Elements Of Data Science - F2023

Week 10: NLP, Sentiment Analysis and Topic Modeling

11/27/2023

TODOs

- Readings:
  - PML Chapter 11: Working with Unlabeled Data - Clustering Analysis, Sections 11.1 and 11.2
  - [Optional] PDSH 5.11 k-Means
  - [Optional] Data Science From Scratch Chap 22: Recommender Systems
- Quiz 10, Due **Mon Dec 4 15th, 11:59pm ET**
- HW3, Due **Fri Dec 1st 11:59pm**

Today

- Pipelines
- NLP
- Sentiment Analysis
- Topic Modeling

Questions?

Environment Setup

Environment Setup

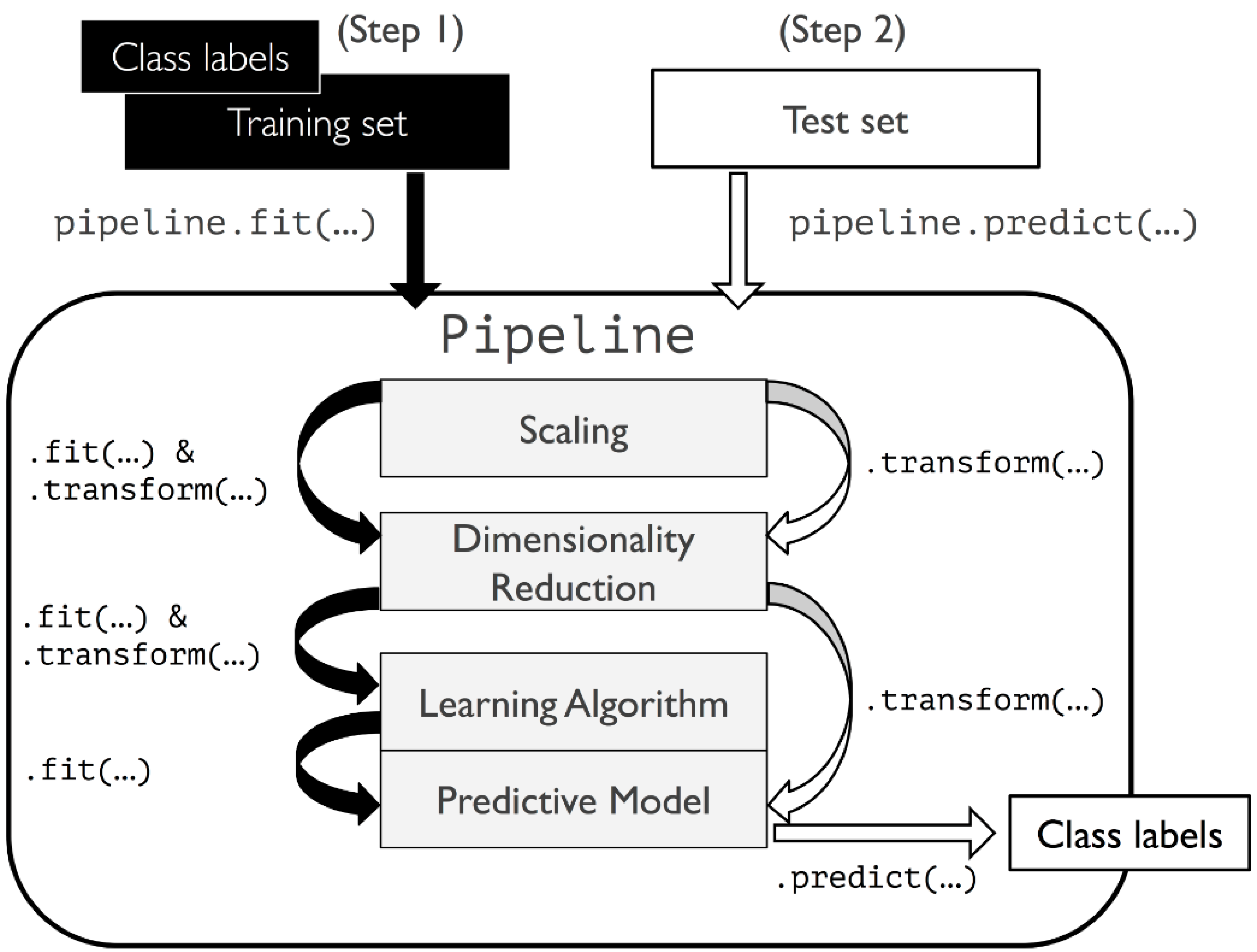
In [1]:

```
import numpy
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

import warnings
warnings.filterwarnings('ignore')


sns.set_style('darkgrid')
%matplotlib inline
```

Pipelines in sklearn

- Pipelines are wrappers used to string together transformers and estimators
    - sequentially apply a series of transforms, eg, `.fit_transform()` and `.transform()`
    - followed by a prediction, eg. `.fit()` and `.predict()`

Pipelines in sklearn



From PML

Binary Classification With All Numeric Features Setup

Binary Classification With All Numeric Features Setup

In [2]:

```
# Example from PML - scaling > feature extraction > classification
from sklearn.datasets import load_breast_cancer
from sklearn.model_selection import train_test_split
bc = load_breast_cancer()
X_bc,y_bc = bc['data'],bc['target']
X_bc_train,X_bc_test,y_bc_train,y_bc_test = train_test_split(X_bc,
                                                             y_bc,
                                                             test_size=0.3,
                                                             stratify=y_bc,
                                                             random_state=123)

# print without scientific notation
numpy.set_printoptions(suppress = True)

print("training set has rows: {} columns: {}".format(*X_bc_train.shape))

# all real valued features
print('Feature names: ',bc.feature_names[:3], ' ...')
print('Corresponding Feature values:', X_bc_train[:1,:3][0].round(2), ' ...')
print('Target names: ', bc.target_names)
```

```
training set has rows: 398 columns: 30
Feature names:  ['mean radius' 'mean texture' 'mean perimeter']  ...
Corresponding Feature values: [12.99 14.23 84.08]  ...
Target names:  ['malignant' 'benign']
```

Pipelines in sklearn

Pipelines in sklearn

In [3]:

```python
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.linear_model import LogisticRegression

# Pipeline: list of (name,object) pairs
pipe1 = Pipeline([('scale',StandardScaler()),                    # scale
                  ('pca',PCA(n_components=15)),                   # reduce dimensions
                  ('lr',LogisticRegression(solver='saga',
                                           max_iter=1000,
                                           random_state=12)),  # classifier
                 ])

pipe1.fit(X_bc_train,y_bc_train)

print(f'train set accuracy: {pipe1.score(X_bc_train,y_bc_train).round(3)}')
print(f'test set accuracy : {pipe1.score(X_bc_test,y_bc_test).round(3)}')
```

```
train set accuracy: 0.987
test set accuracy : 0.982
```

Pipelines in sklearn

In [3]:

```python
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.linear_model import LogisticRegression

# Pipeline: list of (name,object) pairs
pipe1 = Pipeline([('scale',StandardScaler()),                   # scale
                  ('pca',PCA(n_components=15)),                  # reduce dimensions
                  ('lr',LogisticRegression(solver='saga',
                                           max_iter=1000,
                                           random_state=12)),  # classifier
                 ])

pipe1.fit(X_bc_train,y_bc_train)

print(f'train set accuracy: {pipe1.score(X_bc_train,y_bc_train).round(3)}')
print(f'test set accuracy : {pipe1.score(X_bc_test,y_bc_test).round(3)}')
```

```
train set accuracy: 0.987
test set accuracy : 0.982
```

In [4]:

```python
# access pipeline components by name like a dictionary
pipe1['lr'].coef_.round(2)
```

Out[4]:

```
array([[-2.25,  1.41,  0.43, -0.61,  0.97, -0.08, -0.2 ,  0.7 , -1.47,
         0.06,  0.06, -0.5 ,  0.17,  1.07, -0.57]])
```

Pipelines in sklearn

In [3]:

```
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.linear_model import LogisticRegression

# Pipeline: list of (name,object) pairs
pipe1 = Pipeline([('scale',StandardScaler()),                    # scale
                  ('pca',PCA(n_components=15)),                   # reduce dimensions
                  ('lr',LogisticRegression(solver='saga',
                                           max_iter=1000,
                                           random_state=12)),  # classifier
                 ])

pipe1.fit(X_bc_train,y_bc_train)

print(f'train set accuracy: {pipe1.score(X_bc_train,y_bc_train).round(3)}')
print(f'test set accuracy : {pipe1.score(X_bc_test,y_bc_test).round(3)}')
```

```
train set accuracy: 0.987
test set accuracy : 0.982
```

In [4]:

```
# access pipeline components by name like a dictionary
pipe1['lr'].coef_.round(2)
```

Out[4]:

```
array([[-2.25,  1.41,  0.43, -0.61,  0.97, -0.08, -0.2 ,  0.7 , -1.47,
         0.06,  0.06, -0.5 ,  0.17,  1.07, -0.57]])
```

In [5]:

```
pipe1['pca'].components_[0].round(2)
```

Out[5]:

```
array([0.22, 0.08, 0.23, 0.22, 0.14, 0.24, 0.26, 0.26, 0.14, 0.06, 0.21,
       0.  , 0.21, 0.2 , 0.01, 0.17, 0.15, 0.18, 0.04, 0.09, 0.23, 0.08,
       0.24, 0.23, 0.13, 0.21, 0.23, 0.25, 0.12, 0.13])
```

Pipelines in sklearn: GridSearch with Pipelines

Pipelines in sklearn: GridSearch with Pipelines

- specify grid points using 'step name' + '__' (double-underscore) + 'argument'

Pipelines in sklearn: GridSearch with Pipelines

- specify grid points using 'step name' + '__' (double-underscore) + 'argument'

In [6]:
```
from sklearn.exceptions import ConvergenceWarning # needed to supress warnings
from sklearn.utils import parallel_backend        # needed to supress warnings

from sklearn.model_selection import GridSearchCV

# separate step-names and argument-names with double-underscore '__'
params1 = {'pca__n_components':[2,10,15,20],
           'lr__penalty':['none','l1','l2'],
           'lr__C':[0,.01,1,10,100]}

with parallel_backend("multiprocessing"):         # needed to supress warnings
    with warnings.catch_warnings():                # needed to supress warnings
        warnings.filterwarnings("ignore")            # needed to supress warnings

gscv = GridSearchCV(pipe1, params1, cv=3, n_jobs=-1).fit(X_bc_train,y_bc_train)

gscv.best_params_
```

Out[6]:
```
{'lr__C': 1, 'lr__penalty': 'l1', 'pca__n_components': 15}
```

Pipelines in sklearn: GridSearch with Pipelines

- specify grid points using 'step name' + '__' (double-underscore) + 'argument'

In [6]:
```python
from sklearn.exceptions import ConvergenceWarning # needed to supress warnings
from sklearn.utils import parallel_backend        # needed to supress warnings

from sklearn.model_selection import GridSearchCV

# separate step-names and argument-names with double-underscore '__'
params1 = {'pca__n_components':[2,10,15,20],
           'lr__penalty':['none','l1','l2'],
           'lr__C':[0,.01,1,10,100]}

with parallel_backend("multiprocessing"):          # needed to supress warnings
    with warnings.catch_warnings():                 # needed to supress warnings
        warnings.filterwarnings("ignore")            # needed to supress warnings

gscv = GridSearchCV(pipe1, params1, cv=3, n_jobs=-1).fit(X_bc_train,y_bc_train)

gscv.best_params_
```

Out[6]:
```
{'lr__C': 1, 'lr__penalty': 'l1', 'pca__n_components': 15}
```

In [7]:
```python
score = gscv.score(X_bc_test,y_bc_test)
print(f'test set accuracy: {score:0.3f}')
```

```
test set accuracy: 0.977
```

Displaying Pipelines

Displaying Pipelines

In [8]:

```
gscv
```

Out[8]:

▸ **GridSearchCV**

▸ **estimator: Pipeline**

▸ StandardScaler

▸ PCA

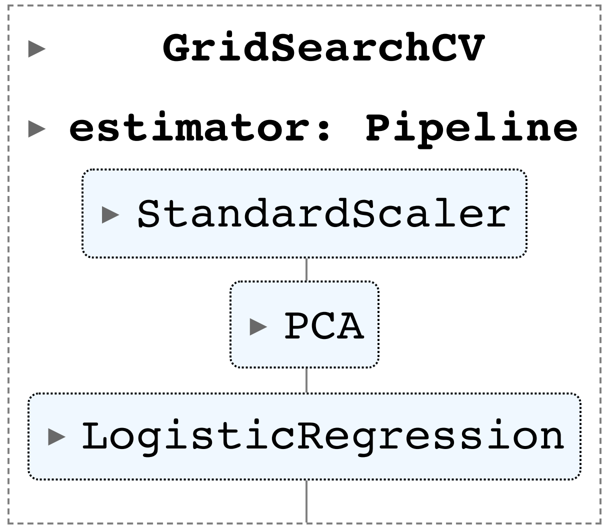
▸ LogisticRegression

Displaying Pipelines

In [8]:

```
gscv
```

Out[8]:



In [9]:

```
print(gscv)
```

```
GridSearchCV(cv=3,
             estimator=Pipeline(steps=[('scale', StandardScaler()),
                                       ('pca', PCA(n_components=15)),
                                       ('lr',
                                        LogisticRegression(max_iter=1000,
                                                           random_state=12,
                                                           solver='saga'))]),
             n_jobs=-1,
             param_grid={'lr__C': [0, 0.01, 1, 10, 100],
                         'lr__penalty': ['none', 'l1', 'l2'],
                         'pca__n_components': [2, 10, 15, 20]})
```

Displaying Pipelines Cont.

Displaying Pipelines Cont.

In [10]:

```
gscv.best_estimator_
```

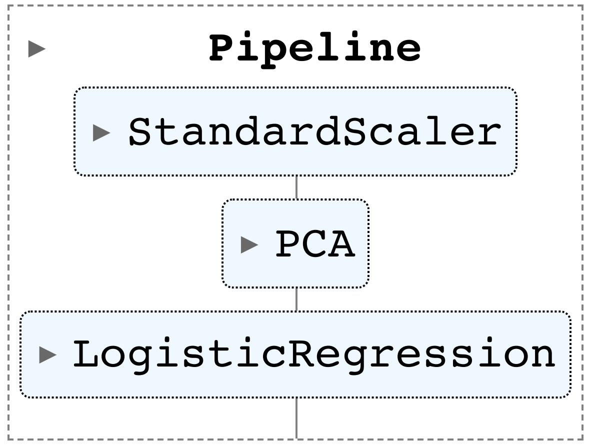
Out[10]:

```
▸ Pipeline
  ▸ StandardScaler
  ▸ PCA
  ▸ LogisticRegression
```

Displaying Pipelines Cont.

In [10]:
```
gscv.best_estimator_
```

Out[10]:



In [11]:
```
print(gscv.best_estimator_)
```

```
Pipeline(steps=[('scale', StandardScaler()), ('pca', PCA(n_components=15)),
                ('lr',
                 LogisticRegression(C=1, max_iter=1000, penalty='l1',
                                    random_state=12, solver='saga'))])
```

Pipelines in sklearn with `make_pipeline`

- shorthand for Pipeline
- step names are lowercase of class names

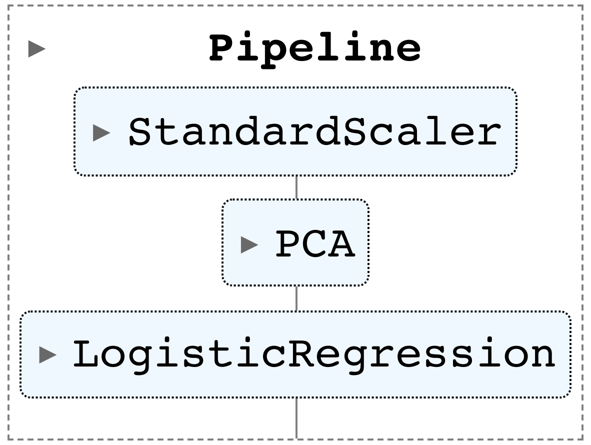
Pipelines in sklearn with `make_pipeline`

- shorthand for Pipeline
- step names are lowercase of class names

In [12]:

```python
from sklearn.pipeline import make_pipeline

# make_pipeline: arguments in order of how they should be applied
pipe2 = make_pipeline(StandardScaler(),                         # center and scale data
                      PCA(n_components=2),                      # extract 2 dimensions
                      LogisticRegression(random_state=123) # classify using logistic regression
                     )
pipe2.fit(X_bc_train,y_bc_train)

pipe2
```

Out[12]:



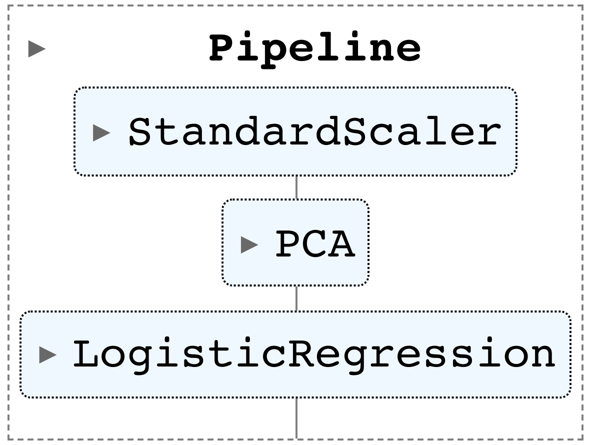
Pipelines in sklearn with `make_pipeline`

- shorthand for Pipeline
- step names are lowercase of class names

In [12]:

```
from sklearn.pipeline import make_pipeline

# make_pipeline: arguments in order of how they should be applied
pipe2 = make_pipeline(StandardScaler(),                        # center and scale data
                      PCA(n_components=2),                     # extract 2 dimensions
                      LogisticRegression(random_state=123) # classify using logistic regression
                     )
pipe2.fit(X_bc_train,y_bc_train)

pipe2
```

Out[12]:



In [13]:

```
pipe2['logisticregression'].coef_.round(2)
```

Out[13]:

```
array([[-1.91,  1.04]])
```

ColumnTransformer

- Transform sets of columns differently as part of a pipeline
- For example: makes it possible to transform categorical and numeric differently

Binary Classification With Mixed Features, Missing Data

Binary Classification With Mixed Features, Missing Data

In [14]:

```python
# from https://scikit-learn.org/stable/auto_examples/compose/plot_column_transformer_mixed_types.html#sphx-glr-auto-examples-cc
titanic_url = ('https://raw.githubusercontent.com/amueller/'
               'scipy-2017-sklearn/091d371/notebooks/datasets/titanic3.csv')
df_titanic = pd.read_csv(titanic_url)[['age','fare','embarked','sex','pclass','survived']]
# Numeric Features:
# - age: float.
# - fare: float.
# Categorical Features:
# - embarked: categories encoded as strings {'C', 'S', 'Q'}.
# - sex: categories encoded as strings {'female', 'male'}.
# - pclass: ordinal integers {1, 2, 3}.
df_titanic.head(1)
```

Out[14]:

	age	fare	embarked	sex	pclass	survived
0	29.0	211.3375	S	female	1	1

Binary Classification With Mixed Features, Missing Data

In [14]:

```python
# from https://scikit-learn.org/stable/auto_examples/compose/plot_column_transformer_mixed_types.html#sphx-glr-auto-examples-co
titanic_url = ('https://raw.githubusercontent.com/amueller/'
               'scipy-2017-sklearn/091d371/notebooks/datasets/titanic3.csv')
df_titanic = pd.read_csv(titanic_url)[['age','fare','embarked','sex','pclass','survived']]
# Numeric Features:
# - age: float.
# - fare: float.
# Categorical Features:
# - embarked: categories encoded as strings {'C', 'S', 'Q'}.
# - sex: categories encoded as strings {'female', 'male'}.
# - pclass: ordinal integers {1, 2, 3}.
df_titanic.head(1)
```

Out[14]:

	age	fare	embarked	sex	pclass	survived
0	29.0	211.3375	S	female	1	1

In [15]:

```python
df_titanic.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1309 entries, 0 to 1308
Data columns (total 6 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1046 non-null   float64
 1   fare      1308 non-null   float64
 2   embarked  1307 non-null   object
 3   sex       1309 non-null   object
 4   pclass    1309 non-null   int64
 5   survived  1309 non-null   int64
dtypes: float64(2), int64(2), object(2)
memory usage: 61.5+ KB
```

ColumnTransformer Cont.

ColumnTransformer Cont.

In [16]:

```
from sklearn.compose import ColumnTransformer
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import OneHotEncoder

# specify columns subset
numeric_features = ['age', 'fare']
# specify pipeline to apply to those columns
numeric_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='median')), # fill missing values with median
    ('scaler', StandardScaler())])                  # scale features
```

ColumnTransformer Cont.

In [16]:

```python
from sklearn.compose import ColumnTransformer
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import OneHotEncoder

# specify columns subset
numeric_features = ['age', 'fare']
# specify pipeline to apply to those columns
numeric_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='median')), # fill missing values with median
    ('scaler', StandardScaler())])                  # scale features
```

In [17]:

```python
categorical_features = ['embarked', 'sex', 'pclass']
categorical_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='constant', fill_value='missing')), # fill missing value with 'missing'
    ('onehot', OneHotEncoder(handle_unknown='ignore'))])                   # one hot encode
```

ColumnTransformer Cont.

In [16]:

```
from sklearn.compose import ColumnTransformer
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import OneHotEncoder

# specify columns subset
numeric_features = ['age', 'fare']
# specify pipeline to apply to those columns
numeric_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='median')), # fill missing values with median
    ('scaler', StandardScaler())])                 # scale features
```

In [17]:

```
categorical_features = ['embarked', 'sex', 'pclass']
categorical_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='constant', fill_value='missing')), # fill missing value with 'missing'
    ('onehot', OneHotEncoder(handle_unknown='ignore'))])                    # one hot encode
```

In [18]:

```
# combine column pipelines
preprocessor = ColumnTransformer(
    transformers=[('num', numeric_transformer, numeric_features),
                  ('cat', categorical_transformer, categorical_features)
                 ])
```

ColumnTransformer Cont.

In [16]:
```
from sklearn.compose import ColumnTransformer
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import OneHotEncoder

# specify columns subset
numeric_features = ['age', 'fare']
# specify pipeline to apply to those columns
numeric_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='median')), # fill missing values with median
    ('scaler', StandardScaler())])                 # scale features
```

In [17]:
```
categorical_features = ['embarked', 'sex', 'pclass']
categorical_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='constant', fill_value='missing')), # fill missing value with 'missing'
    ('onehot', OneHotEncoder(handle_unknown='ignore'))])                   # one hot encode
```

In [18]:
```
# combine column pipelines
preprocessor = ColumnTransformer(
    transformers=[('num', numeric_transformer, numeric_features),
                  ('cat', categorical_transformer, categorical_features)
                 ])
```

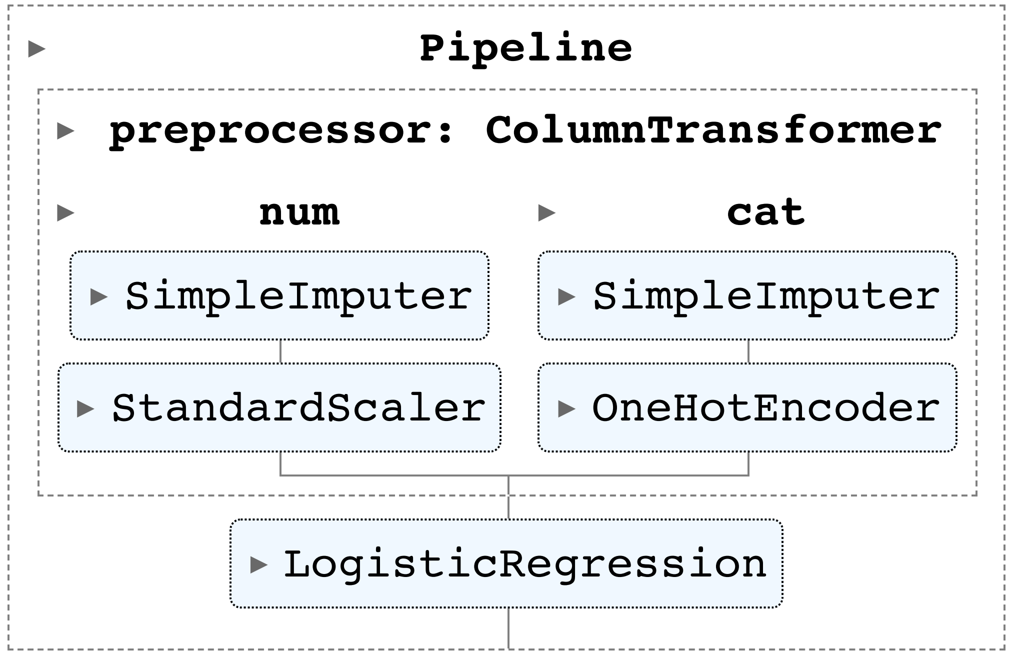
In [19]:
```
# add a final prediction step
pipe3 = Pipeline(steps=[('preprocessor', preprocessor),
                        ('classifier', LogisticRegression(solver='lbfgs', random_state=42))
                       ])
```

ColumnTransformer Cont.

ColumnTransformer Cont.

In [20]:
```python
pipe3
```

Out[20]:



ColumnTransformer Cont.

ColumnTransformer Cont.

In [21]:

```python
X_titanic = df_titanic.drop('survived', axis=1)
y_titanic = df_titanic['survived']

X_titanic_train, X_titanic_test, y_titanic_train, y_titanic_test = train_test_split(X_titanic,
                                                                                    y_titanic,
                                                                                    test_size=0.2,
                                                                                    random_state=142)
pipe3.fit(X_titanic_train, y_titanic_train)
print(f"train set score: {pipe3.score(X_titanic_train, y_titanic_train).round(3)}")
print(f"test set score : {pipe3.score(X_titanic_test, y_titanic_test).round(3)}")
```

```
train set score: 0.796
test set score : 0.756
```

ColumnTransformer Cont.

In [21]:

```
X_titanic = df_titanic.drop('survived', axis=1)
y_titanic = df_titanic['survived']

X_titanic_train, X_titanic_test, y_titanic_train, y_titanic_test = train_test_split(X_titanic,
                                                                                    y_titanic,
                                                                                    test_size=0.2,
                                                                                    random_state=142)
pipe3.fit(X_titanic_train, y_titanic_train)
print(f"train set score: {pipe3.score(X_titanic_train, y_titanic_train).round(3)}")
print(f"test set score : {pipe3.score(X_titanic_test, y_titanic_test).round(3)}")
```

```
train set score: 0.796
test set score : 0.756
```

In [22]:

```
from sklearn.model_selection import GridSearchCV

# grid search deep inside the pipeline
param_grid = {
    'preprocessor__num__imputer__strategy': ['mean', 'median'],
    'classifier__C': [0.1, 1.0, 10, 100],
}

gs_pipeline = GridSearchCV(pipe3, param_grid, cv=3)
gs_pipeline.fit(X_titanic_train, y_titanic_train)
print(f"best test set score from grid search: {gs_pipeline.score(X_titanic_test, y_titanic_test).round(3)}")
print(f"best parameter settings: {gs_pipeline.best_params_}")
```

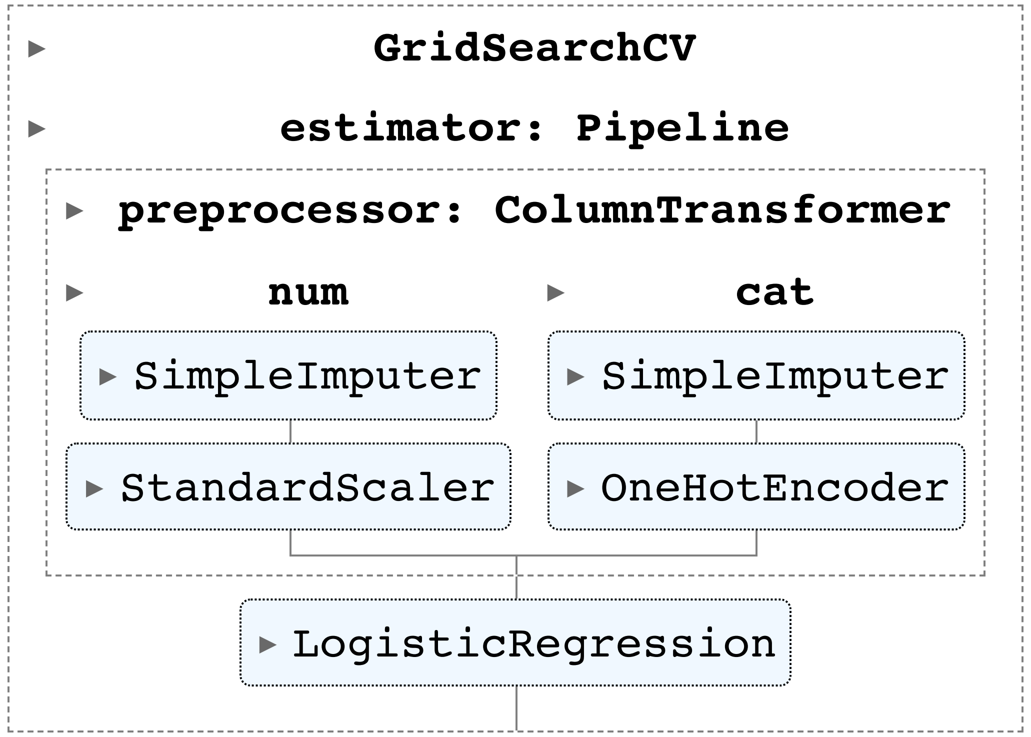
```
best test set score from grid search: 0.752
best parameter settings: {'classifier__C': 0.1, 'preprocessor__num__imputer__strategy': 'mean'}
```

ColumnTransformer Cont.

In [23]:

```
gs_pipeline
```

Out[23]:



Questions re Pipelines?

Natural Language Processing (NLP)

- Analyzing and interacting with natural language
- Python Libraries
  - **sklearn**
  - nltk
  - **spaCy**
  - gensim
  - ...

Natural Language Processing (NLP)

- Many NLP Tasks
    - **sentiment analysis**
    - **topic modeling**
    - entity detection
    - machine translation
    - natural language generation
    - question answering
    - relationship extraction
    - automatic summarization
    - ...

Recall: Python Builtin String Functions

Recall: Python Builtin String Functions

In [24]:
```python
doc = "D.S. is fun!"
doc
```

Out[24]:
```
'D.S. is fun!'
```

Recall: Python Builtin String Functions

In [24]:
```
doc = "D.S. is fun!"
doc
```

Out[24]: 'D.S. is fun!'

In [25]:
```
doc.lower(),doc.upper()       # change capitalization
```

Out[25]: ('d.s. is fun!', 'D.S. IS FUN!')

Recall: Python Builtin String Functions

In [24]:
```python
doc = "D.S. is fun!"
doc
```
Out[24]:
```
'D.S. is fun!'
```

In [25]:
```python
doc.lower(),doc.upper()       # change capitalization
```
Out[25]:
```
('d.s. is fun!', 'D.S. IS FUN!')
```

In [26]:
```python
doc.split() , doc.split('.')  # split a string into parts (default is whitespace)
```
Out[26]:
```
(['D.S.', 'is', 'fun!'], ['D', 'S', ' is fun!'])
```

Recall: Python Builtin String Functions

In [24]:
```
doc = "D.S. is fun!"
doc
```
Out[24]: 'D.S. is fun!'

In [25]:
```
doc.lower(),doc.upper()        # change capitalization
```
Out[25]: ('d.s. is fun!', 'D.S. IS FUN!')

In [26]:
```
doc.split() , doc.split('.')  # split a string into parts (default is whitespace)
```
Out[26]: (['D.S.', 'is', 'fun!'], ['D', 'S', ' is fun!'])

In [27]:
```
' | '.join(['ab','c','d'])      # join items in a list together
```
Out[27]: 'ab | c | d'

Recall: Python Builtin String Functions

In [24]:
```
doc = "D.S. is fun!"
doc
```
Out[24]:
```
'D.S. is fun!'
```

In [25]:
```
doc.lower(),doc.upper()       # change capitalization
```
Out[25]:
```
('d.s. is fun!', 'D.S. IS FUN!')
```

In [26]:
```
doc.split() , doc.split('.')  # split a string into parts (default is whitespace)
```
Out[26]:
```
(['D.S.', 'is', 'fun!'], ['D', 'S', ' is fun!'])
```

In [27]:
```
' | '.join(['ab','c','d'])      # join items in a list together
```
Out[27]:
```
'ab | c | d'
```

In [28]:
```
'|'.join(doc[:5])             # a string itself is treated like a list of characters
```
Out[28]:
```
'D|.|S|.| '
```

Recall: Python Builtin String Functions

In [24]:
```
doc = "D.S. is fun!"
doc
```
Out[24]:
```
'D.S. is fun!'
```

In [25]:
```
doc.lower(),doc.upper()       # change capitalization
```
Out[25]:
```
('d.s. is fun!', 'D.S. IS FUN!')
```

In [26]:
```
doc.split() , doc.split('.')  # split a string into parts (default is whitespace)
```
Out[26]:
```
(['D.S.', 'is', 'fun!'], ['D', 'S', ' is fun!'])
```

In [27]:
```
' | '.join(['ab','c','d'])      # join items in a list together
```
Out[27]:
```
'ab | c | d'
```

In [28]:
```
'|'.join(doc[:5])             # a string itself is treated like a list of characters
```
Out[28]:
```
'D|.|S|.| '
```

In [29]:
```
'  tes t   '.strip()            # remove whitespace from the beginning and end of a string
```
Out[29]:
```
'tes t'
```

Recall: Python Builtin String Functions

```
In [24]: doc = "D.S. is fun!"
         doc
Out[24]: 'D.S. is fun!'

In [25]: doc.lower(),doc.upper()       # change capitalization
Out[25]: ('d.s. is fun!', 'D.S. IS FUN!')

In [26]: doc.split() , doc.split('.')  # split a string into parts (default is whitespace)
Out[26]: (['D.S.', 'is', 'fun!'], ['D', 'S', ' is fun!'])

In [27]: ' | '.join(['ab','c','d'])      # join items in a list together
Out[27]: 'ab | c | d'

In [28]: '|'.join(doc[:5])             # a string itself is treated like a list of characters
Out[28]: 'D|.|S|.| '

In [29]: '  tes t   '.strip()           # remove whitespace from the beginning and end of a string
Out[29]: 'tes t'
```

- and many more, see https://docs.python.org/3.10/library/string.html

NLP: The Corpus

- **corpus:** collection of documents
  - books
  - articles
  - reviews
  - tweets
  - resumes
  - sentences?
  - ...

NLP: Doc Representation

- Documents usually represented as strings
    - string: a sequence (list) of unicode characters

NLP: Doc Representation

- Documents usually represented as strings
    - string: a sequence (list) of unicode characters

In [30]:
```
sample_doc = "D.S. is fun!\nIt's  true."
print(sample_doc)
```

```
D.S. is fun!
It's  true.
```

NLP: Doc Representation

- Documents usually represented as strings
    - string: a sequence (list) of unicode characters

In [30]:
```
sample_doc = "D.S. is fun!\nIt's  true."
print(sample_doc)
```

```
D.S. is fun!
It's  true.
```

In [31]:
```
'|'.join(sample_doc)
```

Out[31]:
```
"D|.|S|.| |i|s| |f|u|n|!|\n|I|t|'|s| | |t|r|u|e|."
```

NLP: Doc Representation

- Documents usually represented as strings
    - string: a sequence (list) of unicode characters

In [30]:
```
sample_doc = "D.S. is fun!\nIt's  true."
print(sample_doc)
```

```
D.S. is fun!
It's  true.
```

In [31]:
```
'|'.join(sample_doc)
```

Out[31]:
```
"D|.|S|.| |i|s| |f|u|n|!|\n|I|t|'|s| | |t|r|u|e|."
```

- Need to split this up into parts (**tokens**)
- Good job for **Regular Expressions**

Aside: Regular Expressions

- Strings that define search patterns over text
- Useful for finding/replacing/grouping
- python `re` library (others available)

Aside: Regular Expressions

- Strings that define search patterns over text
- Useful for finding/replacing/grouping
- python `re` library (others available)

In [32]:
```
print(sample_doc)
```

```
D.S. is fun!
It's  true.
```

Aside: Regular Expressions

- Strings that define search patterns over text
- Useful for finding/replacing/grouping
- python `re` library (others available)

In [32]:
```
print(sample_doc)
```

```
D.S. is fun!
It's  true.
```

In [33]:
```
import re
# Find all of the whitespaces in doc
# '\s+' means "one or more whitespace characters"
re.findall(r'\s+',sample_doc)
```

Out[33]:
```
[' ', ' ', '\n', '  ']
```

Aside: Regular Expressions

Just some of the special character definitions:

- `.` : any single character except newline (r'.' matches 'x')
- `*` : match 0 or more repetitions (r'x*' matches 'x','xx','')
- `+` : match 1 or more repetitions (r'x+' matches 'x','xx')
- `?` : match 0 or 1 repetitions (r'x?' matches 'x' or '')

- `^` : beginning of string (r'^D' matches 'D.S.')
- `$` : end of string (r'fun!$' matches 'DS is fun!'`)

Aside: Regular Expression Cont.

- `[ ]` : a set of characters (^ as first element = not)
- `\s` : whitespace character (Ex: [\t\n\r\f\v])
- `\S` : non-whitespace character (Ex: [^ \t\n\r\f\v])
- `\w` : word character (Ex: [a-zA-Z0-9_])
- `\W` : non-word character
- `\b` : boundary between \w and \W
- and many more!

- See regex101.com for examples and testing

Aside: Regex Python Functions

Aside: Regex Python Functions

In [34]:
```
r'\w*u\w*' # a string of word characters containing the letter 'u'
```

Out[34]:
```
'\\w*u\\w*'
```

Aside: Regex Python Functions

In [34]:
```
r'\w*u\w*' # a string of word characters containing the letter 'u'
```
Out[34]:
```
'\\w*u\\w*'
```

In [35]:
```
re.findall(r'\w*u\w*',sample_doc) # return all substrings that match a pattern
```
Out[35]:
```
['fun', 'true']
```

Aside: Regex Python Functions

In [34]:
```
r'\w*u\w*' # a string of word characters containing the letter 'u'
```
Out[34]:
```
'\\w*u\\w*'
```

In [35]:
```
re.findall(r'\w*u\w*',sample_doc) # return all substrings that match a pattern
```
Out[35]:
```
['fun', 'true']
```

In [36]:
```
re.sub(r'\w*u\w*','XXXX',sample_doc) # substitute all substrings that match a pattern
```
Out[36]:
```
"D.S. is XXXX!\nIt's  XXXX."
```

Aside: Regex Python Functions

```
In [34]: r'\w*u\w*' # a string of word characters containing the letter 'u'
Out[34]: '\\w*u\\w*'

In [35]: re.findall(r'\w*u\w*',sample_doc) # return all substrings that match a pattern
Out[35]: ['fun', 'true']

In [36]: re.sub(r'\w*u\w*','XXXX',sample_doc) # substitute all substrings that match a pattern
Out[36]: "D.S. is XXXX!\nIt's  XXXX."

In [37]: re.split(r'\w*u\w*',sample_doc) # split substrings on a pattern
Out[37]: ['D.S. is ', "!\nIt's  ", '.']
```

NLP: Tokenization

- **tokens:** strings that make up a document ('the', 'cat',...)
- **tokenization:** convert a document into tokens
- **vocabulary:** set of unique tokens (terms) in corpus

NLP: Tokenization

- **tokens:** strings that make up a document ('the', 'cat',...)
- **tokenization:** convert a document into tokens
- **vocabulary:** set of unique tokens (terms) in corpus

In [38]:
```
# split on whitespace
re.split(r'\s+', sample_doc)
```

Out[38]:
```
['D.S.', 'is', 'fun!', "It's", 'true.']
```

NLP: Tokenization

- **tokens:** strings that make up a document ('the', 'cat',...)
- **tokenization:** convert a document into tokens
- **vocabulary:** set of unique tokens (terms) in corpus

In [38]:

```
# split on whitespace
re.split(r'\s+', sample_doc)
```

Out[38]: ['D.S.', 'is', 'fun!', "It's", 'true.']

In [39]:

```
# find tokens of length 2+ word characters
re.findall(r'\b\w\w+\b',sample_doc)
```

Out[39]: ['is', 'fun', 'It', 'true']

NLP: Tokenization

- **tokens:** strings that make up a document ('the', 'cat',...)
- **tokenization:** convert a document into tokens
- **vocabulary:** set of unique tokens (terms) in corpus

In [38]:
```
# split on whitespace
re.split(r'\s+', sample_doc)
```

Out[38]:
```
['D.S.', 'is', 'fun!', "It's", 'true.']
```

In [39]:
```
# find tokens of length 2+ word characters
re.findall(r'\b\w\w+\b',sample_doc)
```

Out[39]:
```
['is', 'fun', 'It', 'true']
```

In [40]:
```
# find tokens of length 2+ non-space characters
re.findall(r"\b\S\S+\b", sample_doc)
```

Out[40]:
```
['D.S', 'is', 'fun', "It's", 'true']
```

NLP: Tokenization

- **tokens:** strings that make up a document ('the', 'cat',...)
- **tokenization:** convert a document into tokens
- **vocabulary:** set of unique tokens (terms) in corpus

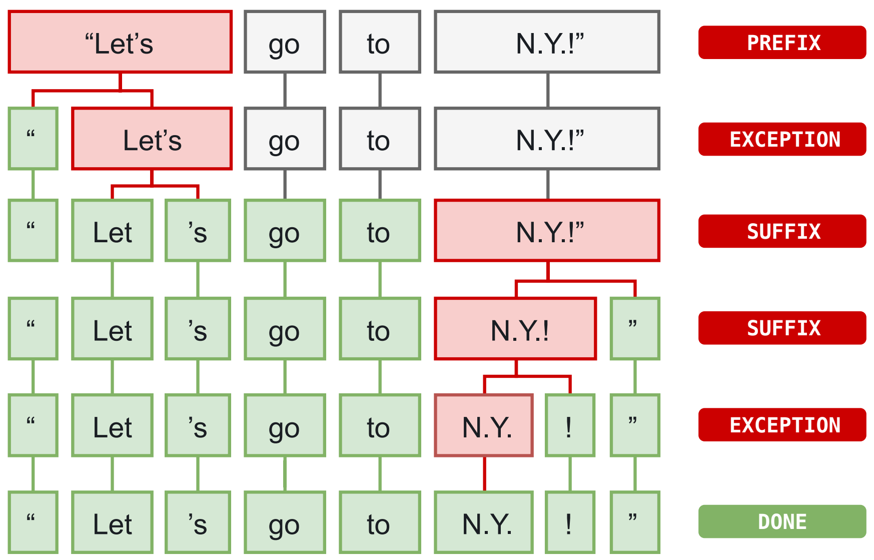
In [38]:
```python
# split on whitespace
re.split(r'\s+', sample_doc)
```
Out[38]:
```
['D.S.', 'is', 'fun!', "It's", 'true.']
```

In [39]:
```python
# find tokens of length 2+ word characters
re.findall(r'\b\w\w+\b',sample_doc)
```
Out[39]:
```
['is', 'fun', 'It', 'true']
```

In [40]:
```python
# find tokens of length 2+ non-space characters
re.findall(r"\b\S\S+\b", sample_doc)
```
Out[40]:
```
['D.S', 'is', 'fun', "It's", 'true']
```

In [41]:
```python
# example vocabulary
set(re.findall(r"\b\S\S+\b", sample_doc))
```
Out[41]:
```
{'D.S', "It's", 'fun', 'is', 'true'}
```

NLP: Tokenization in spaCy



From https://spacy.io/usage/linguistic-features

First, the raw text is split on whitespace characters, similar to text.split(' '). Then, the tokenizer processes the text from left to right. On each substring, it performs two checks:

- Does the substring match a tokenizer exception rule? For example, “don’t” does not contain whitespace, but should be split into two tokens, “do” and “n’t”, while “U.K.” should always remain one token.
- Can a prefix, suffix or infix be split off? For example punctuation like commas, periods, hyphens or

NLP: Other Options for Preprocessing

- lowercase
- remove special characters
- add `<START>`, `<END>` tags
- **lemmatization:** perform morphological analysis
    - 'studies' becomes 'study'
    - 'studying' becomes 'study'

NLP: Bag of Words

- **Bag of Words** (BOW) representation: ignore token order

In [42]:
```
sample_doc
```

Out[42]:
```
"D.S. is fun!\nIt's  true."
```

NLP: Bag of Words

- **Bag of Words** (BOW) representation: ignore token order

In [42]:
```
sample_doc
```

Out[42]: `"D.S. is fun!\nIt's  true."`

In [43]:
```
sample_doc.lower()
```

Out[43]: `"d.s. is fun!\nit's  true."`

NLP: Bag of Words

- **Bag of Words** (BOW) representation: ignore token order

In [42]:
```
sample_doc
```

Out[42]:
```
"D.S. is fun!\nIt's  true."
```

In [43]:
```
sample_doc.lower()
```

Out[43]:
```
"d.s. is fun!\nit's  true."
```

In [44]:
```
sorted(re.findall(r'\b\S\S+\b', sample_doc.lower()))
```

Out[44]:
```
['d.s', 'fun', 'is', "it's", 'true']
```

NLP: n-Grams

- **Unigram:** single token
- **Bigram:** combination of two ordered tokens
- **n-Gram:** combination of n ordered tokens
- The larger *n* is, the larger the vocabulary

NLP: n-Grams

- **Unigram:** single token
- **Bigram:** combination of two ordered tokens
- **n-Gram:** combination of n ordered tokens
- The larger *n* is, the larger the vocabulary

In [45]:

```
# Bigram example:
tokens = '<start> ds is fun ds is great <end>'.split()
print("bigrams     : ",    [tokens[i]+'_'+tokens[i+1] for i in range(len(tokens)-1)])
print("bigram vocab: ",set([tokens[i]+'_'+tokens[i+1] for i in range(len(tokens)-1)]))
```

```
bigrams     :  ['<start>_ds', 'ds_is', 'is_fun', 'fun_ds', 'ds_is', 'is_great', 'great_<end>']
bigram vocab:  {'ds_is', 'is_great', 'is_fun', 'great_<end>', 'fun_ds', '<start>_ds'}
```

NLP: n-Grams

- **Unigram:** single token
- **Bigram:** combination of two ordered tokens
- **n-Gram:** combination of n ordered tokens
- The larger *n* is, the larger the vocabulary

In [45]:

```
# Bigram example:
tokens = '<start> ds is fun ds is great <end>'.split()
print("bigrams      : ",     [tokens[i]+'_'+tokens[i+1] for i in range(len(tokens)-1)])
print("bigram vocab: ",set([tokens[i]+'_'+tokens[i+1] for i in range(len(tokens)-1)]))
```

```
bigrams      :  ['<start>_ds', 'ds_is', 'is_fun', 'fun_ds', 'ds_is', 'is_great', 'great_<end>']
bigram vocab:  {'ds_is', 'is_great', 'is_fun', 'great_<end>', 'fun_ds', '<start>_ds'}
```

In [46]:

```
# Trigrams example:
tokens = '<start> ds is fun ds is great <end>'.split()
['_'.join(tokens[i:i+3]) for i in range(len(tokens)-2)]
```

Out[46]:

```
['<start>_ds_is',
 'ds_is_fun',
 'is_fun_ds',
 'fun_ds_is',
 'ds_is_great',
 'is_great_<end>']
```

NLP: TF and DF

- **Term Frequency:** number of times a term is seen per document
- $\text{tf}(t, d) = \text{count of term } t \text{ in document } d$

NLP: TF and DF

- **Term Frequency:** number of times a term is seen per document
- $\text{tf}(t, d) = \text{count of term } t \text{ in document } d$

In [47]:
```
example_corpus = ['red green blue', 'red blue blue']

#Vocabulary
example_vocab = sorted(set(' '.join(example_corpus).split()))
example_vocab
```

Out[47]:
```
['blue', 'green', 'red']
```

NLP: TF and DF

- **Term Frequency:** number of times a term is seen per document
- $\mathrm{tf}(t, d) = \text{count of term } t \text{ in document } d$

In [47]:

```
example_corpus = ['red green blue', 'red blue blue']

#Vocabulary
example_vocab = sorted(set(' '.join(example_corpus).split()))
example_vocab
```

Out[47]: ['blue', 'green', 'red']

In [48]:

```
#TF
from collections import Counter
example_tf = np.zeros((len(example_corpus),len(example_vocab)))
for i,doc in enumerate(example_corpus):
    for j,term in enumerate(example_vocab):
        example_tf[i,j] = Counter(doc.split())[term]
example_tf = pd.DataFrame(example_tf,index=['doc1','doc2'],columns=example_vocab)
example_tf
```

Out[48]:

	blue	green	red
doc1	1.0	1.0	1.0
doc2	2.0	0.0	1.0

NLP: TF and DF

- **Document Frequency:** number of documents containing each term
  $\mathrm{df}(t) = \text{count of documents containing term } t$

NLP: TF and DF

- **Document Frequency:** number of documents containing each term

$$\text{df}(t) = \text{count of documents containing term } t$$

In [49]:

```
example_tf
```

Out[49]:

	blue	green	red
doc1	1.0	1.0	1.0
doc2	2.0	0.0	1.0

NLP: TF and DF

- **Document Frequency:** number of documents containing each term

$$\text{df}(t) = \text{count of documents containing term } t$$

In [49]:
```
example_tf
```

Out[49]:

	blue	green	red
doc1	1.0	1.0	1.0
doc2	2.0	0.0	1.0

In [50]:
```
#DF
example_df = example_tf.astype(bool).sum(axis=0) # how many documents contain each term (column)
example_df
```

Out[50]:
```
blue     2
green    1
red      2
dtype: int64
```

NLP: Stopwords

- terms that have high (or very low) DF and aren't informative
    - common engish terms (ex: a, the, in,...)
    - domain specific (ex, in class slides: 'data_science')
    - often removed prior to analysis
    - in sklearn
        - `min_df` : integer > 0 : keep terms that occur in at at least n documents
        - `max_df` : float in (0,1] : keep terms that occur in less than max_df% of total documents

NLP: CountVectorizer in sklearn

NLP: CountVectorizer in sklearn

In [51]:

```python
example_corpus = ['blue green red', 'blue green green']

from sklearn.feature_extraction.text import CountVectorizer
cvect = CountVectorizer(lowercase=True,    # default, transform all docs to lowercase
                        ngram_range=(1,1), # default, only unigrams
                        min_df=1,          # default, keep all terms
                        max_df=1.0,        # default, keep all terms
                       )
X_cv = cvect.fit_transform(example_corpus)
X_cv.shape
```

Out[51]:

```
(2, 3)
```

NLP: CountVectorizer in sklearn

In [51]:

```
example_corpus = ['blue green red', 'blue green green']

from sklearn.feature_extraction.text import CountVectorizer
cvect = CountVectorizer(lowercase=True,    # default, transform all docs to lowercase
                        ngram_range=(1,1), # default, only unigrams
                        min_df=1,          # default, keep all terms
                        max_df=1.0,        # default, keep all terms
                       )
X_cv = cvect.fit_transform(example_corpus)
X_cv.shape
```

Out[51]: (2, 3)

In [52]:

```
cvect.vocabulary_ # learned vocabulary, term:index pairs
```

Out[52]: {'blue': 0, 'green': 1, 'red': 2}

NLP: CountVectorizer in sklearn

In [51]:

```
example_corpus = ['blue green red', 'blue green green']

from sklearn.feature_extraction.text import CountVectorizer
cvect = CountVectorizer(lowercase=True,    # default, transform all docs to lowercase
                        ngram_range=(1,1), # default, only unigrams
                        min_df=1,          # default, keep all terms
                        max_df=1.0,        # default, keep all terms
                       )
X_cv = cvect.fit_transform(example_corpus)
X_cv.shape
```

Out[51]: (2, 3)

In [52]:

```
cvect.vocabulary_ # learned vocabulary, term:index pairs
```

Out[52]: {'blue': 0, 'green': 1, 'red': 2}

In [53]:

```
cvect.get_feature_names() # vocabulary, sorted by indexs
```

Out[53]: ['blue', 'green', 'red']

NLP: CountVectorizer in sklearn

In [51]:

```
example_corpus = ['blue green red', 'blue green green']

from sklearn.feature_extraction.text import CountVectorizer
cvect = CountVectorizer(lowercase=True,    # default, transform all docs to lowercase
                        ngram_range=(1,1), # default, only unigrams
                        min_df=1,          # default, keep all terms
                        max_df=1.0,        # default, keep all terms
                       )
X_cv = cvect.fit_transform(example_corpus)
X_cv.shape
```

Out[51]:

```
(2, 3)
```

In [52]:

```
cvect.vocabulary_ # learned vocabulary, term:index pairs
```

Out[52]:

```
{'blue': 0, 'green': 1, 'red': 2}
```

In [53]:

```
cvect.get_feature_names() # vocabulary, sorted by indexs
```

Out[53]:

```
['blue', 'green', 'red']
```

In [54]:

```
X_cv.todense() # term frequencies
```

Out[54]:

```
matrix([[1, 1, 1],
        [1, 2, 0]])
```

NLP: CountVectorizer in sklearn

In [51]:

```python
example_corpus = ['blue green red', 'blue green green']

from sklearn.feature_extraction.text import CountVectorizer
cvect = CountVectorizer(lowercase=True,    # default, transform all docs to lowercase
                        ngram_range=(1,1), # default, only unigrams
                        min_df=1,          # default, keep all terms
                        max_df=1.0,        # default, keep all terms
                       )
X_cv = cvect.fit_transform(example_corpus)
X_cv.shape
```

Out[51]:

```
(2, 3)
```

In [52]:

```python
cvect.vocabulary_ # learned vocabulary, term:index pairs
```

Out[52]:

```
{'blue': 0, 'green': 1, 'red': 2}
```

In [53]:

```python
cvect.get_feature_names() # vocabulary, sorted by indexs
```

Out[53]:

```
['blue', 'green', 'red']
```

In [54]:

```python
X_cv.todense() # term frequencies
```

Out[54]:

```
matrix([[1, 1, 1],
        [1, 2, 0]])
```

In [55]:

```python
cvect.inverse_transform(X_cv) # mapping back to terms via vocabulary mapping
```

Out[55]:

```
[array(['blue', 'green', 'red'], dtype='<U5'),
 array(['blue', 'green'], dtype='<U5')]
```

NLP: TfIdf

- What if some terms are still uninformative?
- Can we downweight terms that occur in many documents?
- **Term Frequency * Inverse Document Frequency (tf-idf)*
    - $\text{tf-idf}(t, d) = \text{tf}(t, d) \times \text{idf}(t)$
    - $\text{idf}(t) = \log \frac{1+n}{1+\text{df}(t)} + 1$

NLP: TfIdf

- What if some terms are still uninformative?
- Can we downweight terms that occur in many documents?
- *Term Frequency * Inverse Document Frequency (tf-idf)*
    - $\text{tf-idf}(t, d) = \text{tf}(t, d) \times \text{idf}(t)$
    - $\text{idf}(t) = \log \frac{1+n}{1+\text{df}(t)} + 1$

In [56]:

```python
from sklearn.feature_extraction.text import TfidfVectorizer

tfidfvect = TfidfVectorizer(norm='l2') # by default, also doing l2 normalization

X_tfidf = tfidfvect.fit_transform(example_corpus)
sorted(tfidfvect.vocabulary_.items(),key=lambda x: x[1])
```

Out[56]:

```
[('blue', 0), ('green', 1), ('red', 2)]
```

NLP: TfIdf

- What if some terms are still uninformative?
- Can we downweight terms that occur in many documents?
- *Term Frequency * Inverse Document Frequency (tf-idf)*
    - $\text{tf-idf}(t,d) = \text{tf}(t,d) \times \text{idf}(t)$
    - $\text{idf}(t) = \log \frac{1+n}{1+\text{df}(t)} + 1$

In [56]:
```
from sklearn.feature_extraction.text import TfidfVectorizer

tfidfvect = TfidfVectorizer(norm='l2') # by default, also doing l2 normalization

X_tfidf = tfidfvect.fit_transform(example_corpus)
sorted(tfidfvect.vocabulary_.items(),key=lambda x: x[1])
```

Out[56]:
```
[('blue', 0), ('green', 1), ('red', 2)]
```

In [57]:
```
X_tfidf.todense().round(2)
```

Out[57]:
```
array([[0.5 , 0.5 , 0.7 ],
       [0.45, 0.89, 0.  ]])
```

NLP: TfIdf

- What if some terms are still uninformative?
- Can we downweight terms that occur in many documents?
- *Term Frequency * Inverse Document Frequency (tf-idf)*
  - $\text{tf-idf}(t,d) = \text{tf}(t,d) \times \text{idf}(t)$
  - $\text{idf}(t) = \log \frac{1+n}{1+\text{df}(t)} + 1$

In [56]:

```
from sklearn.feature_extraction.text import TfidfVectorizer

tfidfvect = TfidfVectorizer(norm='l2') # by default, also doing l2 normalization

X_tfidf = tfidfvect.fit_transform(example_corpus)
sorted(tfidfvect.vocabulary_.items(),key=lambda x: x[1])
```

Out[56]:

```
[('blue', 0), ('green', 1), ('red', 2)]
```

In [57]:

```
X_tfidf.todense().round(2)
```

Out[57]:

```
array([[0.5 , 0.5 , 0.7 ],
       [0.45, 0.89, 0.  ]])
```

In [58]:

```
# can also use to get term frequencies by setting use_idf to False and norm to none
TfidfVectorizer(use_idf=False, norm=None).fit_transform(example_corpus).todense()
```

Out[58]:

```
matrix([[1., 1., 1.],
        [1., 2., 0.]])
```

NLP: Classification Example

NLP: Classification Example

In [59]:

```
from sklearn.datasets import fetch_20newsgroups

ngs = fetch_20newsgroups(categories=['rec.sport.baseball','rec.sport.hockey']) # dataset has 20 categories, only get two

docs_ngs = ngs['data']                          # get documents (emails)
y_ngs = ngs['target']                           # get targets ([0,1])
target_names_ngs = ngs['target_names']          # get target names (['rec.sport.baseball','rec.sport.hockey'])

print(y_ngs[1], target_names_ngs[y_ngs[1]])     # print target int and target name
print('-'*50)                                   # print a string of 50 dashes
print(docs_ngs[0].strip()[:600])                # print beginning characters of first doc, after stripping whitespace
```

```
1 rec.sport.hockey
--------------------------------------------------
From: dougb@comm.mot.com (Doug Bank)
Subject: Re: Info needed for Cleveland tickets
Reply-To: dougb@ecs.comm.mot.com
Organization: Motorola Land Mobile Products Sector
Distribution: usa
Nntp-Posting-Host: 145.1.146.35
Lines: 17

In article <1993Apr1.234031.4950@leland.Stanford.EDU>, bohnert@leland.Stanford.EDU (matthew bohnert) writes:

|> I'm going to be in Cleveland Thursday, April 15 to Sunday, April 18.
|> Does anybody know if the Tribe will be in town on those dates, and
|> if so, who're they playing and if tickets are available?

The tribe will be in town from April 16 to the 19th.
There
```

NLP Example: Transform Docs

NLP Example: Transform Docs

In [60]:

```python
from sklearn.model_selection import train_test_split
docs_ngs_train,docs_ngs_test,y_ngs_train,y_ngs_test = train_test_split(docs_ngs,y_ngs, random_state = 123)

vect = TfidfVectorizer(lowercase=True,
                       min_df=5,           # occur in at least 5 documents
                       max_df=0.8,         # occur in at most 80% of documents
                       token_pattern=r'\b\S\S+\b',  # tokens of at least 2 non-space characters
                       ngram_range=(1,1),  # only unigrams
                       use_idf=False,      # term frequency counts instead of tf-idf
                       norm=None           # do not normalize
                      )
X_ngs_train = vect.fit_transform(docs_ngs_train)
X_ngs_train.shape
```

Out[60]:

```
(897, 3660)
```

NLP Example: Transform Docs

In [60]:

```
from sklearn.model_selection import train_test_split
docs_ngs_train,docs_ngs_test,y_ngs_train,y_ngs_test = train_test_split(docs_ngs,y_ngs, random_state = 123)

vect = TfidfVectorizer(lowercase=True,
                       min_df=5,             # occur in at least 5 documents
                       max_df=0.8,           # occur in at most 80% of documents
                       token_pattern=r'\b\S\S+\b',  # tokens of at least 2 non-space characters
                       ngram_range=(1,1),    # only unigrams
                       use_idf=False,        # term frequency counts instead of tf-idf
                       norm=None             # do not normalize
                      )
X_ngs_train = vect.fit_transform(docs_ngs_train)
X_ngs_train.shape
```

Out[60]:

```
(897, 3660)
```

In [61]:

```
# first few terms in learned vocabulary
list(vect.vocabulary_.items())[:5]
```

Out[61]:

```
[('john', 1781),
 ('hunter', 1637),
 ('white', 3550),
 ('sox', 3059),
 ('mailing', 2029)]
```

NLP Example: Transform Docs

In [60]:

```python
from sklearn.model_selection import train_test_split
docs_ngs_train,docs_ngs_test,y_ngs_train,y_ngs_test = train_test_split(docs_ngs,y_ngs, random_state = 123)

vect = TfidfVectorizer(lowercase=True,
                       min_df=5,           # occur in at least 5 documents
                       max_df=0.8,         # occur in at most 80% of documents
                       token_pattern=r'\b\S\S+\b',  # tokens of at least 2 non-space characters
                       ngram_range=(1,1),  # only unigrams
                       use_idf=False,      # term frequency counts instead of tf-idf
                       norm=None           # do not normalize
                      )
X_ngs_train = vect.fit_transform(docs_ngs_train)
X_ngs_train.shape
```

Out[60]:

```
(897, 3660)
```

In [61]:

```python
# first few terms in learned vocabulary
list(vect.vocabulary_.items())[:5]
```

Out[61]:

```
[('john', 1781),
 ('hunter', 1637),
 ('white', 3550),
 ('sox', 3059),
 ('mailing', 2029)]
```

In [62]:

```python
# first few terms in learned stopword list
list(vect.stop_words_)[:5]
```

Out[62]:

```
["nhl's", 'jozef_stumpel', '3:22', 'sleepers', 'nym,murray']
```

NLP Example: Train and Evaluate Classifier

NLP Example: Train and Evaluate Classifier

In [64]:

```
from sklearn.model_selection import cross_val_score
from sklearn.linear_model import LogisticRegression
from sklearn.dummy import DummyClassifier

scores_dummy = cross_val_score(DummyClassifier(strategy='most_frequent'),X_ngs_train,y_ngs_train)
scores_lr    = cross_val_score(LogisticRegression(),X_ngs_train,y_ngs_train)

print(f'dummy cv accuracy: {scores_dummy.mean().round(2):0.2f} +- {scores_dummy.std().round(2):0.2f}')
print(f'lr    cv accuracy: {scores_lr.mean().round(2):0.2f} +- {scores_lr.std().round(2):0.2f}')
```

```
dummy cv accuracy: 0.52 +- 0.00
lr    cv accuracy: 0.96 +- 0.01
```

NLP Example: Using Pipeline

NLP Example: Using Pipeline

In [65]:

```
from sklearn.pipeline import Pipeline

# Recall: use Pipeline instead of make_pipeline to add names to the steps
#  (name,object) tuple pairs for each step
pipe_ngs1 = Pipeline([('vect',TfidfVectorizer(lowercase=True,
                                              min_df=5,
                                              max_df=0.8,
                                              token_pattern=r'\b\S\S+\b',
                                              ngram_range=(1,1),
                                              use_idf=False,
                                              norm=None )
                      ),
                      ('lr',LogisticRegression())
                     ])

pipe_ngs1.fit(docs_ngs_train,y_ngs_train) # pass in docs, not transformed X

score_ngs1 = pipe_ngs1.score(docs_ngs_train,y_ngs_train).round(2)
print(f'lr pipeline accuracy on training set: {score_ngs1:0.3f}')
```

```
lr pipeline accuracy on training set: 1.000
```

NLP Example: Using Pipeline

In [65]:

```
from sklearn.pipeline import Pipeline

# Recall: use Pipeline instead of make_pipeline to add names to the steps
#  (name,object) tuple pairs for each step
pipe_ngs1 = Pipeline([('vect',TfidfVectorizer(lowercase=True,
                                              min_df=5,
                                              max_df=0.8,
                                              token_pattern=r'\b\S\S+\b',
                                              ngram_range=(1,1),
                                              use_idf=False,
                                              norm=None )
                      ),
                      ('lr',LogisticRegression())
                     ])

pipe_ngs1.fit(docs_ngs_train,y_ngs_train) # pass in docs, not transformed X

score_ngs1 = pipe_ngs1.score(docs_ngs_train,y_ngs_train).round(2)
print(f'lr pipeline accuracy on training set: {score_ngs1:0.3f}')
```

```
lr pipeline accuracy on training set: 1.000
```

In [66]:

```
scores_ngs1 = cross_val_score(pipe_ngs1,docs_ngs_train,y_ngs_train)
print(f'lr pipeline cv accuracy: {scores_ngs1.mean().round(2):0.2f} +- {scores_ngs1.std().round(2):0.2f}')
```

```
lr pipeline cv accuracy: 0.96 +- 0.01
```

NLP Example: Using Pipeline

In [65]:

```
from sklearn.pipeline import Pipeline

# Recall: use Pipeline instead of make_pipeline to add names to the steps
#  (name,object) tuple pairs for each step
pipe_ngs1 = Pipeline([('vect',TfidfVectorizer(lowercase=True,
                                              min_df=5,
                                              max_df=0.8,
                                              token_pattern=r'\b\S\S+\b',
                                              ngram_range=(1,1),
                                              use_idf=False,
                                              norm=None )
                      ),
                      ('lr',LogisticRegression())
                     ])

pipe_ngs1.fit(docs_ngs_train,y_ngs_train) # pass in docs, not transformed X

score_ngs1 = pipe_ngs1.score(docs_ngs_train,y_ngs_train).round(2)
print(f'lr pipeline accuracy on training set: {score_ngs1:0.3f}')
```

```
lr pipeline accuracy on training set: 1.000
```

In [66]:

```
scores_ngs1 = cross_val_score(pipe_ngs1,docs_ngs_train,y_ngs_train)
print(f'lr pipeline cv accuracy: {scores_ngs1.mean().round(2):0.2f} +- {scores_ngs1.std().round(2):0.2f}')
```

```
lr pipeline cv accuracy: 0.96 +- 0.01
```

In [67]:

```
list(pipe_ngs1['vect'].get_feature_names_out())[-5:]
```

Out[67]:

```
['zero', 'zhamnov', 'zhitnik', 'zone', 'zubov']
```

NLP Example: Add Feature Selection

NLP Example: Add Feature Selection

In [68]:

```
from sklearn.feature_selection import SelectFromModel,SelectPercentile

pipe_ngs2 = Pipeline([('vect',TfidfVectorizer(lowercase=True,
                                              min_df=5,
                                              max_df=0.8,
                                              token_pattern='\\b\\S\\S+\\b',
                                              ngram_range=(1,1),
                                              use_idf=False,
                                              norm=None )
                      ),
                      ('fs',SelectFromModel(estimator=LogisticRegression(C=1.0,
                                                                         penalty='l1',
                                                                         solver='liblinear',
                                                                         max_iter=1000,
                                                                         random_state=123
                                                                        ))),
                      ('lr',LogisticRegression(max_iter=10000))
                     ])

pipe_ngs2.fit(docs_ngs_train,y_ngs_train)
print(f'pipeline accuracy on training set: {pipe_ngs2.score(docs_ngs_train,y_ngs_train).round(2):0.2f}')

scores_ngs2 = cross_val_score(pipe_ngs2,docs_ngs_train,y_ngs_train)
print(f'pipeline cv accuracy             : {scores_ngs2.mean().round(2):0.2f} +- {scores_ngs2.std().round(2):0.2f}')
```

```
pipeline accuracy on training set: 1.00
pipeline cv accuracy             : 0.94 +- 0.02
```

NLP Example: Grid Search with Feature Selection

NLP Example: Grid Search with Feature Selection

In [69]:

```
%%time
# NOTE: this may take a minute or so
params_ngs2 = {'vect__use_idf':[True,False],
               'vect__ngram_range':[(1,1),(2,2)],
               'fs__estimator__C':[10,1000],
               'lr__C':[.01,1,100]}

gscv_ngs = GridSearchCV(pipe_ngs2, params_ngs2, cv=2, n_jobs=-1).fit(docs_ngs_train,y_ngs_train)

print(f'gscv_ngs best parameters  : {gscv_ngs.best_params_}')
print(f'gscv_ngs best cv accuracy : {gscv_ngs.best_score_.round(2):0.2f}')
print(f'gscv_ngs test set accuracy: {gscv_ngs.score(docs_ngs_test,y_ngs_test).round(2):0.2f}')
```

```
gscv_ngs best parameters  : {'fs__estimator__C': 1000, 'lr__C': 0.01, 'vect__ngram_range': (1, 1), 'vect__use_idf': True}
gscv_ngs best cv accuracy : 0.96
gscv_ngs test set accuracy: 0.97
CPU times: user 330 ms, sys: 29.2 ms, total: 360 ms
Wall time: 50.9 s
```

Sentiment Analysis and sklearn

- determine sentiment/opinion from unstructured test
- usually positive/negative, but is domain specific
- can be treated as a classification task (with a target, using all of the tools we know)
- can also be treated as a linguistic task (sentence parsing)

- Example: determine sentiment of movie reviews
- see sentiment_analysis_example.ipynb

Topic Modeling

- What topics are our documents composed of?
- How much of each topic does each document contain?
- Can we represent documents using topic weights? (dimensionality reduction!)

Topic Modeling

- What topics are our documents composed of?
- How much of each topic does each document contain?
- Can we represent documents using topic weights? (dimensionality reduction!)

- What is topic modeling?
- How does **Latent Dirichlet Allocation (LDA)** work?
- How to train and use LDA with sklearn?

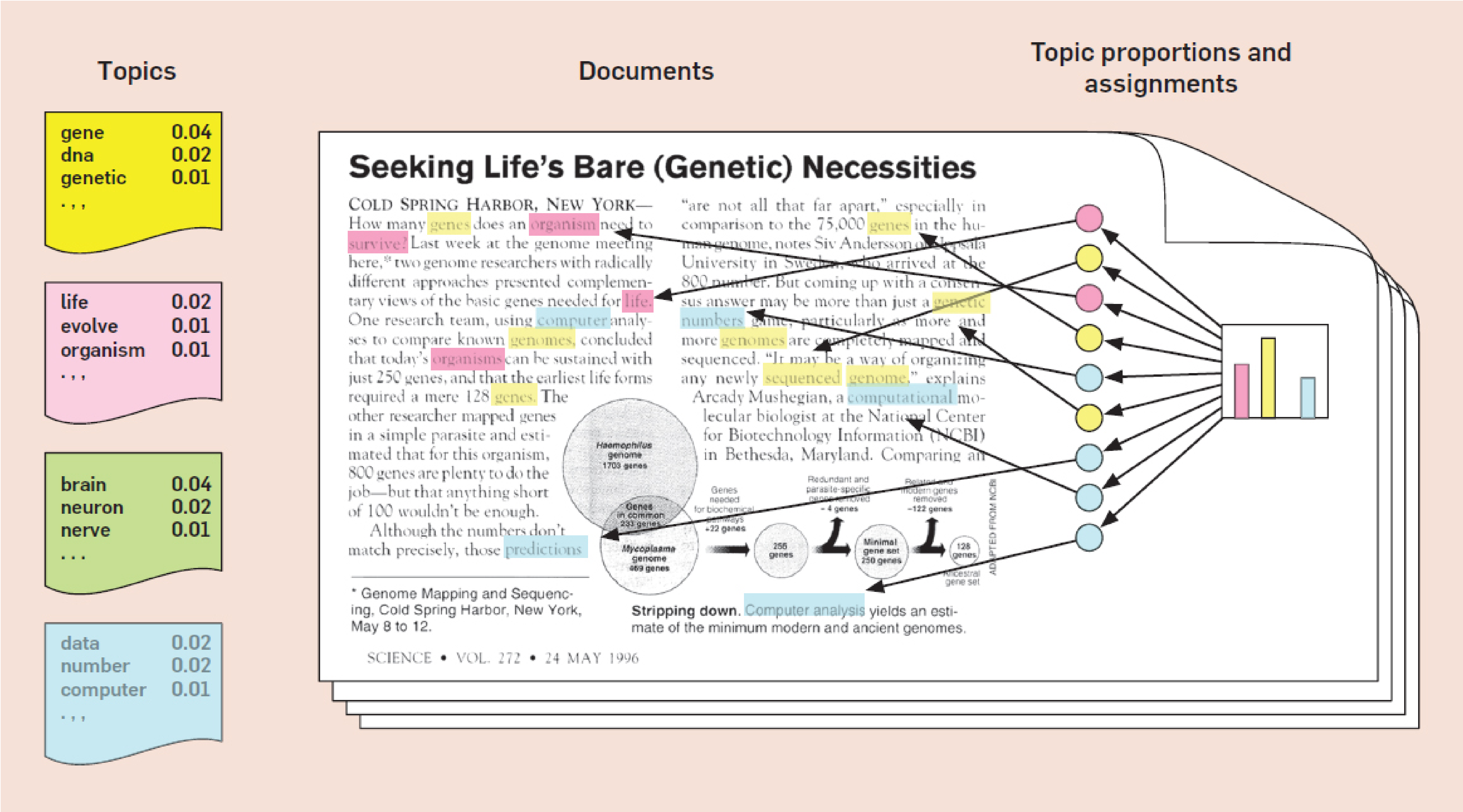
What is Topic Modeling?

- **topic:** a collection of related words
- A document can be composed of several topics

- Given a collection of documents, we can ask:
    - **What terms make up each topic?** (per topic term distribution)
    - **What topics make up each document?** (per document topic distribution)

Topic Modeling with Latent Dirichlet Allocation (LDA)

- Unsupervised method for determining topics and topic assignments



From David Blei

Two Important Matrices Learned by LDA

- the **per topic term distributions** aka φ (phi)

Two Important Matrices Learned by LDA

- the **per topic term distributions** aka φ (phi)

In [70]:

```
topics = ['topic1','topic2']
vocab = ['cat','baseball','play']
phi = pd.DataFrame([[0.4,.2,.4],[0.2,.4,.4]],columns=vocab,index=topics)
phi
```

Out[70]:

	cat	baseball	play
topic1	0.4	0.2	0.4
topic2	0.2	0.4	0.4

Two Important Matrices Learned by LDA

- the **per topic term distributions** aka φ (phi)

In [70]:
```
topics = ['topic1','topic2']
vocab = ['cat','baseball','play']
phi = pd.DataFrame([[0.4,.2,.4],[0.2,.4,.4]],columns=vocab,index=topics)
phi
```

Out[70]:

	cat	baseball	play
topic1	0.4	0.2	0.4
topic2	0.2	0.4	0.4

- the **per document TOPIC distributions** aka θ (theta)

Two Important Matrices Learned by LDA

- the **per topic term distributions** aka φ (phi)

In [70]:
```
topics = ['topic1','topic2']
vocab = ['cat','baseball','play']
phi = pd.DataFrame([[0.4,.2,.4],[0.2,.4,.4]],columns=vocab,index=topics)
phi
```

Out[70]:

	cat	baseball	play
topic1	0.4	0.2	0.4
topic2	0.2	0.4	0.4

- the **per document TOPIC distributions** aka θ (theta)

In [71]:
```
topics = ['topic1','topic2']
docs = ['doc1','doc2']
theta = pd.DataFrame([[0.1,.9],[.5,.5]],columns=topics,index=docs)
theta
```

Out[71]:

	topic1	topic2
doc1	0.1	0.9
doc2	0.5	0.5

Topic Modeling: Example

- Given the data and the number of topics we want

Topic Modeling: Example

- Given the data and the number of topics we want

In [72]:

```python
corpus = ['the dog and cat played tennis',
          'tennis and baseball are sports',
          'a dog or a cat can be a pet']

M = 3 # the number of documents

vocab = ['baseball','cat','dog','pet','played','tennis']

V = len(vocab) # size of vocabulary

K = 2 # our guess about the number of topics

print(f'{M = :}\n{V = :}\n{K = :}')
```

```
M = 3
V = 6
K = 2
```

Topic Modeling: Example

- Guessing some **per topic term distributions** (φ) given the documents and vocab

Topic Modeling: Example

- Guessing some **per topic term distributions** (φ) given the documents and vocab

In [73]:
```
print(vocab)
```

```
['baseball', 'cat', 'dog', 'pet', 'played', 'tennis']
```

Topic Modeling: Example

- Guessing some **per topic term distributions** (φ) given the documents and vocab

In [73]:
```python
print(vocab)
```

```
['baseball', 'cat', 'dog', 'pet', 'played', 'tennis']
```

In [74]:
```python
# the probability of each term given topic 1 (high for sports terms)
topic_1 = [.33,   0,   0,   0, .33, .33]

# the probability of each term given topic 2 (high for pet terms)
topic_2 = [  0, .25, .25, .25, .25,   0]

# per topic term distributions
phi = pd.DataFrame([topic_1, topic_2],columns=vocab,
                   index=['topic_'+str(x) for x in range(1,K+1)])

phi
```

Out[74]:

	baseball	cat	dog	pet	played	tennis
topic_1	0.33	0.00	0.00	0.00	0.33	0.33
topic_2	0.00	0.25	0.25	0.25	0.25	0.00

Topic Modeling: Example

- Guessing the **per document topic distributions** θ given the **topics**

Topic Modeling: Example

- Guessing the **per document topic distributions** θ given the **topics**

In [75]:
```
# Given our guess about phi
display(phi)
# And the corpus
corpus
```

	baseball	cat	dog	pet	played	tennis
topic_1	0.33	0.00	0.00	0.00	0.33	0.33
topic_2	0.00	0.25	0.25	0.25	0.25	0.00

Out[75]:
```
['the dog and cat played tennis',
 'tennis and baseball are sports',
 'a dog or a cat can be a pet']
```

Topic Modeling: Example

- Guessing the **per document topic distributions** θ given the **topics**

In [75]:

```
# Given our guess about phi
display(phi)
# And the corpus
corpus
```

	baseball	cat	dog	pet	played	tennis
topic_1	0.33	0.00	0.00	0.00	0.33	0.33
topic_2	0.00	0.25	0.25	0.25	0.25	0.00

Out[75]:

```
['the dog and cat played tennis',
 'tennis and baseball are sports',
 'a dog or a cat can be a pet']
```

In [76]:

```
# generate a guess about per document topic distributions
theta = pd.DataFrame([[.50, .50],
                      [.99, .01],
                      [.01, .99]],
                     columns=['topic_'+str(x) for x in range(1,K+1)],
                     index=['doc_'+str(x) for x in range(1,M+1)])
theta
```

Out[76]:

	topic_1	topic_2
doc_1	0.50	0.50
doc_2	0.99	0.01
doc_3	0.01	0.99

Topic Modeling With LDA

- Given
  - a set of documents
  - a number of topics K
- Learn
  - the **per topic term distributions** φ **(phi)**, size: $K \times V$
  - the **per document topic distributions** θ **(theta)**, size: $M \times K$
- How to learn φ and θ:
  - Latent Dirichlet Allocation (LDA)
  - generative statistical model
  - Blei, D., Ng, A., Jordan, M. Latent Dirichlet allocation. J. Mach. Learn. Res. 3 (Jan 2003)

Topic Modeling With LDA

- Uses for φ (phi), the per topic term distributions:
  - infering labels for topics
  - word clouds

- Uses for θ (theta), the per document topic distributions:
  - dimensionality reduction
  - clustering
  - similarity

LDA with sklearn

LDA with sklearn

In [77]:

```
# load data from all 20 newsgroups
newsgroups = fetch_20newsgroups()
ngs_all = newsgroups.data
len(ngs_all)
```

Out[77]: 11314

LDA with sklearn

In [77]:
```
# load data from all 20 newsgroups
newsgroups = fetch_20newsgroups()
ngs_all = newsgroups.data
len(ngs_all)
```

Out[77]: 11314

In [78]:
```
# transform documents using tf-idf
tfidf = TfidfVectorizer(token_pattern=r'\b[a-zA-Z0-9-][a-zA-Z0-9-]+\b',min_df=50, max_df=.2)
X_tfidf = tfidf.fit_transform(ngs_all)
X_tfidf.shape
```

Out[78]: (11314, 4256)

LDA with sklearn

In [77]:
```
# load data from all 20 newsgroups
newsgroups = fetch_20newsgroups()
ngs_all = newsgroups.data
len(ngs_all)
```

Out[77]:
```
11314
```

In [78]:
```
# transform documents using tf-idf
tfidf = TfidfVectorizer(token_pattern=r'\b[a-zA-Z0-9-][a-zA-Z0-9-]+\b',min_df=50, max_df=.2)
X_tfidf = tfidf.fit_transform(ngs_all)
X_tfidf.shape
```

Out[78]:
```
(11314, 4256)
```

In [79]:
```
feature_names = tfidf.get_feature_names()
print(feature_names[:10])
print(feature_names[-10:])
```

```
['00', '000', '01', '02', '03', '04', '05', '06', '07', '08']
['yours', 'yourself', 'ysu', 'zealand', 'zero', 'zeus', 'zip', 'zone', 'zoo', 'zuma']
```

LDA with sklearn Cont.

LDA with sklearn Cont.

In [80]:

```python
from sklearn.decomposition import LatentDirichletAllocation

# create model with 20 topics
lda = LatentDirichletAllocation(n_components=20,  # the number of topics
                                n_jobs=-1,         # use all cpus
                                random_state=123) # for reproducability

# learn phi (lda.components_) and theta (X_lda)
# this will take a while!
X_lda = lda.fit_transform(X_tfidf)
```

LDA with sklearn Cont.

In [80]:

```
from sklearn.decomposition import LatentDirichletAllocation

# create model with 20 topics
lda = LatentDirichletAllocation(n_components=20,  # the number of topics
                                n_jobs=-1,         # use all cpus
                                random_state=123) # for reproducability

# learn phi (lda.components_) and theta (X_lda)
# this will take a while!
X_lda = lda.fit_transform(X_tfidf)
```

In [81]:

```
ngs_all[100][:100]
```

Out[81]:

```
'From: tchen@magnus.acs.ohio-state.edu (Tsung-Kun Chen)\nSubject: ** Software forsale (lots) **\nNntp-P'
```

LDA with sklearn Cont.

In [80]:

```
from sklearn.decomposition import LatentDirichletAllocation

# create model with 20 topics
lda = LatentDirichletAllocation(n_components=20,  # the number of topics
                                n_jobs=-1,         # use all cpus
                                random_state=123) # for reproducability

# learn phi (lda.components_) and theta (X_lda)
# this will take a while!
X_lda = lda.fit_transform(X_tfidf)
```

In [81]:

```
ngs_all[100][:100]
```

Out[81]:

```
'From: tchen@magnus.acs.ohio-state.edu (Tsung-Kun Chen)\nSubject: ** Software forsale (lots) **\nNntp-P'
```

In [82]:

```
X_lda[100].round(2) # lda representation of document_100
```

Out[82]:

```
array([0.01, 0.01, 0.01, 0.01, 0.08, 0.01, 0.01, 0.01, 0.01, 0.01, 0.01,
       0.01, 0.01, 0.01, 0.79, 0.01, 0.01, 0.01, 0.01, 0.01])
```

LDA with sklearn Cont.

In [80]:
```
from sklearn.decomposition import LatentDirichletAllocation

# create model with 20 topics
lda = LatentDirichletAllocation(n_components=20,  # the number of topics
                                n_jobs=-1,        # use all cpus
                                random_state=123) # for reproducability

# learn phi (lda.components_) and theta (X_lda)
# this will take a while!
X_lda = lda.fit_transform(X_tfidf)
```

In [81]:
```
ngs_all[100][:100]
```

Out[81]:
```
'From: tchen@magnus.acs.ohio-state.edu (Tsung-Kun Chen)\nSubject: ** Software forsale (lots) **\nNntp-P'
```

In [82]:
```
X_lda[100].round(2) # lda representation of document_100
```

Out[82]:
```
array([0.01, 0.01, 0.01, 0.01, 0.08, 0.01, 0.01, 0.01, 0.01, 0.01, 0.01,
       0.01, 0.01, 0.01, 0.79, 0.01, 0.01, 0.01, 0.01, 0.01])
```

In [83]:
```
# Note: since this is unsupervised, these numbers may change
np.argsort(X_lda[100])[::-1][:3] # the top topics of document_100
```

Out[83]:
```
array([14,  4, 11])
```

LDA: Per Topic Term Distributions

LDA: Per Topic Term Distributions

In [85]:
```
print_top_words(lda,feature_names,5)
```

```
Topic  0: udel rpi princeton phoenix delaware
Topic  1: turkish armenian armenians armenia serdar
Topic  2: god his jesus because why
Topic  3: pitt geb gordon banks cadre
Topic  4: cwru cleveland ohio-state magnus acs
Topic  5: rit isc rochester ring testing
Topic  6: psu psuvm penn gay men
Topic  7: caltech keith sgi livesey cco
Topic  8: stratus wpi sw jewish cdt
Topic  9: msg gatech prism utexas georgia
Topic 10: fraser sfu portal simon craig
Topic 11: window mit motif lcs x11r5
Topic 12: indiana duke lehigh captain ns1
Topic 13: henry toronto uga ai zoo
Topic 14: windows card drive thanks dos
Topic 15: columbia cunixb gerald cc alchemy
Topic 16: ca new cs use anyone
Topic 17: alaska ti dseg aurora nsmca
Topic 18: radar detector su 4th quote
Topic 19: cramer optilink clayton virginia carleton
```

LDA Review

- What did we learn?
    - per document topic distributions
    - per topic term distributions

- What can we use this for?
    - Dimensionality Reduction/Feature Extraction!
    - investigate topics (much like PCA components)

Other NLP Features

- Part of Speech tags
- Dependency Parsing
- Entity Detection
- Word Vectors
- See spaCy!

Using spaCy for NLP

Using spaCy for NLP

In [86]:

```python
import spacy

# uncomment the line below the first time you run this cell
#%run -m spacy download en_core_web_sm
try:

    nlp = spacy.load("en_core_web_sm")

except OSError as e:
    print('Need to run the following line in a new cell:')
    print('%run -m spacy download en_core_web_sm')
    print('or the following line from the commandline with eods-f20 activated:')
    print('python -m spacy download en_core_web_sm')

parsed = nlp("N.Y.C. isn't in New Jersey.")
'|'.join([token.text for token in parsed])
```

```
2023-11-27 21:43:08.314505: I tensorflow/core/platform/cpu_feature_guard.cc:193] This TensorFlow binary is optimized with oneAP
I Deep Neural Network Library (oneDNN) to use the following CPU instructions in performance-critical operations:  AVX2 FMA
To enable them in other operations, rebuild TensorFlow with the appropriate compiler flags.
```

Out[86]:

```
"N.Y.C.|is|n't|in|New|Jersey|."
```

spaCy: Part of Speech Tagging

spaCy: Part of Speech Tagging

In [87]:

```
doc = nlp("Apple is looking at buying U.K. startup for $1 billion.")

print(f"{'text':7s} {'lemma':7s} {'pos':5s} {'is_stop'}")
print('-'*30)
for token in doc:
    print(f'{token.text:7s} {token.lemma_:7s} {token.pos_:5s} {token.is_stop}')
```

```
text    lemma   pos   is_stop
------------------------------
Apple   Apple   PROPN False
is      be      AUX   True
looking look    VERB  False
at      at      ADP   True
buying  buy     VERB  False
U.K.    U.K.    PROPN False
startup startup VERB  False
for     for     ADP   True
$       $       SYM   False
1       1       NUM   False
billion billion NUM   False
.       .       PUNCT False
```

spaCy: Part of Speech Tagging

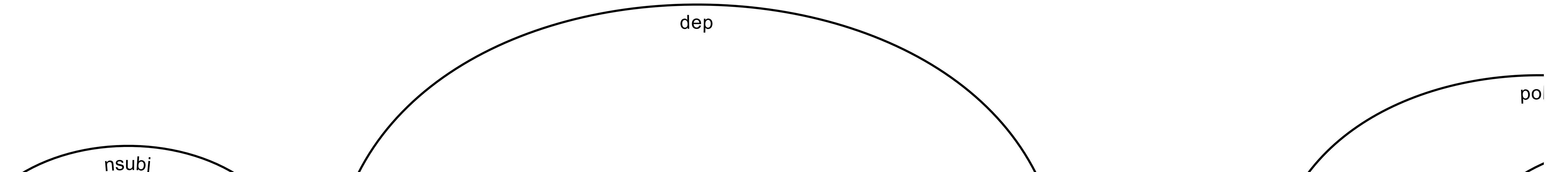
In [87]:

```python
doc = nlp("Apple is looking at buying U.K. startup for $1 billion.")

print(f"{'text':7s} {'lemma':7s} {'pos':5s} {'is_stop'}")
print('-'*30)
for token in doc:
    print(f'{token.text:7s} {token.lemma_:7s} {token.pos_:5s} {token.is_stop}')
```

```
text    lemma   pos   is_stop
------------------------------
Apple   Apple   PROPN False
is      be      AUX   True
looking look    VERB  False
at      at      ADP   True
buying  buy     VERB  False
U.K.    U.K.    PROPN False
startup startup VERB  False
for     for     ADP   True
$       $       SYM   False
1       1       NUM   False
billion billion NUM   False
.       .       PUNCT False
```

In [88]:

```python
from spacy import displacy
displacy.render(doc, style="dep")
```



spaCy: Part of Speech Tagging

spaCy: Entity Detection

spaCy: Entity Detection

```
In [89]:  1 [(ent.text,ent.label_) for ent in doc.ents]

Out[89]: [('Apple', 'ORG'), ('U.K.', 'GPE'), ('$1 billion', 'MONEY')]
```

spaCy: Entity Detection

In [89]:
```
[(ent.text,ent.label_) for ent in doc.ents]
```

Out[89]:
```
[('Apple', 'ORG'), ('U.K.', 'GPE'), ('$1 billion', 'MONEY')]
```

In [90]:
```
displacy.render(doc, style="ent")
```

Apple ORG is looking at buying U.K. GPE startup for $1 billion MONEY .

spaCy: Word Vectors

- word2vec
- shallow neural net
- predict a word given the surrounding context (SkipGram or CBOW)
- words used in similar context should have similar vectors

spaCy: Word Vectors

- word2vec
- shallow neural net
- predict a word given the surrounding context (SkipGram or CBOW)
- words used in similar context should have similar vectors

In [91]:

```
# Need either the _md or _lg models to get vector information
# Note: this takes a while!
# %run -m spacy download en_core_web_md
```

spaCy: Word Vectors

- word2vec
- shallow neural net
- predict a word given the surrounding context (SkipGram or CBOW)
- words used in similar context should have similar vectors

In [91]:
```
# Need either the _md or _lg models to get vector information
# Note: this takes a while!
# %run -m spacy download en_core_web_md
```

In [92]:
```
nlp = spacy.load('en_core_web_md') # _lg has a larger vocabulary

doc = nlp('Baseball is played on a diamond.')
doc[0].text, doc[0].vector.shape, list(doc[0].vector[:3])
```

Out[92]:
```
('Baseball', (300,), [0.55838, 0.42791, -0.11687])
```

spaCy: Multiple Documents

spaCy: Multiple Documents

In [93]:

```
# Use nlp.pipe to transform multiple docs at once
docs = list(nlp.pipe(['Baseball is played on a diamond.',
                      'Hockey is played on ice.',
                      'Diamonds are clear as ice.']))
```

spaCy: Multiple Documents

In [93]:

```python
# Use nlp.pipe to transform multiple docs at once
docs = list(nlp.pipe(['Baseball is played on a diamond.',
                      'Hockey is played on ice.',
                      'Diamonds are clear as ice.']))
```

In [94]:

```python
# using average of token vectors for each document.
np.array([['{:.2f}'.format(docs[i].similarity(docs[j])) for j in range(3)]
          for i in range(3)])
```

Out[94]:

```
array([['1.00', '0.85', '0.76'],
       ['0.85', '1.00', '0.77'],
       ['0.76', '0.77', '1.00']], dtype='<U4')
```

Learning Sequences

- Hidden Markov Models
- Conditional Random Fields
- Recurrant Neural Networks
- LSTM
- GPT3
- BERT
- Transformers (MLP 16.4)

NLP Review

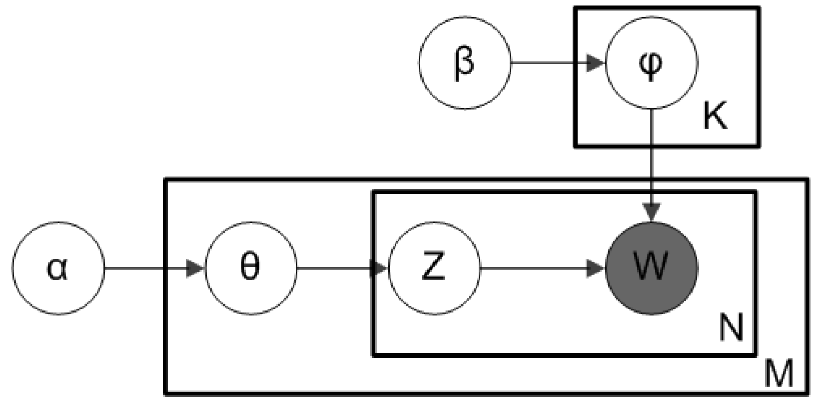
- corpus, tokens, vocabulary, terms, n-grams, stopwords
- tokenization
- term frequency (TF), document frequency (DF)
- TF vs TF-IDF
- sentiment analysis
- topic modeling

- POS
- Dependency Parsing
- Entity Extraction
- Word Vectors

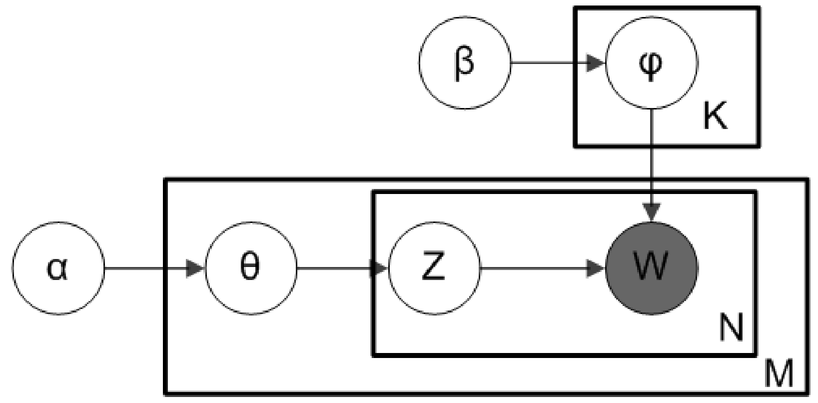
Questions?

Appendix: LDA Plate Diagram

Appendix: LDA Plate Diagram



Appendix: LDA Plate Diagram



K : number of topics

φ : per topic term distributions

β : parameters for word distribution die factory, length = V (size of vocab)

M : number of documents

N : number of words/tokens in each document

θ : per document topic distributions

α : parameters for topic die factory, length = K (number of topics)

z : topic indexes

w : observed tokens