LDA QUI FONCTIONNE MAIS AVEC DES MESSAGES D'ERREUR

Dans le domaine du traitement automatique des langues, l'allocation de Dirichlet latente (de l'anglais Latent Dirichlet Allocation) ou LDA est un modèle génératif probabiliste permettant d'expliquer des ensembles d'observations, par le moyen de groupes non observés, eux-mêmes définis par des similarités de données.

In [3]:
```
!pip install --upgrade gensim --user
!pip install pyldavis --user
!pip install python-Levenshtein
```

```
Requirement already satisfied: gensim in c:\users\tomdu\anaconda3\lib\site-packages (4.0.1)
Requirement already satisfied: Cython==0.29.21 in c:\users\tomdu\anaconda3\lib\site-packages (from gensim) (0.29.21)
Requirement already satisfied: scipy>=0.18.1 in c:\users\tomdu\anaconda3\lib\site-packages (from gensim) (1.5.2)
Requirement already satisfied: smart-open>=1.8.1 in c:\users\tomdu\anaconda3\lib\site-packages (from gensim) (3.0.0)
Requirement already satisfied: numpy>=1.11.3 in c:\users\tomdu\appdata\roaming\python\python38\site-packages (from gensim) (1.20.2)
Requirement already satisfied: requests in c:\users\tomdu\anaconda3\lib\site-packages (from smart-open>=1.8.1->gensim) (2.24.0)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in c:\users\tomdu\anaconda3\lib\site-packages (from requests->smart-open>=1.8.1->gensim) (1.25.11)
Requirement already satisfied: chardet<4,>=3.0.2 in c:\users\tomdu\anaconda3\lib\site-packages (from requests->smart-open>=1.8.1->gensim) (3.0.4)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\tomdu\anaconda3\lib\site-packages (from requests->smart-open>=1.8.1->gensim) (2020.6.20)
Requirement already satisfied: idna<3,>=2.5 in c:\users\tomdu\anaconda3\lib\site-packages (from requests->smart-open>=1.8.1->gensim) (2.10)
WARNING: Value for scheme.headers does not match. Please report this to <https://github.com/pypa/pip/issues/9617>
distutils: C:\Users\tomdu\AppData\Roaming\Python\Python38\Include\UNKNOWN
sysconfig: C:\Users\tomdu\AppData\Roaming\Python\Python38\Include
WARNING: Additional context:
user = True
home = None
root = None
prefix = None
WARNING: You are using pip version 21.1; however, version 21.1.3 is available.
You should consider upgrading via the 'c:\users\tomdu\anaconda3\python.exe -m pip install --upgrade pip' command.
Requirement already satisfied: pyldavis in c:\users\tomdu\appdata\roaming\python\python38\site-packages (3.3.1)
Requirement already satisfied: pandas>=1.2.0 in c:\users\tomdu\appdata\roaming\python\python38\site-packages (from pyldavis) (1.2.4)
Requirement already satisfied: gensim in c:\users\tomdu\anaconda3\lib\site-packages (from pyldavis) (4.0.1)
Requirement already satisfied: setuptools in c:\users\tomdu\anaconda3\lib\site-packages (from pyldavis) (50.3.1.post20201107)
Requirement already satisfied: numexpr in c:\users\tomdu\anaconda3\lib\site-packages (from pyldavis) (2.7.1)
Requirement already satisfied: scikit-learn in c:\users\tomdu\anaconda3\lib\site-packages (from pyldavis) (0.23.2)
Requirement already satisfied: funcy in c:\users\tomdu\appdata\roaming\python\python38\site-packages (from pyldavis) (1.15)
Requirement already satisfied: sklearn in c:\users\tomdu\appdata\roaming\python\python38\site-packages (from pyldavis) (0.0)
Requirement already satisfied: joblib in c:\users\tomdu\anaconda3\lib\site-packages (from pyldavis) (0.17.0)
Requirement already satisfied: numpy>=1.20.0 in c:\users\tomdu\appdata\roaming\python\python38\site-packages (from pyldavis) (1.20.2)
