

Spiced Data Science

Week 4: Text Classification



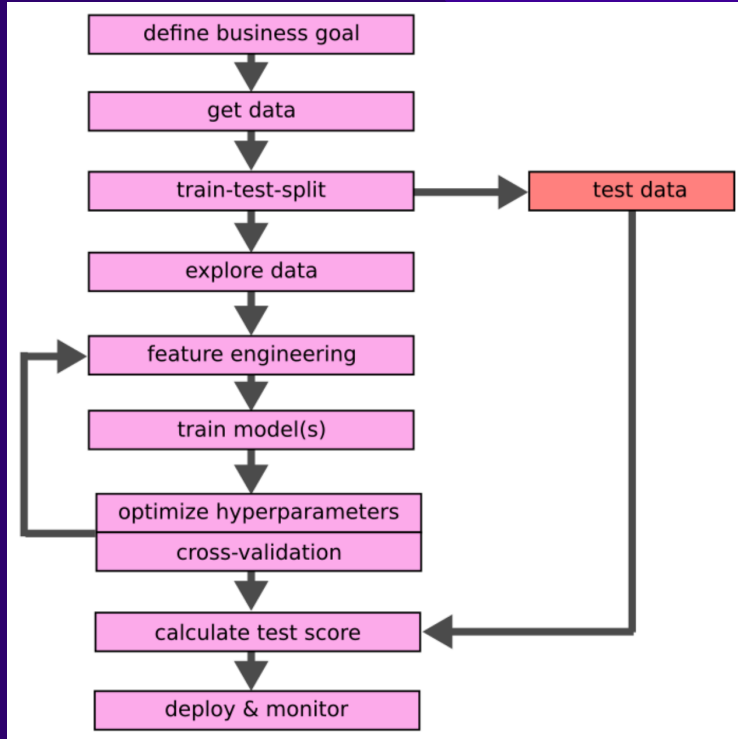
Project of the Week:

```
🍺 $ python song_classifier.py -t 'inconceivable hummingbird strawberries'
```

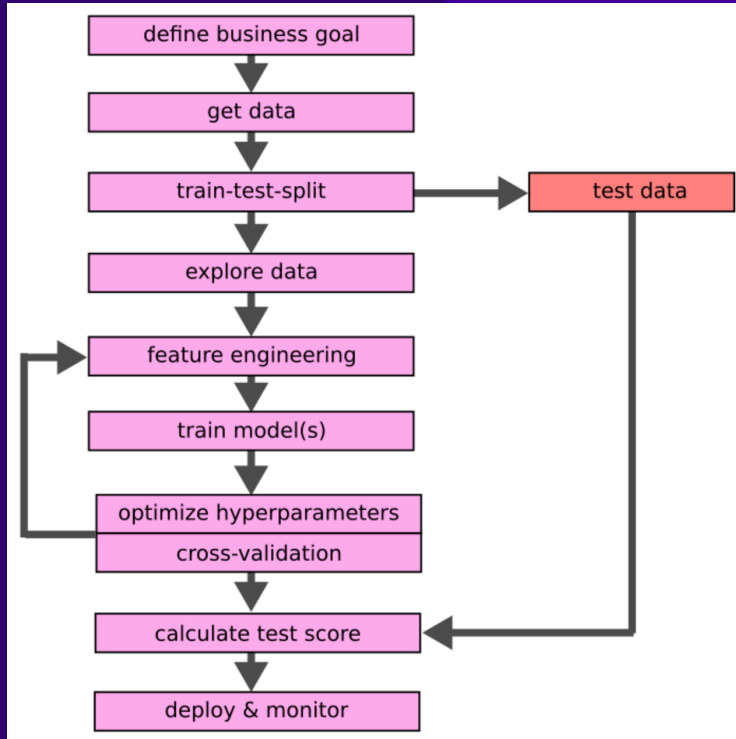
The model predicts that the sample text belongs to Fleet Foxes, with a probability of 80.3%.

```
🍺 $ |
```

Roadmap for Text Classification



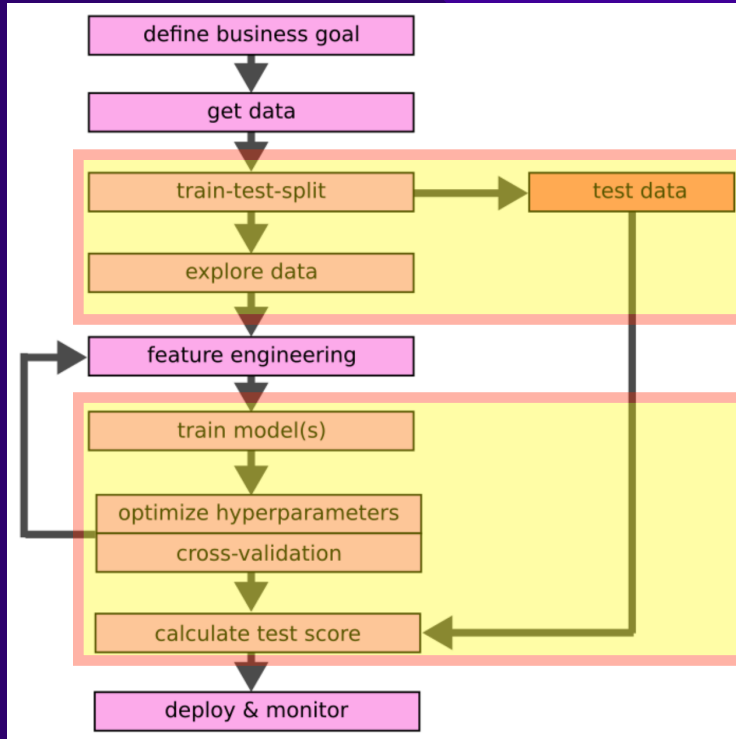
Roadmap for Text Classification



Your Schedule:

	AM	PM
Mon.	Career Workshop	Graduation
Tues.	Web Scraping	Regular Expressions
Wed.	Parsing HTML	Bag of Words
Thurs.	Class Imbalance	Writing Python Functions + CLI
Fri.	Naive Bayes	Project Review / Presentations

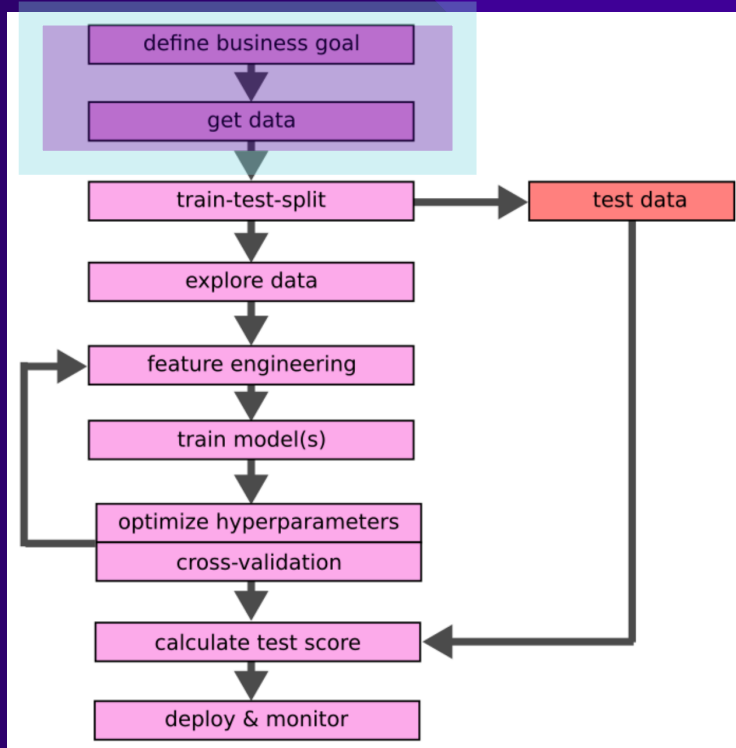
Roadmap for Text Classification



Your Schedule:

	AM	PM
Mon.	Career Workshop	Graduation
Tues.	Web Scraping	Regular Expressions
Wed.	Parsing HTML	Bag of Words
Thurs.	Class Imbalance	Writing Python Functions + CLI
Fri.	Naive Bayes	Project Review / Presentations

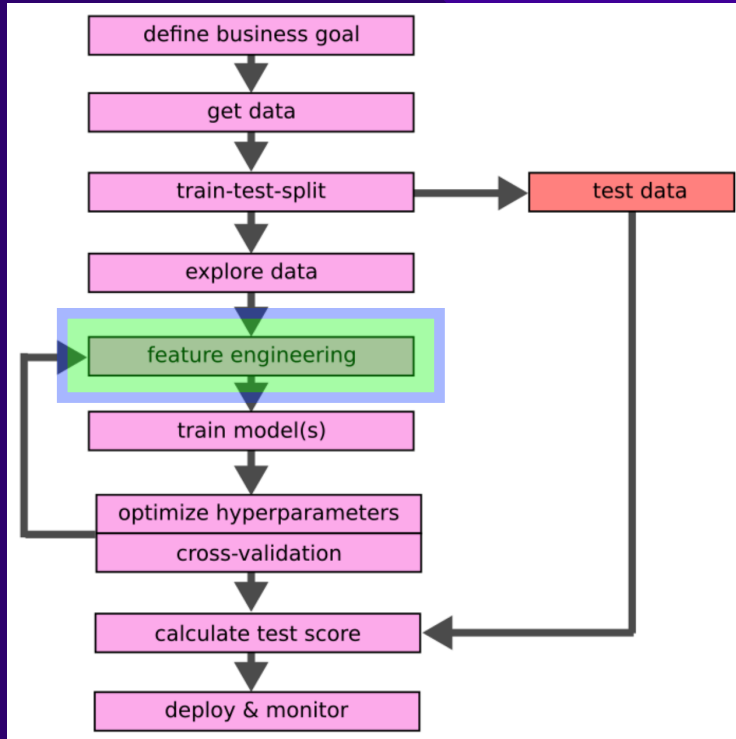
Roadmap for Text Classification



Your Schedule:

	AM	PM
Mon.	Career Workshop	Graduation
Tues.	Web Scraping	Regular Expressions
Wed.	Parsing HTML	Bag of Words
Thurs.	Class Imbalance	Writing Python Functions + CLI
Fri.	Naive Bayes	Project Review / Presentations

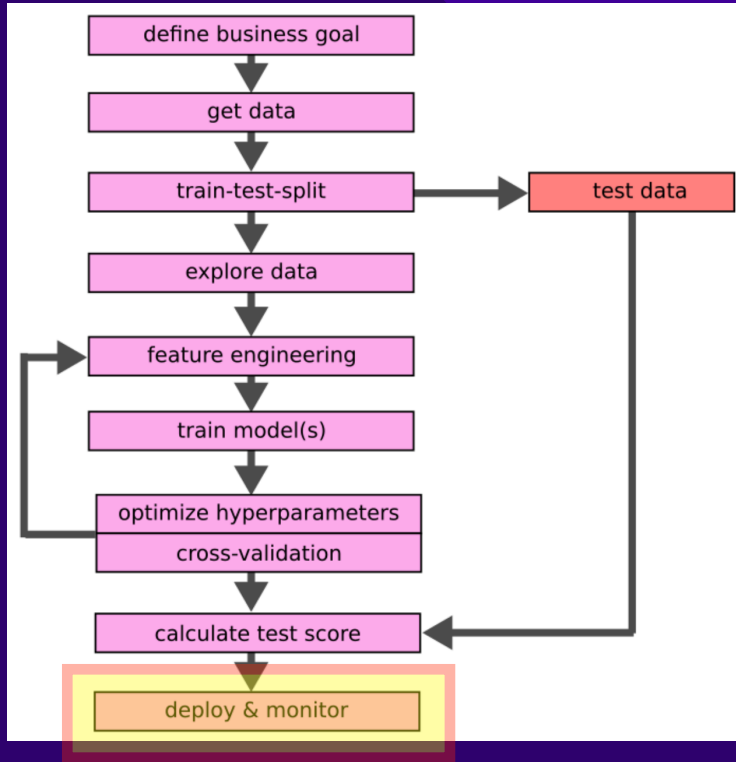
Roadmap for Text Classification



Your Schedule:

	AM	PM
Mon.	Career Workshop	Graduation
Tues.	Web Scraping	Regular Expressions
Wed.	Parsing HTML	Bag of Words
Thurs.	Class Imbalance	Writing Python Functions + CLI
Fri.	Naive Bayes	Project Review / Presentations

