Lesson 24 (Bag of Words II)

Consider the collection of documents given below where each document is a sentence.

	Document
0	Jack and Jill went up the hill.
1	Jack and Jill are siblings.
2	The quick brown fox jumped over the lazy dog.

(a) Use `CountVectorizer` to construct a *bag-of-words* representation of the document data. The output is a *sparse matrix*. A *sparse matrix* consists of the indices and values of non-zero elements of a matrix. The method `.toarray()` converts sparse matrices to ordinary matrices.

- How many non-zero elements are in the *sparse matrix* generated by `CountVectorizer`? Use `.nnz`.

(b) Use Pandas to construct a *document term matrix* for the document data. The column headings of a document term matrix are the unique words in the documents and the values in the matrix are word counts.

- `vectorizer.get_feature_names()` generates a list of unique words used in the documents.
- `word_counts.toarray()` converts the word counts to an ordinary matrix (only possible for a small vocabulary of words).
- Use the `pd.DataFrame()` command to create a data frame `df` of word counts.

(c) What is the vocabulary size of these documents?

(d) *Stop words* are words like “the” and “and” that are normally not useful in natural language processing. These words can be removed by using the option `stop_words='english'` in `CountVectorizer`. Compute a new document term matrix with stop words removed.

- What is the vocabulary size with stop words removed?

(e) Compute a Term Frequency Matrix with stop words removed.

(f) Use a correlation matrix to determine which documents are the most similar and which are the most dissimilar to each other. Repeat using cosine similiarity instead. (Use the commands below.)

```
from sklearn.metrics.pairwise import cosine_similarity

cosine_similarity(TF)
```

(g) A bag of words representation does not preserve word order information. Some local word order information can be preserved by using word *ngrams*. Use the option `ngram_range=(1,2)` to add word 2grams to the vocabulary.

- What are 2grams?
- What is the size of the vocabulary if 2grams are included?