Requirement already satisfied: scipy in c:\users\tomdu\anaconda3\lib\site-packages (from pyldavis) (1.5.2)
Requirement already satisfied: jinja2 in c:\users\tomdu\anaconda3\lib\site-packages (from pyldavis) (2.11.2)
Requirement already satisfied: future in c:\users\tomdu\anaconda3\lib\site-packages (from pyldavis) (0.18.2)
Requirement already satisfied: pytz>=2017.3 in c:\users\tomdu\anaconda3\lib\site-packages (from pandas>=1.2.0->pyldavis) (2020.1)
Requirement already satisfied: python-dateutil>=2.7.3 in c:\users\tomdu\anaconda3\lib\site-packages (from pandas>=1.2.0->pyldavis) (2.8.1)
Requirement already satisfied: six>=1.5 in c:\users\tomdu\anaconda3\lib\site-packages (from python-dateutil>=2.7.3->pandas>=1.2.0->pyldavis) (1.15.0)
Requirement already satisfied: smart-open>=1.8.1 in c:\users\tomdu\anaconda3\lib\site-packages (from gensim->pyldavis) (3.0.0)
Requirement already satisfied: Cython==0.29.21 in c:\users\tomdu\anaconda3\lib\site-packages (from gensim->pyldavis) (0.29.21)
Requirement already satisfied: requests in c:\users\tomdu\anaconda3\lib\site-packages (from smart-open>=1.8.1->gensim->pyldavis) (2.24.0)
Requirement already satisfied: MarkupSafe>=0.23 in c:\users\tomdu\anaconda3\lib\site-packages (from jinja2->pyldavis) (1.1.1)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\tomdu\anaconda3\lib\site-packages (from requests->smart-open>=1.8.1->gensim->pyldavis) (2020.6.20)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in c:\users\tomdu\anaconda3\lib\site-packages (from requests->smart-open>=1.8.1->gensim->pyldavis)
(1.25.11)
Requirement already satisfied: idna<3,>=2.5 in c:\users\tomdu\anaconda3\lib\site-packages (from requests->smart-open>=1.8.1->gensim->pyldavis) (2.10)
Requirement already satisfied: chardet<4,>=3.0.2 in c:\users\tomdu\anaconda3\lib\site-packages (from requests->smart-open>=1.8.1->gensim->pyldavis) (3.0.4)
Requirement already satisfied: threadpoolctl>=2.0.0 in c:\users\tomdu\anaconda3\lib\site-packages (from scikit-learn->pyldavis) (2.1.0)
WARNING: Value for scheme.headers does not match. Please report this to <https://github.com/pypa/pip/issues/9617>
distutils: C:\Users\tomdu\AppData\Roaming\Python\Python38\Include\UNKNOWN
sysconfig: C:\Users\tomdu\AppData\Roaming\Python\Python38\Include
WARNING: Additional context:
user = True
home = None
root = None
prefix = None
WARNING: You are using pip version 21.1; however, version 21.1.3 is available.
You should consider upgrading via the 'c:\users\tomdu\anaconda3\python.exe -m pip install --upgrade pip' command.
Requirement already satisfied: python-Levenshtein in c:\users\tomdu\anaconda3\lib\site-packages (0.12.2)
Requirement already satisfied: setuptools in c:\users\tomdu\anaconda3\lib\site-packages (from python-Levenshtein) (50.3.1.post20201107)
WARNING: Value for scheme.headers does not match. Please report this to <https://github.com/pypa/pip/issues/9617>
distutils: c:\users\tomdu\anaconda3\Include\UNKNOWN
sysconfig: c:\users\tomdu\anaconda3\Include
WARNING: Additional context:
user = False
home = None
root = None
prefix = None
WARNING: Value for scheme.headers does not match. Please report this to <https://github.com/pypa/pip/issues/9617>
distutils: c:\users\tomdu\anaconda3\Include\UNKNOWN
sysconfig: c:\users\tomdu\anaconda3\Include
WARNING: Additional context:
user = False
home = None
root = None
prefix = None
WARNING: You are using pip version 21.1; however, version 21.1.3 is available.
You should consider upgrading via the 'c:\users\tomdu\anaconda3\python.exe -m pip install --upgrade pip' command.
```

