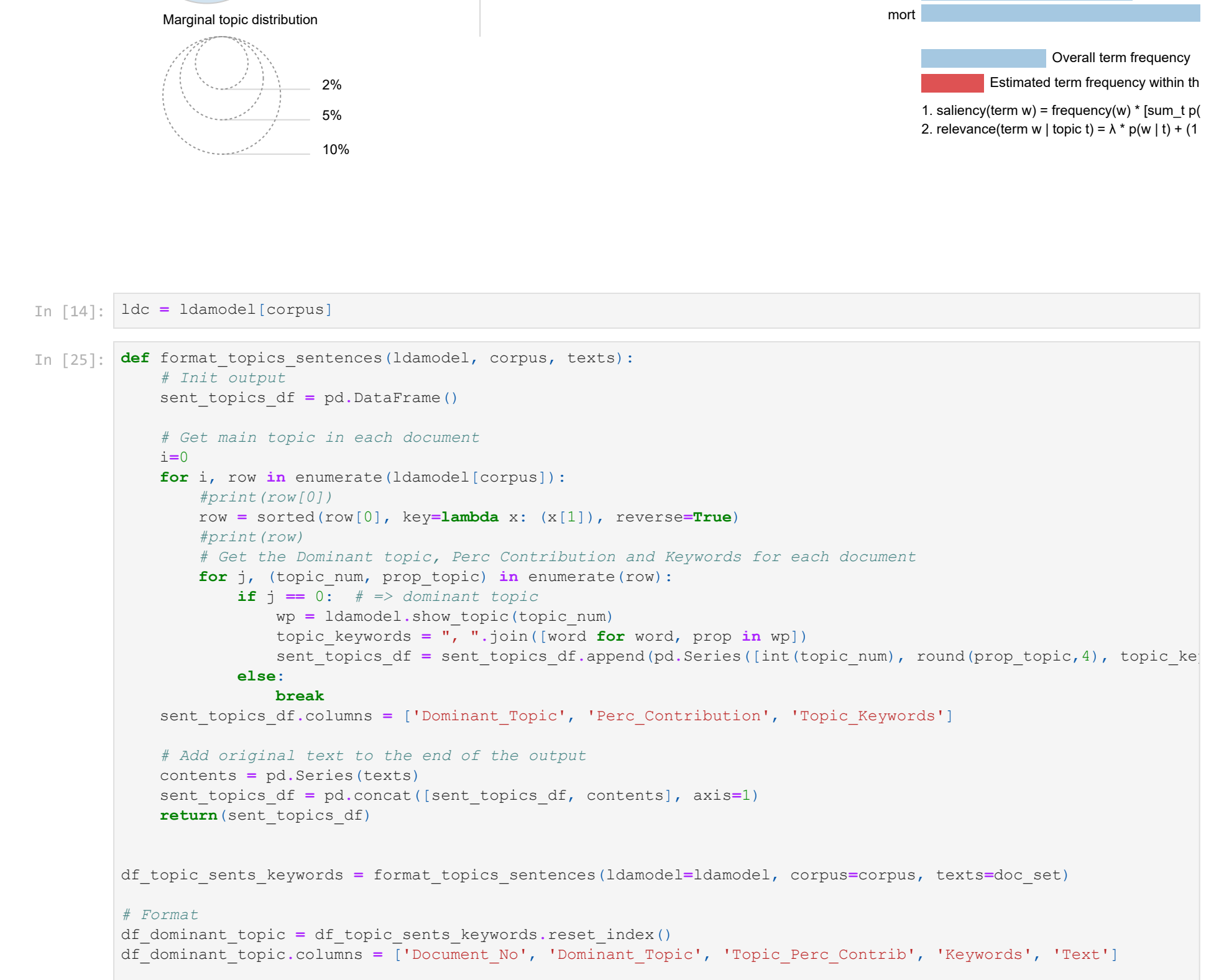
In [12]:
```python
ldamodel.show_topics(num_topics=40,formatted=False)
```

Out[12]:
```
[(0,
  [('roman', 0.011101952),
   ('dont', 0.008585664),
   ('histoire', 0.007922815),
   ('auteur', 0.005627373),
   ('oeuvre', 0.005332705),
   ('livre', 0.0052453353),
   ('litterature', 0.005192275),
   ('annees', 0.00496529),
   ('personnages', 0.0047211247),
   ('recit', 0.0047044616)]),
 (1,
  [('photo', 0.10563041),
   ('ile', 0.06708874),
   ('mer', 0.06300689),
   ('port', 0.03232534),
   ('cm', 0.021750832),
   ('avion', 0.019874325),
   ('matricule', 0.014484915),
   ('ambre', 0.011962216),
   ('ulysse', 0.011766081),
   ('quentin', 0.011090266)]),
 (2,
  [('sexe', 0.011813908),
   ('masculin', 0.010036173),
   ('animal', 0.009403951),
   ('croire', 0.00875692),
   ('lucie', 0.007320646),
   ('dune', 0.006791551),
   ('sonner', 0.0067498265),
   ('feminin', 0.0063092643),
   ('baleine', 0.006126382),
   ('violet', 0.005972298)]),
 (3,
  [('anna', 0.036052696),
   ('victor', 0.029417023),
   ('hugo', 0.026904456),
   ('didier', 0.016143993),
   ('robin', 0.01599371),
   ('nicolas', 0.014850031),
   ('monique', 0.014482595),
   ('lou', 0.014392301),
   ('roger', 0.014165274),
   ('perec', 0.013537045)]),
 (4,
  [('blog', 0.047791142),
   ('da', 0.044379514),
   ('blogspot', 0.034695435),
   ('blogs', 0.03297753),
   ('jose', 0.029070895),
   ('ana', 0.024379136),
   ('didier', 0.023496967),
   ('bill', 0.022934926),
   ('net', 0.02167708),
   ('maria', 0.020801922)]),
 (5,
  [('berri', 0.00981315),
   ('mort', 0.0097363),
   ('uqam', 0.00747298),
   ('tait', 0.0070196637),
   ('moscou', 0.0065944208),
   ('connaissez', 0.006521534),
   ('effacement', 0.006508368),
   ('chaussee', 0.006124028),
   ('suite', 0.005412718),
   ('bus', 0.004652498)]),
 (6,
  [('police', 0.04343523),
   ('https', 0.02326987),
   ('facebook', 0.019532032),
   ('vote', 0.019363359),
   ('people', 0.017353207),
   ('org', 0.016730448),
   ('justice', 0.016691575),
   ('social', 0.014368969),
   ('voter', 0.013335323),
   ('elections', 0.011087669)]),
 (7,
  [('pays', 0.017626511),
   ('france', 0.01674816),
   ('contre', 0.013170539),
   ('guerre', 0.0128056),
   ('francais', 0.009851971),
   ('politique', 0.009653286),
   ('monde', 0.009026532),
   ('etre', 0.006364158),
   ('etat', 0.0063582165),
   ('europe', 0.0057254247)]),
 (8,
  [('pen', 0.019782037),
   ('chiffres', 0.01702911),
   ('ira', 0.016228285),
   ('ales', 0.015969828),
   ('tourne', 0.014993431),