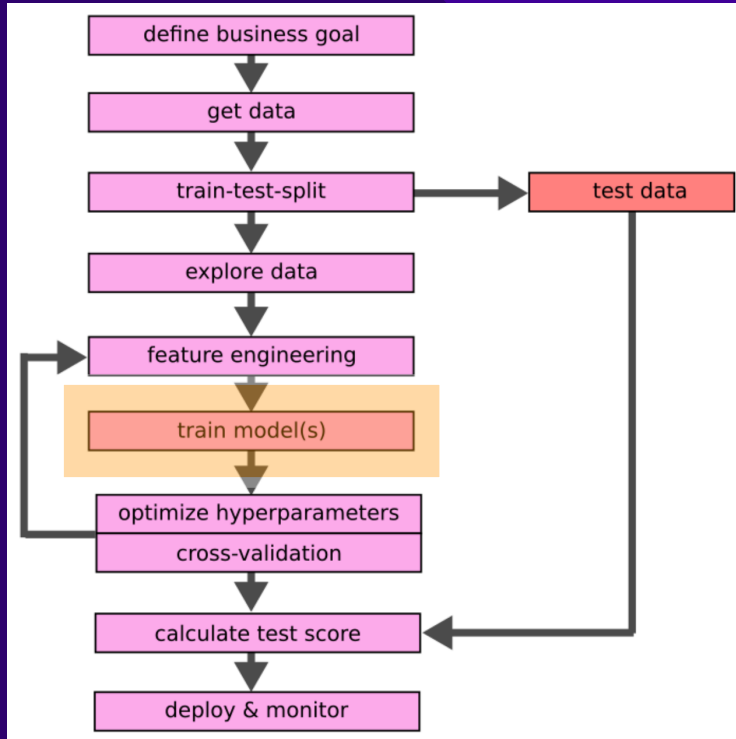
Roadmap for Text Classification



Your Schedule:

	AM	PM
Mon.	Career Workshop	Graduation
Tues.	Web Scraping	Regular Expressions
Wed.	Parsing HTML	Bag of Words
Thurs.	Class Imbalance	Writing Python Functions + CLI
Fri.	Naive Bayes	Project Review / Presentations

Roadmap for Text Classification



Your Schedule:

	AM	PM
Mon.	Career Workshop	Graduation
Tues.	Web Scraping	Regular Expressions
Wed.	Parsing HTML	Bag of Words
Thurs.	Class Imbalance	Writing Python Functions + CLI
Fri.	Naive Bayes	Project Review / Presentations

More Detailed Roadmap

Scrape and parse lyrics



Get text data into tabular form



Further text cleaning



Vectorize text



Train model



Add command-line interface

More Detailed Roadmap

Scrape and parse lyrics



Get text data into tabular form



Further text cleaning



Vectorize text



Train model



Add command-line interface

Tools used:

- requests + Regular Expressions
- requests + BeautifulSoup (Parsing HTML)
- Scrapy (bonus!)

More Detailed Roadmap

Scrape and parse lyrics



Get text data into tabular form



Further text cleaning



Vectorize text



Train model



Add command-line interface

List of Lyrics (X)

["hello darkness my old friend...",
"...we will we will rock you...",
"...i see a little silhouette of a man...",
"you are my fire the one desire..."]

List of Artists (y)

["simon and garfunkel",
"queen",
"queen",
"backstreet boys"]

More Detailed Roadmap

Scrape and parse lyrics



Get text data into tabular form



Further text cleaning



Vectorize text



Train model



Add command-line interface

Tools used:

- Lemmatization
- RegEx / lots of string cleaning!!

More Detailed Roadmap

Scrape and parse lyrics



Get text data into tabular form



Further text cleaning



Vectorize text



Train model



Add command-line interface

Bag of Words (BOW) Lesson:

- CountVectorizer
- TfidfTransformer

- (or simply the TfidfVectorizer)

More Detailed Roadmap

Scrape and parse lyrics



Get text data into tabular form



Further text cleaning



Vectorize text



Train model



Add command-line interface

Bag of Words (BOW) Lesson:

- `CountVectorizer`
- `TfidfTransformer`

- (or simply the `TfidfVectorizer`)

Class Imbalance techniques
(e.g. SMOTE) can then be applied
on the vectorized dataframe.

More Detailed Roadmap

Scrape and parse lyrics



Get text data into tabular form



Further text cleaning



Vectorize text



Train model



Add command-line interface

Lots of Options:

- LogisticRegression
- RandomForest
- GradientBoostingClassifier

- ***NEW: Naive Bayes!***

More Detailed Roadmap

Scrape and parse lyrics



Get text data into tabular form



Further text cleaning



Vectorize text



Train model



Add command-line interface

Important Lessons:

- **Writing Python Functions**
(laying a solid foundation)
- **Command-Line-Interface**