In [4]:
```
pip install python-Levenshtein
```

```
Requirement already satisfied: python-Levenshtein in c:\users\tomdu\anaconda3\lib\site-packages (0.12.2)
Requirement already satisfied: setuptools in c:\users\tomdu\anaconda3\lib\site-packages (from python-Levenshtein) (50.3.1.post20201107)
Note: you may need to restart the kernel to use updated packages.
WARNING: Value for scheme.headers does not match. Please report this to <https://github.com/pypa/pip/issues/9617>
distutils: C:\Users\tomdu\anaconda3\Include\UNKNOWN
sysconfig: C:\Users\tomdu\anaconda3\Include
WARNING: Additional context:
user = False
home = None
root = None
prefix = None
WARNING: Value for scheme.headers does not match. Please report this to <https://github.com/pypa/pip/issues/9617>
distutils: C:\Users\tomdu\anaconda3\Include\UNKNOWN
sysconfig: C:\Users\tomdu\anaconda3\Include
WARNING: Additional context:
user = False
home = None
root = None
prefix = None
WARNING: You are using pip version 21.1; however, version 21.1.3 is available.
You should consider upgrading via the 'C:\Users\tomdu\anaconda3\python.exe -m pip install --upgrade pip' command.
```

In [5]:
```
! pip install pyLDAvis
```

```
Requirement already satisfied: pyLDAvis in c:\users\tomdu\appdata\roaming\python\python38\site-packages (3.3.1)
Requirement already satisfied: numexpr in c:\users\tomdu\anaconda3\lib\site-packages (from pyLDAvis) (2.7.1)
Requirement already satisfied: pandas>=1.2.0 in c:\users\tomdu\appdata\roaming\python\python38\site-packages (from pyLDAvis) (1.2.4)
Requirement already satisfied: scipy in c:\users\tomdu\anaconda3\lib\site-packages (from pyLDAvis) (1.5.2)
Requirement already satisfied: joblib in c:\users\tomdu\anaconda3\lib\site-packages (from pyLDAvis) (0.17.0)
Requirement already satisfied: scikit-learn in c:\users\tomdu\anaconda3\lib\site-packages (from pyLDAvis) (0.23.2)
Requirement already satisfied: setuptools in c:\users\tomdu\anaconda3\lib\site-packages (from pyLDAvis) (50.3.1.post20201107)
```

```
Requirement already satisfied: numpy>=1.20.0 in c:\users\tomdu\appdata\roaming\python\python38\site-packages (from pyLDAvis) (1.20.2)
Requirement already satisfied: jinja2 in c:\users\tomdu\anaconda3\lib\site-packages (from pyLDAvis) (2.11.2)
Requirement already satisfied: funcy in c:\users\tomdu\appdata\roaming\python\python38\site-packages (from pyLDAvis) (1.15)
Requirement already satisfied: sklearn in c:\users\tomdu\appdata\roaming\python\python38\site-packages (from pyLDAvis) (0.0)
Requirement already satisfied: future in c:\users\tomdu\anaconda3\lib\site-packages (from pyLDAvis) (0.18.2)
Requirement already satisfied: gensim in c:\users\tomdu\anaconda3\lib\site-packages (from pyLDAvis) (4.0.1)
Requirement already satisfied: pytz>=2017.3 in c:\users\tomdu\anaconda3\lib\site-packages (from pandas>=1.2.0->pyLDAvis) (2020.1)
Requirement already satisfied: python-dateutil>=2.7.3 in c:\users\tomdu\anaconda3\lib\site-packages (from pandas>=1.2.0->pyLDAvis) (2.8.1)
Requirement already satisfied: six>=1.5 in c:\users\tomdu\anaconda3\lib\site-packages (from python-dateutil>=2.7.3->pandas>=1.2.0->pyLDAvis) (1.15.0)
Requirement already satisfied: Cython==0.29.21 in c:\users\tomdu\anaconda3\lib\site-packages (from gensim->pyLDAvis) (0.29.21)
Requirement already satisfied: smart-open>=1.8.1 in c:\users\tomdu\anaconda3\lib\site-packages (from gensim->pyLDAvis) (3.0.0)
Requirement already satisfied: requests in c:\users\tomdu\anaconda3\lib\site-packages (from smart-open>=1.8.1->gensim->pyLDAvis) (2.24.0)
Requirement already satisfied: MarkupSafe>=0.23 in c:\users\tomdu\anaconda3\lib\site-packages (from jinja2->pyLDAvis) (1.1.1)
Requirement already satisfied: chardet<4,>=3.0.2 in c:\users\tomdu\anaconda3\lib\site-packages (from requests->smart-open>=1.8.1->gensim->pyLDAvis) (3.0.4)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in c:\users\tomdu\anaconda3\lib\site-packages (from requests->smart-open>=1.8.1->gensim->pyLDAvis)
(1.25.11)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\tomdu\anaconda3\lib\site-packages (from requests->smart-open>=1.8.1->gensim->pyLDAvis) (2020.6.20)
Requirement already satisfied: idna<3,>=2.5 in c:\users\tomdu\anaconda3\lib\site-packages (from requests->smart-open>=1.8.1->gensim->pyLDAvis) (2.10)
Requirement already satisfied: threadpoolctl>=2.0.0 in c:\users\tomdu\anaconda3\lib\site-packages (from scikit-learn->pyLDAvis) (2.1.0)
WARNING: Value for scheme.headers does not match. Please report this to <https://github.com/pypa/pip/issues/9617>
distutils: c:\users\tomdu\anaconda3\Include\UNKNOWN
sysconfig: c:\users\tomdu\anaconda3\Include
WARNING: Additional context:
user = False
home = None
root = None
prefix = None
WARNING: Value for scheme.headers does not match. Please report this to <https://github.com/pypa/pip/issues/9617>
distutils: c:\users\tomdu\anaconda3\Include\UNKNOWN
sysconfig: c:\users\tomdu\anaconda3\Include
WARNING: Additional context:
user = False
home = None
root = None
prefix = None
WARNING: You are using pip version 21.1; however, version 21.1.3 is available.
You should consider upgrading via the 'c:\users\tomdu\anaconda3\python.exe -m pip install --upgrade pip' command.
```