   ('figaro', 0.014924936),
   ('portier', 0.014909508),
   ('planes', 0.014603688),
   ('mondial', 0.014366594),
   ('pareils', 0.012010158)]),
 (9,
  [('pain', 0.029873764),
   ('table', 0.029772151),
   ('manger', 0.018066229),
   ('cuisine', 0.013598998),
   ('mange', 0.013487048),
   ('huile', 0.012613796),
   ('soupe', 0.011575765),
   ('repas', 0.011465147),
   ('cafe', 0.011324211),
   ('vin', 0.011085428)]),
 (10,
  [('clown', 0.025304463),
   ('flash', 0.022583785),
   ('cirque', 0.017678037),
   ('cherie', 0.017394362),
   ('verres', 0.015859766),
   ('jeff', 0.013888322),
   ('jimmy', 0.013690372),
   ('sonia', 0.013133268),
   ('teste', 0.012663293),
   ('hop', 0.012433876)]),
 (11,
  [('sherbrooke', 0.011265041),
   ('paade', 0.007153103),
   ('attila', 0.006909236),
   ('amical', 0.006747978),
   ('prisonniers', 0.006465918),
   ('snoutre', 0.006445079),
   ('esti', 0.0057940406),
   ('samson', 0.005322526),
   ('cristal', 0.0049359864),
   ('oui', 0.004620105)]),
 (12,
  [('michel', 0.13745925),
   ('menard', 0.052007463),
   ('web', 0.036099542),
   ('cliquant', 0.03518885),
   ('elise', 0.03286806),
   ('sanchez', 0.02109452),
   ('detacheurs', 0.015068993),
   ('interruption', 0.013867157),
   ('souviendrai', 0.013320693),
   ('mamemoure', 0.011957005)]),
 (13,
  [('po', 0.10196769),
   ('kevin', 0.03823895),
   ('dormir', 0.030404245),
   ('oeil', 0.025264606),
   ('bunker', 0.015513225),
   ('conserver', 0.014066189),
   ('feuilleton', 0.012938899),
   ('hebdo', 0.012613367),
   ('charlie', 0.010775536),
   ('petrole', 0.009717751)]),
 (14,
  [('tete', 0.0054453597),
   ('vieux', 0.005355363),
   ('dit', 0.004616103),
   ('main', 0.003622188),
   ('bois', 0.003529901),
   ('jusqu', 0.0034843239),
   ('pieds', 0.003195614),
   ('petite', 0.003127610),
   ('chien', 0.0030531099),
   ('eau', 0.0030297441)]),
 (15,
  [('etre', 0.017184015),
   ('dire', 0.00878966),
   ('chose', 0.006883013),
   ('faut', 0.006784626),
   ('reponse', 0.006512394),
   ('toujours', 0.0064358295),
   ('autres', 0.0060326327),
   ('temps', 0.005993726),
   ('jamais', 0.005583673),
   ('dont', 0.005414577)]),
 (16,
  [('terre', 0.027109569),
   ('espace', 0.01646082),
   ('lumiere', 0.010953558),
   ('eau', 0.01084253),
   ('feu', 0.01082302),
   ('ciel', 0.008784806),
   ('ombre', 0.008669026),
   ('image', 0.008132656),
   ('oiseaux', 0.0075235516),
   ('arbre', 0.0069327066)]),
 (17,
  [('mas', 0.03615788),
   ('dias', 0.02752002),
   ('annie', 0.025030555),
   ('claro', 0.024726804),
   ('jourde', 0.022501536),
   ('ups', 0.021729162),
   ('video', 0.016804708),
   ('vida', 0.01447008),
   ('twitter', 0.013254809),
   ('dia', 0.012649873)]),
 (18,
  [('new', 0.044183303),
   ('york', 0.028162936),
   ('billy', 0.02026076),
   ('john', 0.017526416),
   ('albin', 0.015736518),
   ('richard', 0.015175309),
   ('court', 0.014827885),
   ('us', 0.013863526),
   ('william', 0.012290369),
   ('poet', 0.011636457)]),
 (19,
  [('ville', 0.03822032),
   ('rue', 0.036508527),
   ('paris', 0.036504604),
   ('poste', 0.033857275),
   ('saint', 0.018485365),
   ('prenoms', 0.017562455),
   ('carte', 0.015100711),
   ('mercredi', 0.0111594135),
   ('gare', 0.010197011),
   ('place', 0.009625371)]),
 (20,
  [('etes', 0.022733731),
   ('like', 0.017674943),
   ('vote', 0.017049538),
   ('monsieur', 0.015762266),
   ('don', 0.012953114),
   ('zero', 0.012058582),
   ('color', 0.011976959),
   ('boudou', 0.01104442),
   ('commission', 0.010055038),
   ('dan', 0.010006335)]),
 (21,
  [('alberto', 0.03257417),
   ('singe', 0.026756318),
   ('pp', 0.023928482),
   ('ronds', 0.022651684),
   ('fabien', 0.02162022),
   ('carlo', 0.019853277),
   ('big', 0.01726185),
   ('simone', 0.016228972),
   ('jene', 0.015177432),
   ('dion', 0.014304489)]),
 (22,
  [('flaubert', 0.016030053),
   ('zoe', 0.014461725),
   ('ronda', 0.014348427),
   ('jen', 0.011903159),
   ('helene', 0.011164072),
   ('marde', 0.010882738),
   ('sou', 0.009122027),
   ('leo', 0.008476625),
   ('balzac', 0.007839109),
   ('galets', 0.006913106)]),
 (23,
  [('musique', 0.06872127),
   ('chanson', 0.027771836),
   ('album', 0.01666328),
   ('chansons', 0.0151069295),
   ('rock', 0.012930039),
   ('piano', 0.012609311),
   ('voix', 0.012311879),
   ('ecouter', 0.011517609),
   ('notes', 0.011093418),
   ('jouer', 0.010537321)]),
 (24,
  [('film', 0.12836154),
   ('cinema', 0.060517896),
   ('films', 0.04212035),
   ('voir', 0.02021951),
   ('prix', 0.019950459),
   ('eme', 0.0152905015),
   ('mc', 0.012214608),
   ('bande', 0.011731129),
   ('polar', 0.011364644),
   ('ecran', 0.011267191)]),
 (25,
  [('etre', 0.010304897),
   ('temps', 0.009720486),
   ('puis', 0.008019847),
   ('poste', 0.00615475),
   ('pis', 0.005720505),
   ('quelques', 0.0055577517),
   ('voir', 0.0055313725),
   ('petit', 0.00519967),
   ('dire', 0.0049159047),
   ('sais', 0.0046760733)]),
 (26,
  [('jean', 0.052546784),
   ('pierre', 0.020562198),
   ('marie', 0.019727176),
   ('dominique', 0.019079002),
   ('francois', 0.018158479),
   ('pascale', 0.016836517),
   ('thomas', 0.015579885),
   ('laurent', 0.015509267),
   ('mathieu', 0.014910009),
   ('eric', 0.012879365)]),
 (27,
  [('information', 0.006404726),
   ('mme', 0.0063094613),
   ('anti', 0.0057870178),
   ('help', 0.0049152425),
   ('bilan', 0.0046702987),
   ('pression', 0.0040328195),
   ('strategie', 0.0038576403),
   ('plan', 0.003555109),
   ('robot', 0.0034956188),
   ('opinion', 0.0032962847)]),
 (28,
  [('livre', 0.039784137),
   ('commentaire', 0.035757348),
   ('lire', 0.025142977),
   ('livres', 0.020235589),
   ('texte', 0.017553007),
   ('lecture', 0.016143316),
   ('auteur', 0.014648168),
   ('ecrire', 0.01434796),
   ('lu', 0.011205039),
   ('ecriture', 0.011138941)]),
 (29,
  [('cartes', 0.06113293),
   ('japonais', 0.016243046),
   ('usa', 0.015528237),
   ('circuit', 0.014800466),
   ('postales', 0.014341312),
   ('japon', 0.012966949),
   ('brique', 0.011768036),
   ('malle', 0.009846795),
   ('pont', 0.008749446),
   ('balneaire', 0.008409815)]),
 (30,
  [('bin', 0.024310058),
   ('quidam', 0.00864893),
   ('medecin', 0.008494701),
   ('fit', 0.0064377007),
   ('docteur', 0.0062160073),
   ('recre', 0.0057989787),
   ('hopital', 0.005157824),
   ('orthographe', 0.004696835),
   ('jour', 0.004676036),
   ('prit', 0.004323449)]),
 (31,
  [('black', 0.043804944),
   ('go', 0.026279418),
   ('re', 0.02444408),
   ('change', 0.024304202),
   ('man', 0.023472274),
   ('will', 0.022895504),
   ('can', 0.021011198),
   ('one', 0.020277724),
   ('now', 0.018367691),
   ('time', 0.017689068)]),
 (32,
  [('dit', 0.020421939),
   ('ans', 0.018394057),
   ('pere', 0.015933914),
   ('mere', 0.015705744),
   ('fille', 0.012785979),
   ('vie', 0.011627079),
   ('aime', 0.011499694),
   ('sais', 0.00955296),
   ('enfants', 0.009439512),
   ('jeune', 0.009349012)]),
 (33,
  [('nuit', 0.00919319),
   ('yeux', 0.009016523),
   ('corps', 0.0078747645),
   ('jour', 0.0067957076),
   ('ciel', 0.006271554),
   ('mots', 0.006125003),
   ('temps', 0.005898598),
   ('voix', 0.0057349117),
   ('vent', 0.005674794),
   ('vers', 0.0051622298)]),
 (34,
  [('photos', 0.0338626),
   ('art', 0.031978444),
   ('exposition', 0.025861122),
   ('musee', 0.02399846),
   ('act', 0.020101586),
   ('photographie', 0.017413449),
   ('artiste', 0.016809084),
   ('johnson', 0.015408865),
   ('images', 0.013240184),
   ('street', 0.013060853)]),
 (35,
  [('edmond', 0.050326426),
   ('dr', 0.018269936),
   ('boris', 0.01806194),
   ('adam', 0.010033128),
   ('visibilite', 0.0097950455),
   ('chou', 0.009568162),
   ('innocent', 0.009297292),
   ('recherche', 0.008808638),
   ('eve', 0.008206173),
   ('daniele', 0.0081537336)]),
 (36,
  [('chat', 0.019092156),
   ('matin', 0.016220162),
   ('blanc', 0.015981901),
   ('rouge', 0.014025664),
   ('petit', 0.013913314),
   ('neige', 0.013548078),
   ('rose', 0.012810685),
   ('jardin', 0.012238647),
   ('hiver', 0.011994887),
   ('noir', 0.011867278)]),
 (37,
  [('poesie', 0.05272058),
   ('poete', 0.033578306),
   ('poeme', 0.03263667),
   ('poemes', 0.02128032),
   ('poetes', 0.014492928),
   ('editions', 0.01042907),
   ('mots', 0.008142182),
   ('pierre', 0.007785834),
   ('vers', 0.007695649),
   ('recueil', 0.0074246496)]),
 (38,
  [('theatre', 0.021654895),
   ('saison', 0.014631087),
   ('scene', 0.013766229),
   ('spectacle', 0.009517956),
   ('semaine', 0.009245773),
   ('samedi', 0.009245657),
   ('annee', 0.009084101),
   ('public', 0.008290013),
   ('vendredi', 0.007998084),
   ('ete', 0.007551393)]),
 (39,
  [('monde', 0.018956311),
   ('vie', 0.018897047),
   ('homme', 0.014170518),
   ('amour', 0.008187717),
   ('mort', 0.008030022),
   ('etre', 0.007702533),
   ('temps', 0.006993648),
   ('corps', 0.006515796),
   ('hommes', 0.0058392677),
   ('ame', 0.005487328)])]
```

In [13]:
```python
pyLDAvis.enable_notebook()
vis = pyLDAvis.gensim_models.prepare(ldamodel, gensim.matutils.corpus2csc(corpus), dictionary=ldamodel.id2word)
pyLDAvis.display(vis)
```

```
C:\Users\brune\anaconda3\envs\CoursHN\lib\site-packages\pyLDAvis\_prepare.py:247: FutureWarning: In a future 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[13]:



In [14]:
```python
ldc = ldamodel[corpus]
```

In [25]:
```python
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]):
        #print(row[0])
        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_ke
            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(50)
```

```
C:\Users\brune\AppData\Local\Temp\ipykernel_20372\2926149746.py:16: FutureWarning: The frame.append method 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_index=True)
```

Out[25]:

	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text
0	0	33	0.3164	nuit, yeux, corps, jour, ciel, mots, temps, vo...	le premier jour du premier mois de l'année res...
1	1	0	0.0250	roman, dont, histoire, auteur, œuvre, livre, l...	
2	2	15	0.6686	etre, dire, chose, faut, reponse, toujours, au...	âmes perduescauses perduesun constat s'impose ...
3	3	33	0.3797	nuit, yeux, corps, jour, ciel, mots, temps, vo...	pour y croîtrej'ai enterré quelque chose de mo...
4	4	0	0.0250	roman, dont, histoire, auteur, œuvre, livre, l...	
5	5	33	0.2747	nuit, yeux, corps, jour, ciel, mots, temps, vo...	des sentiments le poursuivent qui lui cuisent ...
6	6	39	0.4677	monde, vie, homme, amour, mort, etre, temps, c...	y a des journées parfaites et fébriles où j'ar...
7	7	15	0.4668	etre, dire, chose, faut, reponse, toujours, au...	je ne suis ni normand ni féministe ni chanteur...
8	8	16	0.3261	terre, espace, lumiere, eau, feu, ciel, ombre,...	je ne sais pas ce qui me peine et je sais à p...
9	9	33	0.4009	nuit, yeux, corps, jour, ciel, mots, temps, vo...	sans les méchants les saints ne le sont guère ...
10	10	15	0.5665	etre, dire, chose, faut, reponse, toujours, au...	si parler n'est pas agir alors les mots ne peu...
11	11	32	0.4658	dit, ans, pere, mere, fille, vie, aime, sais, ...	les questions me perdent et les réponses me fo...
12	12	33	0.3167	nuit, yeux, corps, jour, ciel, mots, temps, vo...	la magie parfois c'est arrêter d'être subtil l...
13	13	15	0.2664	etre, dire, chose, faut, reponse, toujours, au...	tu as toujours cette facilité à aller chercher...
14	14	15	0.2872	etre, dire, chose, faut, reponse, toujours, au...	Toujours pas de nouvelles de mon frère, recue...
15	15	15	0.4321	etre, dire, chose, faut, reponse, toujours, au...	je vous préviens j'ai une terrible envie de ba...
16	16	33	0.8926	nuit, yeux, corps, jour, ciel, mots, temps, vo...	je regarde le ciel et la mer derrière une fenê...
17	17	0	0.0250	roman, dont, histoire, auteur, œuvre, livre, l...	
18	18	39	0.2748	monde, vie, homme, amour, mort, etre, temps, c...	DANGERTOUTES LES FAMILLES NUISENTDANGERTOUTES...
19	19	0	0.0250	roman, dont, histoire, auteur, œuvre, livre, l...	
20	20	32	0.3971	dit, ans, pere, mere, fille, vie, aime, sais, ...	la ville se referme sur l'angoisse et le dram...
21	21	9	0.3025	pain, table, manger, cuisine, mange, huile, so...	-à Christine Spadaccini aucune couleur au...
22	22	33	0.4348	nuit, yeux, corps, jour, ciel, mots, temps, vo...	même si le crachin bleu reste jouablema chevil...
23	23	0	0.0250	roman, dont, histoire, auteur, œuvre, livre, l...	
24	24	33	0.3494	nuit, yeux, corps, jour, ciel, mots, temps, vo...	nous sommes graines que le vent pousse nous so...
25	25	1	0.3418	photo, ile, mer, port, cm, avion, matricule, a...	Ajouter une légende
26	26	0	0.0250	roman, dont, histoire, auteur, œuvre, livre, l...	
27	27	30	0.2595	bin, quidam, medecin, fit, docteur, recre, hop...	Stella se souvient en jouant on apprend autan...
28	28	39	0.4560	monde, vie, homme, amour, mort, etre, temps, c...	elle cherche un moyen de contacter les anges ...
29	29	0	0.0250	roman, dont, histoire, auteur, œuvre, livre, l...	
30	30	25	0.2899	etre, temps, puis, poste, pis, quelques, voir,...	peu importe la mode tout finit par puer la vie...
31	31	33	0.3522	nuit, yeux, corps, jour, ciel, mots, temps, vo...	Aullène, au fin fond de l'Ohio, le froid et ...
32	32	0	0.0250	roman, dont, histoire, auteur, œuvre, livre, l...	
33	33	39	0.3190	monde, vie, homme, amour, mort, etre, temps, c...	les grands mots ont aveuglé le monde ce petit ...
34	34	0	0.0250	roman, dont, histoire, auteur, œuvre, livre, l...	
35	35	25	0.3712	etre, temps, puis, poste, pis, quelques, voir,...	sous un ciel sans fin je rêvasse comme un idi...
36	36	33	0.4347	nuit, yeux, corps, jour, ciel, mots, temps, vo...	l'automne éclate comme une guerre et je récup...
37	37	0	0.0250	roman, dont, histoire, auteur, œuvre, livre, l...	
38	38	15	0.3108	etre, dire, chose, faut, reponse, toujours, au...	la ville entière est comme çaun abri impossib...
39	39	0	0.0250	roman, dont, histoire, auteur, œuvre, livre, l...	
40	40	33	0.3816	nuit, yeux, corps, jour, ciel, mots, temps, vo...	ciel gris mer mauvedes choses étranges arrive...
41	41	0	0.0250	roman, dont, histoire, auteur, œuvre, livre, l...	
42	42	32	0.1925	dit, ans, pere, mere, fille, vie, aime, sais, ...	un boulot normalj'ai essayéune femme des enfan...
43	43	15	0.2421	etre, dire, chose, faut, reponse, toujours, au...	tu es sympa mais tu es hors-sujet tu es sym...
44	44	15	0.4053	etre, dire, chose, faut, reponse, toujours, au...	chômeur parasite probablement pédé la vente ch...
45	45	0	0.0250	roman, dont, histoire, auteur, œuvre, livre, l...	
46	46	33	0.3201	nuit, yeux, corps, jour, ciel, mots, temps, vo...	l'automne fait du sacré bon boulot le monde es...
47	47	32	0.2763	dit, ans, pere, mere, fille, vie, aime, sais, ...	mauvais présages que les étoiles filantes dans...
48	48	15	0.2716	etre, dire, chose, faut, reponse, toujours, au...	je suis le veninje suis la bonne nouvelle je s...
49	49	15	0.4426	etre, dire, chose, faut, reponse, toujours, au...	les gens changent même si les histoires resten...

In [26]:
```python
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[26]:

	Topic_Num	Topic_Perc_Contrib	Keywords	Text
0	0	0.9607	roman, dont, histoire, auteur, œuvre, livre, l...	c'est le soir - je retire le texte écrit ce ma...
1	1	0.6750	photo, ile, mer, port, cm, avion, matricule, ...	photo F. Laborde
2	2	0.8376	sexe, masculin, animal, croire, lucie, dune, ...	A trop grave, manquer les heurs.A trop léger, ...
3	3	0.8966	anna, victor, hugo, didier, robin, nicolas, mo...	02 juillet 2020. 05 juillet 2020...
4	4	0.8205	blog, da, blogspot, blogs, jose, ana, didier, ...	Ana pourquoi tu t 'appelles Angeles Ana Angele...
5	5	0.8478	berri, mort, uqam, tait, moscou, connaissez, e...	interruption au milieu - reste l'oiseau bec su...
6	6	0.7375	police, https, facebook, vote, people, org, ju...	Credit: NASA, ESA, F. Paresce (INAF-IASF, Bolo...
7	7	0.8124	pays, france, contre, guerre, francais, politi...	J'ai été arrêté à seulement 807 années de pr...
8	8	0.6796	pen, chiffres, ira, ales, tourne, figaro, port...	un homme marche au bord du chemin blanc , s'ar...
9	9	0.9639	pain, table, manger, cuisine, mange, huile, so...	A cassé B cassé C cassé D cassé E cassé F cass...
10	10	0.6015	clown, flash, cirque, cherie, verres, jeff, ji...	Un clown de ciel grisun clown assassinun clown...
11	11	0.6750	sherbrooke, paade, attila, amical, prisonniers...	Aidez-le, de grâce.
12	12	0.9994	michel, menard, web, cliquant, elise, sanchez,...	Un blog rétroublié, des triptyques qui s'épui...
13	13	0.8050	po, kevin, dormir, oeil, bunker, conserver, fe...	sortir de là tout pixavant la prostration à c...
14	14	0.9652	tete, vieux, dit, main, bois, jusqu, pieds, pe...	Je n'ai jamais été drôle, c'est ainsi, le comi...
15	15	0.9557	etre, dire, chose, faut, reponse, toujours, au...	http://www.youtube.com/watch?feature=endscreen...
16	16	0.9965	terre, espace, lumiere, eau, feu, ciel, ombre,...	Combien de fois faut-il que je vous le dise, m...
17	17	0.9419	mas, dias, annie, claro, jourde, ups, video, v...	noche noche noche noche noche noche noche noc...
18	18	0.8800	new, york, billy, john, albin, richard, court,...	SINEAD o' CONNOR et cette version by TANYA TU...
19	19	0.8781	ville, rue, paris, poste, saint, prenoms, cart...	MERCREDI 30 SEPTEMBRE 2015à 20hau CENTRE WALLO...
20	20	0.6750	etes, like, vote, monsieur, don, zero, color, ...	(pour Dan Fante)
21	21	0.8375	alberto, singe, pp, ronds, fabien, carlo, big...	by Neil Krug
22	22	0.8375	flaubert, zoe, ronda, jen, helene, marde, sou,...	Clochard céleste, portrait 7, Miroslav tichy s...
23	23	0.8344	musique, chanson, album, chansons, rock, piano...	Il y a soit un fleuve voix soit un espace voix...
24	24	0.6750	film, cinema, films, voir, prix, eme, mc, band...	Il y auront de faucon voir les chosesque de...
25	25	0.9729	etre, temps, puis, poste, pis, quelques, voir,...	*Après avoir longuement hésité, vous avez opté...
26	26	0.8823	jean, pierre, marie, dominique, francois, pasc...	J'ai le plaisir et le privilège d'être en comp...
27	27	0.7562	information, mme, anti, help, bilan, pression,...	La nouvelle Créée est arrivée
28	28	0.9114	livre, commentaire, lire, livres, texte, lectu...	Et cette fois-ci c'est sur La Cause Littéraire...
29	29	0.9652	cartes, japonais, usa, circuit, postales, japo...	A Ray of LightA Ray of EnergyA Ray of Inspirat...
30	30	0.8374	bin, quidam, medecin, fit, docteur, recre, hop...	Ce nouveau souvenirse sera bientôt plusqu'un m...
31	31	0.8375	black, go, re, change, man, will, can, one, no...	Don't tell anybody but I think I just saw Kilg...
32	32	0.9554	dit, ans, pere, mere, fille, vie, aime, sais, ...	Je voulais me rendre sur le blog 807. J'ai cli...
33	33	0.9983	nuit, yeux, corps, jour, ciel, mots, temps, vo...	la voix devient nuit - la voix devient nuit -...
34	34	0.7562	photos, art, exposition, musee, act, photograp...	vitrine de la galerie mam_rouen_mainteantplus...
35	35	0.5125	edmond, dr, boris, adam, visibilite, chou, inn...	Des robinetsfuientet agacenttes lavabosde...
36	36	0.9025	chat, matin, blanc, rouge, petit, neige, rose...	ciel bleu coupe fleurs ciel bleu coupe fleurs ...
37	37	0.9961	poesie, poete, poeme, poemes, poetes, editions...	he de ver te he de ver te he de ver te he de v...
38	38	0.8607	theatre, saison, scene, spectacle, semaine, sa...	Thème : "Beau_belle"regain du frisible faire "...
39	39	0.9798	monde, vie, homme, amour, mort, etre, temps, c...	à Ana Angeles : vivre vivre vivre vivre vivre ...

In [27]:
```python
# Number of Documents for Each Topic
topic_counts = df_topic_sents_keywords['Dominant_Topic'].value_counts()
topic_counts
```

Out[27]:

```
33    24600
25    17002
0     15230
15    14978
32     6348
14     4681
39     3877
28     3759
7      3719
38     2769
36     2710
16     1795
26     1520
37     1353
31     1336
19     1321
23      559
18      525
6       335
9       329
30      233
34      221
4       220
1       177
24      171
5       156
17      142
2       126
27      104
22      101
13       73
11       66
3        66
29       57
20       53
35       45
21       45
8        36
10       35
12       22
Name: Dominant_Topic, dtype: int64
```

In [ ]:

In [ ]: