In [1]:

```
doc1 = "Sugar is bad to consume. My sister likes to have sugar, but not my father."
doc2 = "My father spends a lot of time driving my sister around to dance practice."
doc3 = "Doctors suggest that driving may cause increased stress and blood pressure."
doc4 = "Sometimes I feel pressure to perform well at school, but my father never seems to
drive my sister to do better."
doc5 = "Health experts say that Sugar is not good for your lifestyle."

doc_complete = [doc1, doc2, doc3, doc4, doc5]
```

In [2]:

```
from nltk.corpus import stopwords
from nltk.stem.wordnet import WordNetLemmatizer
import string
stop = set(stopwords.words('english'))
exclude = set(string.punctuation)
lemma = WordNetLemmatizer()

def clean(doc):
    stop_free = ' '.join([i for i in doc.lower().split() if i not in stop])
    punc_free = ''.join([ch for ch in stop_free if ch not in exclude])
    normalized = ' '.join(lemma.lemmatize(word) for word in punc_free.split())
    return normalized
doc_clean = [clean(doc).split() for doc in doc_complete]
```

In [3]:

```
import gensim
from gensim import corpora
dictionary = corpora.Dictionary(doc_clean)
doc_term_matrix = [dictionary.doc2bow(doc) for doc in doc_clean]
```

In [4]:

```
Lda = gensim.models.ldamodel.LdaModel
ldamodel = Lda(doc_term_matrix, num_topics = 3, id2word = dictionary, passes=50)
```

In [5]:

```
topics=ldamodel.print_topics(num_topics=3, num_words=4)
for topic in topics:
    print(topic)
```

```
(0, '0.050*"driving" + 0.050*"increased" + 0.050*"stress" + 0.050*"cause"')
(1, '0.056*"pressure" + 0.056*"never" + 0.056*"seems" + 0.056*"school"')
(2, '0.085*"sugar" + 0.084*"sister" + 0.084*"father" + 0.048*"lot"')
```

In [ ]:

```

```