In [6]:
```
import gensim as gs
import pandas as pd
import pyLDAvis
pyLDAvis.enable_notebook()
```

```
C:\Users\tomdu\anaconda3\lib\site-packages\sklearn\linear_model\_least_angle.py:34: DeprecationWarning: `np.float` is a deprecated alias for the builtin `float`. To sile
nce this warning, use `float` by itself. Doing this will not modify any behavior and is safe. If you specifically wanted the numpy scalar type, use `np.float64` here.
Deprecated in NumPy 1.20; for more details and guidance: https://numpy.org/devdocs/release/1.20.0-notes.html#deprecations
  method='lar', copy_X=True, eps=np.finfo(np.float).eps,
C:\Users\tomdu\anaconda3\lib\site-packages\sklearn\linear_model\_least_angle.py:164: DeprecationWarning: `np.float` is a deprecated alias for the builtin `float`. To sil
ence this warning, use `float` by itself. Doing this will not modify any behavior and is safe. If you specifically wanted the numpy scalar type, use `np.float64` here.
Deprecated in NumPy 1.20; for more details and guidance: https://numpy.org/devdocs/release/1.20.0-notes.html#deprecations
  method='lar', copy_X=True, eps=np.finfo(np.float).eps,
C:\Users\tomdu\anaconda3\lib\site-packages\sklearn\linear_model\_least_angle.py:281: DeprecationWarning: `np.float` is a deprecated alias for the builtin `float`. To sil
ence this warning, use `float` by itself. Doing this will not modify any behavior and is safe. If you specifically wanted the numpy scalar type, use `np.float64` here.
Deprecated in NumPy 1.20; for more details and guidance: https://numpy.org/devdocs/release/1.20.0-notes.html#deprecations
  eps=np.finfo(np.float).eps, copy_Gram=True, verbose=0,
C:\Users\tomdu\anaconda3\lib\site-packages\sklearn\linear_model\_least_angle.py:865: DeprecationWarning: `np.float` is a deprecated alias for the builtin `float`. To sil
ence this warning, use `float` by itself. Doing this will not modify any behavior and is safe. If you specifically wanted the numpy scalar type, use `np.float64` here.
Deprecated in NumPy 1.20; for more details and guidance: https://numpy.org/devdocs/release/1.20.0-notes.html#deprecations
  eps=np.finfo(np.float).eps, copy_X=True, fit_path=True,
C:\Users\tomdu\anaconda3\lib\site-packages\sklearn\linear_model\_least_angle.py:1121: DeprecationWarning: `np.float` is a deprecated alias for the builtin `float`. To si
lence this warning, use `float` by itself. Doing this will not modify any behavior and is safe. If you specifically wanted the numpy scalar type, use `np.float64` here.
Deprecated in NumPy 1.20; for more details and guidance: https://numpy.org/devdocs/release/1.20.0-notes.html#deprecations
  eps=np.finfo(np.float).eps, copy_X=True, fit_path=True,
C:\Users\tomdu\anaconda3\lib\site-packages\sklearn\linear_model\_least_angle.py:1149: DeprecationWarning: `np.float` is a deprecated alias for the builtin `float`. To si
lence this warning, use `float` by itself. Doing this will not modify any behavior and is safe. If you specifically wanted the numpy scalar type, use `np.float64` here.
Deprecated in NumPy 1.20; for more details and guidance: https://numpy.org/devdocs/release/1.20.0-notes.html#deprecations
  eps=np.finfo(np.float).eps, positive=False):
C:\Users\tomdu\anaconda3\lib\site-packages\sklearn\linear_model\_least_angle.py:1379: DeprecationWarning: `np.float` is a deprecated alias for the builtin `float`. To si
lence this warning, use `float` by itself. Doing this will not modify any behavior and is safe. If you specifically wanted the numpy scalar type, use `np.float64` here.
Deprecated in NumPy 1.20; for more details and guidance: https://numpy.org/devdocs/release/1.20.0-notes.html#deprecations
  max_n_alphas=1000, n_jobs=None, eps=np.finfo(np.float).eps,
C:\Users\tomdu\anaconda3\lib\site-packages\sklearn\linear_model\_least_angle.py:1621: DeprecationWarning: `np.float` is a deprecated alias for the builtin `float`. To si
lence this warning, use `float` by itself. Doing this will not modify any behavior and is safe. If you specifically wanted the numpy scalar type, use `np.float64` here.
Deprecated in NumPy 1.20; for more details and guidance: https://numpy.org/devdocs/release/1.20.0-notes.html#deprecations
  max_n_alphas=1000, n_jobs=None, eps=np.finfo(np.float).eps,
C:\Users\tomdu\anaconda3\lib\site-packages\sklearn\linear_model\_least_angle.py:1755: DeprecationWarning: `np.float` is a deprecated alias for the builtin `float`. To si
lence this warning, use `float` by itself. Doing this will not modify any behavior and is safe. If you specifically wanted the numpy scalar type, use `np.float64` here.
Deprecated in NumPy 1.20; for more details and guidance: https://numpy.org/devdocs/release/1.20.0-notes.html#deprecations
  eps=np.finfo(np.float).eps, copy_X=True, positive=False):
C:\Users\tomdu\anaconda3\lib\site-packages\sklearn\decomposition\_lda.py:28: DeprecationWarning: `np.float` is a deprecated alias for the builtin `float`. To silence thi
s warning, use `float` by itself. Doing this will not modify any behavior and is safe. If you specifically wanted the numpy scalar type, use `np.float64` here.
Deprecated in NumPy 1.20; for more details and guidance: https://numpy.org/devdocs/release/1.20.0-notes.html#deprecations
  EPS = np.finfo(np.float).eps
```

In [7]:
```
dataset = pd.read_csv("insurance_QA_train.csv", sep=";")
```

```
C:\Users\tomdu\anaconda3\lib\site-packages\ipykernel\ipkernel.py:287: DeprecationWarning: `should_run_async` will not call `transform_cell` automatically in the future.
Please pass the result to `transformed_cell` argument and any exception that happen during thetransform in `preprocessing_exc_tuple` in IPython 7.17 and above.
  and should_run_async(code)
```

In [8]:
```
dataset.head()
```

```
C:\Users\tomdu\anaconda3\lib\site-packages\ipykernel\ipkernel.py:287: DeprecationWarning: `should_run_async` will not call `transform_cell` automatically in the future.
Please pass the result to `transformed_cell` argument and any exception that happen during thetransform in `preprocessing_exc_tuple` in IPython 7.17 and above.
  and should_run_async(code)
```

Out[8]:

	Question	Answer
0	be Life Insurance exempt from creditor in Conn...	provided that you have name a primarybeneficia...
1	which country have the good retirement plan?	not many sovereign nation have retirement plan...
2	how do I purchase homeowner insurance?	it simple really contact a broker like myself ...
3	when be Medicare primary or secondary?	the factor that determine whether Medicare be ...
4	do smoking void Life Insurance?	if you buy a life insurance policy as a nonsmo...

In [9]:
```
docs = list(dataset["Question"] + dataset["Answer"])
```

```
C:\Users\tomdu\anaconda3\lib\site-packages\ipykernel\ipkernel.py:287: DeprecationWarning: `should_run_async` will not call `transform_cell` automatically in the future.
Please pass the result to `transformed_cell` argument and any exception that happen during thetransform in `preprocessing_exc_tuple` in IPython 7.17 and above.
  and should_run_async(code)
```

In [10]:
```
# Tokenize the documents.
from nltk.tokenize import RegexpTokenizer

# Split the documents into tokens.
tokenizer = RegexpTokenizer(r'\w+')
for idx in range(len(docs)):
    docs[idx] = docs[idx].lower()  # Convert to lowercase.
    docs[idx] = tokenizer.tokenize(docs[idx])  # Split into words.
```

```
# Remove numbers, but not words that contain numbers.
docs = [[token for token in doc if not token.isnumeric()] for doc in docs]

# Remove words that are only one character.
docs = [[token for token in doc if len(token) > 1] for doc in docs]
```

```
C:\Users\tomdu\anaconda3\lib\site-packages\ipykernel\ipkernel.py:287: DeprecationWarning: `should_run_async` will not call `transform_cell` automatically in the future.
Please pass the result to `transformed_cell` argument and any exception that happen during thetransform in `preprocessing_exc_tuple` in IPython 7.17 and above.
  and should_run_async(code)
C:\Users\tomdu\anaconda3\lib\site-packages\sklearn\feature_extraction\image.py:172: DeprecationWarning: `np.int` is a deprecated alias for the builtin `int`. To silence
this warning, use `int` by itself. Doing this will not modify any behavior and is safe. When replacing `np.int`, you may wish to use e.g. `np.int64` or `np.int32` to spe
cify the precision. If you wish to review your current use, check the release note link for additional information.
Deprecated in NumPy 1.20; for more details and guidance: https://numpy.org/devdocs/release/1.20.0-notes.html#deprecations
  dtype=np.int):
```

In [11]:
```
docs[97]
```

```
C:\Users\tomdu\anaconda3\lib\site-packages\ipykernel\ipkernel.py:287: DeprecationWarning: `should_run_async` will not call `transform_cell` automatically in the future.
Please pass the result to `transformed_cell` argument and any exception that happen during thetransform in `preprocessing_exc_tuple` in IPython 7.17 and above.
  and should_run_async(code)
```

