In [1]:

```
raw_docs = ["Here are some very simple basic sentences.",
"They won't be very interesting, I'm afraid.",
"The point of these examples is to _learn how basic text cleaning works_ on *very simple*
data."]
```

In [2]:

```
# Tokenizing text into bags of words
from nltk.tokenize import word_tokenize
tokenized_docs = [word_tokenize(doc) for doc in raw_docs]
print(tokenized_docs)
```

```
[['Here', 'are', 'some', 'very', 'simple', 'basic', 'sentences', '.'], ['They', 'wo', "n'
t", 'be', 'very', 'interesting', ',', 'I', "'m", 'afraid', '.'], ['The', 'point', 'of', '
these', 'examples', 'is', 'to', '_learn', 'how', 'basic', 'text', 'cleaning', 'works_', '
on', '*', 'very', 'simple', '*', 'data', '.']]
```

In [3]:

```
# Removing punctuation
import re
import string
regex = re.compile('[%s]' % re.escape(string.punctuation)) #see documentation here: http
://docs.python.org/2/library/string.html

tokenized_docs_no_punctuation = []

for review in tokenized_docs:
    new_review = []
    for token in review:
        new_token = regex.sub(u'', token)
        if not new_token == u'':
            new_review.append(new_token)

    tokenized_docs_no_punctuation.append(new_review)

print(tokenized_docs_no_punctuation)
```

```
[['Here', 'are', 'some', 'very', 'simple', 'basic', 'sentences'], ['They', 'wo', 'nt', 'b
e', 'very', 'interesting', 'I', 'm', 'afraid'], ['The', 'point', 'of', 'these', 'examples
', 'is', 'to', 'learn', 'how', 'basic', 'text', 'cleaning', 'works', 'on', 'very', 'simpl
e', 'data']]
```

In [5]:

```
nltk.download('stopwords')
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-5-e3fc0c9c9a89> in <module>
----> 1 nltk.download('stopwords')

NameError: name 'nltk' is not defined
```

In [6]:

```
import nltk
```

In [7]:

```
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\hp\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping corpora\stopwords.zip.
```

Out[7]:

```
True
```

In [8]:

```
# Cleaning text of stopwords
from nltk.corpus import stopwords

tokenized_docs_no_stopwords = []

for doc in tokenized_docs_no_punctuation:
    new_term_vector = []
    for word in doc:
        if not word in stopwords.words('english'):
            new_term_vector.append(word)

    tokenized_docs_no_stopwords.append(new_term_vector)

print(tokenized_docs_no_stopwords)
```

```
[['Here', 'simple', 'basic', 'sentences'], ['They', 'wo', 'nt', 'interesting', 'I', 'afra
id'], ['The', 'point', 'examples', 'learn', 'basic', 'text', 'cleaning', 'works', 'simple
', 'data']]
```

In [9]:

```
# Stemming and Lemmatizing
from nltk.stem.porter import PorterStemmer
from nltk.stem.snowball import SnowballStemmer
from nltk.stem.wordnet import WordNetLemmatizer

porter = PorterStemmer()
snowball = SnowballStemmer('english')
wordnet = WordNetLemmatizer()

preprocessed_docs = []

for doc in tokenized_docs_no_stopwords:
    final_doc = []
    for word in doc:
        final_doc.append(porter.stem(word))
        #final_doc.append(snowball.stem(word))
        #final_doc.append(wordnet.lemmatize(word))

    preprocessed_docs.append(final_doc)

print(preprocessed_docs)
```

```
[['here', 'simpl', 'basic', 'sentenc'], ['they', 'wo', 'nt', 'interest', 'I', 'afraid'],
['the', 'point', 'exampl', 'learn', 'basic', 'text', 'clean', 'work', 'simpl', 'data']]
```

In [ ]: