*You are currently looking at **version 1.0** of this notebook. To download notebooks and datafiles, as well as get help on Jupyter notebooks in the Coursera platform, visit the Jupyter Notebook FAQ (https://www.coursera.org/learn/python-text-mining/resources/d9pwm) course resource.*

Assignment 4 - Document Similarity & Topic Modelling

Part 1 - Document Similarity

For the first part of this assignment, you will complete the functions `doc_to_synsets` and `similarity_score` which will be used by `document_path_similarity` to find the path similarity between two documents.

The following functions are provided:

- **`convert_tag:`** converts the tag given by `nltk.pos_tag` to a tag used by `wordnet.synsets`. You will need to use this function in `doc_to_synsets`.
- **`document_path_similarity:`** computes the symmetrical path similarity between two documents by finding the synsets in each document using `doc_to_synsets`, then computing similarities using `similarity_score`.

You will need to finish writing the following functions:

- **`doc_to_synsets:`** returns a list of synsets in document. This function should first tokenize and part of speech tag the document using `nltk.word_tokenize` and `nltk.pos_tag`. Then it should find each tokens corresponding synset using `wn.synsets(token, wordnet_tag)`. The first synset match should be used. If there is no match, that token is skipped.
- **`similarity_score:`** returns the normalized similarity score of a list of synsets (s1) onto a second list of synsets (s2). For each synset in s1, find the synset in s2 with the largest similarity value. Sum all of the largest similarity values together and normalize this value by dividing it by the number of largest similarity values found. Be careful with data types, which should be floats. Missing values should be ignored.

Once `doc_to_synsets` and `similarity_score` have been completed, submit to the autograder which will run `test_document_path_similarity` to test that these functions are running correctly.

Do not modify the functions `convert_tag`, `document_path_similarity`, and `test_document_path_similarity`.

In [1]:
```
import numpy as np
import nltk
from nltk.corpus import wordnet as wn
import pandas as pd

nltk.download('punkt')
nltk.download('averaged_perceptron_tagger')
nltk.download('wordnet')
```

```
[nltk_data] Downloading package punkt to /home/jovyan/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     /home/jovyan/nltk_data...
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
[nltk_data] Downloading package wordnet to /home/jovyan/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```

Out[1]: True

In [2]:
```
def convert_tag(tag):
    """Convert the tag given by nltk.pos_tag to the tag used by wordnet.synsets"""

    tag_dict = {'N': 'n', 'J': 'a', 'R': 'r', 'V': 'v'}
    try:
        return tag_dict[tag[0]]
    except KeyError:
        return None

def doc_to_synsets(doc):
    """
    Returns a list of synsets in document.

    Tokenizes and tags the words in the document doc.
    Then finds the first synset for each word/tag combination.
    If a synset is not found for that combination it is skipped.

    Args:
        doc: string to be converted

    Returns:
        list of synsets

    Example:
        doc_to_synsets('Fish are nvqjp friends.')
        Out: [Synset('fish.n.01'), Synset('be.v.01'), Synset('friend.n.01')]
    """

    # Your Code Here

    tokenized_doc = nltk.word_tokenize(doc)

    tagged_doc = nltk.pos_tag(tokenized_doc)

    wordnet_tags = [convert_tag(tag[1]) for tag in tagged_doc]

    synsets = [wn.synsets(token, tag)[0] for token,
               tag in zip(tokenized_doc, wordnet_tags) if len(wn.synsets(token, tag)) > 0]

    return synsets

def similarity_score(s1, s2):
    """
    Calculate the normalized similarity score of s1 onto s2

    For each synset in s1, finds the synset in s2 with the largest similarity value.
    Sum of all of the largest similarity values and normalize this value by dividing it by the
    number of largest similarity values found.

    Args:
        s1, s2: list of synsets from doc_to_synsets

    Returns:
        normalized similarity score of s1 onto s2

    Example:
        synsets1 = doc_to_synsets('I like cats')
        synsets2 = doc_to_synsets('I like dogs')
        similarity_score(synsets1, synsets2)
        Out: 0.73333333333333339
    """

    # Your Code Here

    max_similarity= []

    for synsets in s1:
        sim = [synsets.path_similarity(x) for x in s2 if synsets.path_similarity(x) is not None]

        if sim:
            max_similarity.append(max(sim))

    return np.mean(max_similarity)
```

```
def document_path_similarity(doc1, doc2):
    """Finds the symmetrical similarity between doc1 and doc2"""

    synsets1 = doc_to_synsets(doc1)
    synsets2 = doc_to_synsets(doc2)

    return (similarity_score(synsets1, synsets2) + similarity_score(synsets2, synsets1)) / 2
```

test_document_path_similarity

Use this function to check if doc_to_synsets and similarity_score are correct.

This function should return the similarity score as a float.

In [3]:
```
def test_document_path_similarity():
    doc1 = 'This is a function to test document_path_similarity.'
    doc2 = 'Use this function to see if your code in doc_to_synsets \
    and similarity_score is correct!'
    return document_path_similarity(doc1, doc2)

test_document_path_similarity()
```

Out[3]:
```
0.55426587301587305
```

`paraphrases` is a DataFrame which contains the following columns: `Quality`, `D1`, and `D2`.

`Quality` is an indicator variable which indicates if the two documents `D1` and `D2` are paraphrases of one another (1 for paraphrase, 0 for not paraphrase).

In [4]:
```
# Use this dataframe for questions most_similar_docs and label_accuracy
paraphrases = pd.read_csv('paraphrases.csv')
paraphrases.head()
```

Out[4]:

	Quality	D1	D2
0	1	Ms Stewart, the chief executive, was not expec...	Ms Stewart, 61, its chief executive officer an...
1	1	After more than two years' detention under the...	After more than two years in detention by the ...
2	1	"It still remains to be seen whether the reven...	"It remains to be seen whether the revenue rec...
3	0	And it's going to be a wild ride," said Allan ...	Now the rest is just mechanical," said Allan H...
4	1	The cards are issued by Mexico's consulates to...	The card is issued by Mexico's consulates to i...

---

most_similar_docs

Using `document_path_similarity`, find the pair of documents in paraphrases which has the maximum similarity score.

This function should return a tuple `(D1, D2, similarity_score)`

In [5]:
```
def most_similar_docs():

    # Your Code Here

    similarity = np.array([document_path_similarity(row['D1'], row['D2']) for _, row in paraphrases.iterrows
()])

    index = np.nanargmax(similarity)

    similarity_score = np.nanmax(similarity)

    D1 = paraphrases['D1'][index]
    D2 = paraphrases['D2'][index]

    return (D1, D2, similarity_score)

most_similar_docs()
```

Out[5]:
```
('"Indeed, Iran should be put on notice that efforts to try to remake Iraq in their image will be aggressivel
y put down," he said.',
 '"Iran should be on notice that attempts to remake Iraq in Iran\'s image will be aggressively put down," he
said.\n',
 0.97530864197530864)
```

label_accuracy

Provide labels for the twenty pairs of documents by computing the similarity for each pair using `document_path_similarity`. Let the classifier rule be that if the score is greater than 0.75, label is paraphrase (1), else label is not paraphrase (0). Report accuracy of the classifier using scikit-learn's accuracy_score.

This function should return a float.

In [6]:
```
def label_accuracy():
    from sklearn.metrics import accuracy_score

    # Your Code Here

    similarity = [document_path_similarity(row['D1'], row['D2']) for _, row in paraphrases.iterrows()]

    predictions = list(map(lambda x: 1 if x > 0.75 else 0, similarity))

    return accuracy_score(predictions, paraphrases['Quality'])

label_accuracy()
```

Out[6]:
```
0.80000000000000004
```

Part 2 - Topic Modelling

For the second part of this assignment, you will use Gensim's LDA (Latent Dirichlet Allocation) model to model topics in `newsgroup_data`. You will first need to finish the code in the cell below by using gensim.models.ldamodel.LdaModel constructor to estimate LDA model parameters on the corpus, and save to the variable `ldamodel`. Extract 10 topics using `corpus` and `id_map`, and with `passes=25` and `random_state=34`.

In [7]:
```
import pickle
import gensim
from sklearn.feature_extraction.text import CountVectorizer

# Load the list of documents
with open('newsgroups', 'rb') as f:
    newsgroup_data = pickle.load(f)

# Use CountVectorizor to find three letter tokens, remove stop_words,
# remove tokens that don't appear in at least 20 documents,
# remove tokens that appear in more than 20% of the documents
vect = CountVectorizer(min_df=20, max_df=0.2, stop_words='english',
                       token_pattern='(?u)\\b\\w\\w\\w+\\b')
# Fit and transform
X = vect.fit_transform(newsgroup_data)

# Convert sparse matrix to gensim corpus.
corpus = gensim.matutils.Sparse2Corpus(X, documents_columns=False)

# Mapping from word IDs to words (To be used in LdaModel's id2word parameter)
id_map = dict((v, k) for k, v in vect.vocabulary_.items())
```

In [8]:
```
# Use the gensim.models.ldamodel.LdaModel constructor to estimate
# LDA model parameters on the corpus, and save to the variable `ldamodel`

# Your code here:
ldamodel =  gensim.models.ldamodel.LdaModel(corpus, num_topics=10, id2word=id_map, passes=25, random_state=34)
```

lda_topics

Using `ldamodel`, find a list of the 10 topics and the most significant 10 words in each topic. This should be structured as a list of 10 tuples where each tuple takes on the form:

`(9, '0.068*"space" + 0.036*"nasa" + 0.021*"science" + 0.020*"edu" + 0.019*"data" + 0.017*"shuttle" + 0.015*"launch" + 0.015*"available" + 0.014*"center" + 0.014*"sci"')`

for example.

This function should return a list of tuples.

In [9]:
```
def lda_topics():

    # Your Code Here

    return list(ldamodel.show_topics(num_topics=10, num_words=10))

lda_topics()
```

Out[9]:
```
[(0,
  '0.056*"edu" + 0.043*"com" + 0.033*"thanks" + 0.022*"mail" + 0.021*"know" + 0.020*"does" + 0.014*"info" +
0.012*"monitor" + 0.010*"looking" + 0.010*"don"'),
 (1,
  '0.024*"ground" + 0.018*"current" + 0.018*"just" + 0.013*"want" + 0.013*"use" + 0.011*"using" + 0.011*"use
d" + 0.010*"power" + 0.010*"speed" + 0.010*"output"'),
 (2,
  '0.061*"drive" + 0.042*"disk" + 0.033*"scsi" + 0.030*"drives" + 0.028*"hard" + 0.028*"controller" + 0.027
*"card" + 0.020*"rom" + 0.018*"floppy" + 0.017*"bus"'),
 (3,
  '0.023*"time" + 0.015*"atheism" + 0.014*"list" + 0.013*"left" + 0.012*"alt" + 0.012*"faq" + 0.012*"probabl
y" + 0.011*"know" + 0.011*"send" + 0.010*"months"'),
 (4,
  '0.025*"car" + 0.016*"just" + 0.014*"don" + 0.014*"bike" + 0.012*"good" + 0.011*"new" + 0.011*"think" + 0.0
10*"year" + 0.010*"cars" + 0.010*"time"'),
 (5,
  '0.030*"game" + 0.027*"team" + 0.023*"year" + 0.017*"games" + 0.016*"play" + 0.012*"season" + 0.012*"player
s" + 0.012*"win" + 0.011*"hockey" + 0.011*"good"'),
 (6,
  '0.017*"information" + 0.014*"help" + 0.014*"medical" + 0.012*"new" + 0.012*"use" + 0.012*"000" + 0.012*"re
search" + 0.011*"university" + 0.010*"number" + 0.010*"program"'),
 (7,
  '0.022*"don" + 0.021*"people" + 0.018*"think" + 0.017*"just" + 0.012*"say" + 0.011*"know" + 0.011*"does" +
0.011*"good" + 0.010*"god" + 0.009*"way"'),
 (8,
  '0.034*"use" + 0.023*"apple" + 0.020*"power" + 0.016*"time" + 0.015*"data" + 0.015*"software" + 0.012*"pin"
+ 0.012*"memory" + 0.012*"simms" + 0.012*"port"'),
 (9,
  '0.068*"space" + 0.036*"nasa" + 0.021*"science" + 0.020*"edu" + 0.019*"data" + 0.017*"shuttle" + 0.015*"lau
nch" + 0.015*"available" + 0.014*"center" + 0.014*"sci"')]
```

topic_distribution

For the new document `new_doc`, find the topic distribution. Remember to use vect.transform on the the new doc, and Sparse2Corpus to convert the sparse matrix to gensim corpus.

This function should return a list of tuples, where each tuple is `(#topic, probability)`

In [10]:
```
new_doc = ["\n\nIt's my understanding that the freezing will start to occur because \
of the\ngrowing distance of Pluto and Charon from the Sun, due to it's\nelliptical orbit. \
It is not due to shadowing effects. \n\n\nPluto can shadow Charon, and vice-versa.\n\nGeorge \
Krumins\n-- "]
```

In [11]:
```
def topic_distribution():

    # Your Code Here

    Xnew = vect.transform(new_doc)
    corpus_new = gensim.matutils.Sparse2Corpus(Xnew, documents_columns=False)

    return list(ldamodel.get_document_topics(corpus_new))

topic_distribution()
```

Out[11]:
```
[[(0, 0.020001831829864054),
  (1, 0.02000204822465949),
  (2, 0.02000000083212411),
  (3, 0.49631042031583406),
  (4, 0.020002764135450282),
  (5, 0.020002856012202509),
  (6, 0.020001696374813884),
  (7, 0.020001367510038583),
  (8, 0.020001848233010606),
  (9, 0.34367516653200242)]]
```

topic_names

From the list of the following given topics, assign topic names to the topics you found. If none of these names best matches the topics you found, create a new 1-3 word "title" for the topic.

Topics: Health, Science, Automobiles, Politics, Government, Travel, Computers & IT, Sports, Business, Society & Lifestyle, Religion, Education.

This function should return a list of 10 strings.

In [12]:
```
def topic_names():

    # Your Code Here

    topic_names = ['Health', 'Science', 'Automobiles', 'Politics', 'Government', 'Travel', 'Computers & IT', Sp
orts',
                   'Business', 'Society & Lifestyle', 'Religion', 'Education']

    topics = lda_topics()

    results = []

    for _, dist in topics:
        similarity = []

        for topic in topic_names:
            similarity.append(document_path_similarity(dist, topic))

        best_topic = sorted(zip(similarity, topic_names))[-1][1]

        results.append(best_topic)

    return results

topic_names()
```

Out[12]:
```
['Politics',
 'Education',
 'Education',
 'Politics',
 'Automobiles',
 'Education',
 'Education',
 'Society & Lifestyle',
 'Education',
 'Science']
```