Out[11]:
```
['how',
 'determine',
 'term',
 'life',
 'insurance',
 'amount',
 'many',
 'life',
 'insurance',
 'expert',
 'recommend',
 'about',
 'to',
 'time',
 'your',
 'annual',
 'salary',
 'for',
 'the',
 'amount',
 'of',
 'term',
 'life',
 'insurance',
 'that',
 'you',
 'may',
 'need',
 'however',
 'the',
 'figure',
 'will',
 'depend',
 'on',
 'your',
 'financial',
 'need',
 'and',
 'goal',
 'you',
 'have',
 'factor',
 'in',
 'that',
 'mortgage',
 'education',
 'expense',
 'for',
 'child',
 'basically',
 'whatever',
 'your',
 'family',
 'will',
 'need',
 'survive',
 'and',
 'thrive',
 'when',
 'you',
 'be',
 'gone',
 'of',
 'course',
 'since',
 'term',
 'life',
 'insurance',
 'be',
 'for',
 'specific',
 'period',
 'of',
 'time',
 'you',
 'will',
 'need',
 'calculate',
 'how',
 'long',
 'you',
 'need',
 'this',
 'certain',
 'amount',
 'of',
 'money',
 'as',
 'well']
```

In [12]:
```
# Remove words that are only one character.
docs = [[token for token in doc if len(token) > 3] for doc in docs]
```

```
C:\Users\tomdu\anaconda3\lib\site-packages\ipykernel\ipkernel.py:287: DeprecationWarning: `should_run_async` will not call `transform_cell` automatically in the future.
Please pass the result to `transformed_cell` argument and any exception that happen during thetransform in `preprocessing_exc_tuple` in IPython 7.17 and above.
  and should_run_async(code)
```

In [13]:
```
docs[97]
```

```
C:\Users\tomdu\anaconda3\lib\site-packages\ipykernel\ipkernel.py:287: DeprecationWarning: `should_run_async` will not call `transform_cell` automatically in the future.
Please pass the result to `transformed_cell` argument and any exception that happen during thetransform in `preprocessing_exc_tuple` in IPython 7.17 and above.
```

```
  and should_run_async(code)
```

Out[13]:
```
['determine',
 'term',
 'life',
 'insurance',
 'amount',
 'many',
 'life',
 'insurance',
 'expert',
 'recommend',
 'about',
 'time',
 'your',
 'annual',
 'salary',
 'amount',
 'term',
 'life',
 'insurance',
 'that',
 'need',
 'however',
 'figure',
 'will',
 'depend',
 'your',
 'financial',
 'need',
 'goal',
 'have',
 'factor',
 'that',
 'mortgage',
 'education',
 'expense',
 'child',
 'basically',
 'whatever',
 'your',
 'family',
 'will',
 'need',
 'survive',
 'thrive',
 'when',
 'gone',
 'course',
 'since',
 'term',
 'life',
 'insurance',
 'specific',
 'period',
 'time',
 'will',
 'need',
 'calculate',
 'long',
 'need',
 'this',
 'certain',
 'amount',
 'money',
 'well']
```

In [14]:
```
#!pip install --user -U nltk
import nltk
nltk.download()
```

```
C:\Users\tomdu\anaconda3\lib\site-packages\ipykernel\ipkernel.py:287: DeprecationWarning: `should_run_async` will not call `transform_cell` automatically in the future.
Please pass the result to `transformed_cell` argument and any exception that happen during thetransform in `preprocessing_exc_tuple` in IPython 7.17 and above.
  and should_run_async(code)
showing info https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/index.xml
```

Out[14]:
```
True
```

In [15]:
```
# Lemmatize the documents.
from nltk.stem.wordnet import WordNetLemmatizer

lemmatizer = WordNetLemmatizer()
docs = [[lemmatizer.lemmatize(token) for token in doc] for doc in docs]
```

```
C:\Users\tomdu\anaconda3\lib\site-packages\ipykernel\ipkernel.py:287: DeprecationWarning: `should_run_async` will not call `transform_cell` automatically in the future.
Please pass the result to `transformed_cell` argument and any exception that happen during thetransform in `preprocessing_exc_tuple` in IPython 7.17 and above.
  and should_run_async(code)
```

In [16]:
```
# Compute bigrams.
from gensim.models import Phrases

# Add bigrams and trigrams to docs (only ones that appear 20 times or more).
bigram = Phrases(docs, min_count=20)
for idx in range(len(docs)):
    for token in bigram[docs[idx]]:
        if '_' in token:
            # Token is a bigram, add to document.
            docs[idx].append(token)
```

```
C:\Users\tomdu\anaconda3\lib\site-packages\ipykernel\ipkernel.py:287: DeprecationWarning: `should_run_async` will not call `transform_cell` automatically in the future.
Please pass the result to `transformed_cell` argument and any exception that happen during thetransform in `preprocessing_exc_tuple` in IPython 7.17 and above.
  and should_run_async(code)
```

In [17]:
```
# Remove rare and common tokens.
from gensim.corpora import Dictionary

# Create a dictionary representation of the documents.
dictionary = Dictionary(docs)

# Filter out words that occur less than 20 documents, or more than 50% of the documents.
dictionary.filter_extremes(no_below=20, no_above=0.4)
```

```
C:\Users\tomdu\anaconda3\lib\site-packages\ipykernel\ipkernel.py:287: DeprecationWarning: `should_run_async` will not call `transform_cell` automatically in the future.
Please pass the result to `transformed_cell` argument and any exception that happen during thetransform in `preprocessing_exc_tuple` in IPython 7.17 and above.
  and should_run_async(code)
```

In [18]:
```
# Bag-of-words representation of the documents.
corpus = [dictionary.doc2bow(doc) for doc in docs]
```

```
C:\Users\tomdu\anaconda3\lib\site-packages\ipykernel\ipkernel.py:287: DeprecationWarning: `should_run_async` will not call `transform_cell` automatically in the future.
Please pass the result to `transformed_cell` argument and any exception that happen during thetransform in `preprocessing_exc_tuple` in IPython 7.17 and above.
  and should_run_async(code)
```

In [19]:
```
print('Number of unique tokens: %d' % len(dictionary))
print('Number of documents: %d' % len(corpus))
```

```
Number of unique tokens: 2542
Number of documents: 12887
```

```
C:\Users\tomdu\anaconda3\lib\site-packages\ipykernel\ipkernel.py:287: DeprecationWarning: `should_run_async` will not call `transform_cell` automatically in the future.
Please pass the result to `transformed_cell` argument and any exception that happen during thetransform in `preprocessing_exc_tuple` in IPython 7.17 and above.
  and should_run_async(code)
```

In [20]:
```
# Train LDA model.
from gensim.models import LdaModel

# Set training parameters.
num_topics = 5
chunksize = 200
passes = 2
iterations = 10
eval_every = None  # Don't evaluate model perplexity, takes too much time.

# Make a index to word dictionary.
temp = dictionary[0]  # This is only to "load" the dictionary.
id2word = dictionary.id2token

lda_model = LdaModel(
    corpus=corpus,
    id2word=id2word,
    chunksize=chunksize,
    alpha='auto',
    eta='auto',
    iterations=iterations,
    num_topics=num_topics,
    passes=passes,
    eval_every=eval_every
)
```

```
C:\Users\tomdu\anaconda3\lib\site-packages\ipykernel\ipkernel.py:287: DeprecationWarning: `should_run_async` will not call `transform_cell` automatically in the future.
Please pass the result to `transformed_cell` argument and any exception that happen during thetransform in `preprocessing_exc_tuple` in IPython 7.17 and above.
  and should_run_async(code)
```

In [21]:
```
top_topics = lda_model.top_topics(corpus) #, num_words=20)

# Average topic coherence is the sum of topic coherences of all topics, divided by the number of topics.
avg_topic_coherence = sum([t[1] for t in top_topics]) / num_topics
print('Average topic coherence: %.4f.' % avg_topic_coherence)

from pprint import pprint
pprint(top_topics)
```

```
C:\Users\tomdu\anaconda3\lib\site-packages\ipykernel\ipkernel.py:287: DeprecationWarning: `should_run_async` will not call `transform_cell` automatically in the future.
Please pass the result to `transformed_cell` argument and any exception that happen during thetransform in `preprocessing_exc_tuple` in IPython 7.17 and above.
  and should_run_async(code)
Average topic coherence: -1.9932.
[([(0.12614127, 'life'),
   (0.01966556, 'term'),
   (0.019394673, 'cash'),
   (0.01894011, 'value'),
   (0.018666994, 'death'),
   (0.012816806, 'whole'),
   (0.012449187, 'beneficiary'),
   (0.011088992, 'cash_value'),
   (0.01017957, 'whole_life'),
   (0.01004031, 'when'),
   (0.009540585, 'benefit'),
   (0.009373764, 'premium'),
   (0.009226393, 'what'),
   (0.00883872, 'amount'),
   (0.008817394, 'take'),
   (0.008777546, 'year'),
   (0.008550702, 'loan'),
   (0.008188961, 'from'),
   (0.0081020035, 'insured'),
   (0.007476705, 'need')],
  -1.6102195433079032),
 ([(0.025017295, 'company'),
   (0.015590106, 'good'),
   (0.014857169, 'coverage'),
   (0.013260337, 'they'),
   (0.012420598, 'agent'),
   (0.011980588, 'there'),
   (0.011702982, 'this'),
   (0.010669974, 'what'),
   (0.010209811, 'rate'),
   (0.009815579, 'state'),
   (0.0086553935, 'need'),
   (0.008586847, 'carrier'),
   (0.008339755, 'claim'),
   (0.008303673, 'more'),
   (0.007927701, 'cost'),
   (0.007514473, 'auto'),
   (0.007502808, 'many'),
   (0.007462304, 'question'),
   (0.0068000844, 'find'),
   (0.0067807827, 'their')],
  -1.6258585782660822),
 ([(0.047499254, 'health'),
   (0.02884378, 'care'),
   (0.026057137, 'disability'),
   (0.026022209, 'term'),
   (0.023791455, 'long'),
   (0.018547922, 'long_term'),
   (0.01648582, 'benefit'),
   (0.011322177, 'period'),
   (0.011092854, 'plan'),
   (0.01088838, 'group'),
   (0.010788704, 'premium'),
   (0.010006704, 'their'),
   (0.009104127, 'year'),
   (0.009025803, 'individual'),
   (0.0087220045, 'qualify'),
   (0.008701975, 'employer'),
   (0.008677473, 'when'),
   (0.008602606, 'coverage'),
   (0.008316472, 'employee'),
   (0.0076151066, 'income')],
  -1.7598578763294337),
 ([(0.061937653, 'medicare'),
   (0.05254399, 'cover'),
   (0.025860807, 'coverage'),
   (0.024658907, 'part'),
   (0.019144941, 'renter'),
   (0.01591726, 'damage'),
```

```
  (0.014648288, 'property'),
  (0.014119978, 'homeowner'),
  (0.013676797, 'home'),
  (0.012480358, 'deductible'),
  (0.0123406835, 'plan'),
  (0.010949498, 'medicare_part'),
  (0.008975725, 'personal'),
  (0.008937432, 'provide'),
  (0.0081316475, 'loss'),
  (0.007397514, 'there'),
  (0.007185958, 'most'),
  (0.007157672, 'under'),
  (0.006917353, 'from'),
  (0.006889715, 'other')],
 -2.3022934185636736),
([(0.06392838, 'plan'),
  (0.032373574, 'annuity'),
  (0.023428453, 'retirement'),
  (0.013514823, 'medigap'),
  (0.010457887, 'income'),
  (0.008814811, 'please'),
  (0.0083389, 'what'),
  (0.008196491, 'product'),
  (0.007849567, 'medigap_plan'),
  (0.0076112812, 'free'),
  (0.0073126797, 'feel'),
  (0.007276072, 'retirement_plan'),
  (0.0068751574, 'question'),
  (0.0065418575, 'benefit'),
  (0.006433936, 'variable'),
  (0.0063957674, 'rate'),
  (0.0063319774, 'account'),
  (0.006290932, 'money'),
  (0.0061202995, 'from'),
  (0.0059608086, 'there')],
 -2.667890566273875)]
```

In [22]:
```
!pip install pyLDAvis.gensim --user
```

```
C:\Users\tomdu\anaconda3\lib\site-packages\ipykernel\ipkernel.py:287: DeprecationWarning: `should_run_async` will not call `transform_cell` automatically in the future.
Please pass the result to `transformed_cell` argument and any exception that happen during thetransform in `preprocessing_exc_tuple` in IPython 7.17 and above.
  and should_run_async(code)
ERROR: Could not find a version that satisfies the requirement pyLDAvis.gensim (from versions: none)
ERROR: No matching distribution found for pyLDAvis.gensim
WARNING: You are using pip version 21.1; however, version 21.1.3 is available.
You should consider upgrading via the 'c:\users\tomdu\anaconda3\python.exe -m pip install --upgrade pip' command.
```

In [23]:
```
import pyLDAvis.gensim_models
```

```
C:\Users\tomdu\anaconda3\lib\site-packages\ipykernel\ipkernel.py:287: DeprecationWarning: `should_run_async` will not call `transform_cell` automatically in the future.
Please pass the result to `transformed_cell` argument and any exception that happen during thetransform in `preprocessing_exc_tuple` in IPython 7.17 and above.
  and should_run_async(code)
```

In [24]:
```
vis = pyLDAvis.gensim_models.prepare(lda_model, corpus, dictionary)
```

```
C:\Users\tomdu\anaconda3\lib\site-packages\ipykernel\ipkernel.py:287: DeprecationWarning: `should_run_async` will not call `transform_cell` automatically in the future.
Please pass the result to `transformed_cell` argument and any exception that happen during thetransform in `preprocessing_exc_tuple` in IPython 7.17 and above.
  and should_run_async(code)
```

In [25]:
```
vis
```

```
C:\Users\tomdu\anaconda3\lib\site-packages\ipykernel\ipkernel.py:287: DeprecationWarning: `should_run_async` will not call `transform_cell` automatically in the future.
Please pass the result to `transformed_cell` argument and any exception that happen during thetransform in `preprocessing_exc_tuple` in IPython 7.17 and above.
  and should_run_async(code)
```

Out[25]: