! **Try again once you are ready**
TO PASS 80% or higher

Try again

GRADE
40%

Analyze Text Data with Yellowbrick

LATEST SUBMISSION GRADE
40%

1. Which of the following are somechallenges of modelling text data? (Select all that apply) — 1 / 1 point

- [x] Text data is very high dimensional.

✓ **Correct**
Correct! E.g. One dimension for every word (token) in the corpus!

- [x] Text data is often sparsely distributed.

✓ **Correct**
Correct! E.g. documents vary in length, most instances (documents) may be mostly zeros.

- [x] Text data has some features that are more important than others.

✓ **Correct**
Correct! Eg. the "of" dimensions vs. the "basketball" dimension when clustering sports articles.

2. In text analysis, instances are entiredocuments, which can vary in length from quotes or tweets to entire books, but whose vectors are always of a uniform length. Each property of the vector representation is called a: — 0 / 1 point

- (•) Target
- ( ) Feature

! **Incorrect**
Incorrect. Please review Tasks 1 and 2 of the hands-on component of this project.

3. You are pre-processing text data to create a sentiment analyzer. The corpus of documents is stored as a column **data** in a pandas DataFrame **corpus**. How do you tokenize the documents using scikit-learn's TfidfVectorizer? — 0 / 1 point

(•)
```
1 from sklearn.feature_extraction.text import TfidfVectorizer
2
3 vectorizer = TfidfVectorizer()
4 tokenized_docs = vectorizer.predict(corpus.data)
```

( )
```
1 from sklearn.feature_extraction.text import TfidfVectorizer
2
3 vectorizer = TfidfVectorizer()
4 tokenized_docs = vectorizer.fit(corpus.data)
```

( )
```
1 from sklearn.feature_extraction.text import TfidfVectorizer
2
3 vectorizer = TfidfVectorizer()
4 tokenized_docs = vectorizer.fit_transform(corpus.data)
```

! **Incorrect**
Incorrect. The predict method is called on fitted scikit-learn models. Please review Task 2 of the hands-on component of this project.

4. Euclidean distance is an ideal choice of metric for sparse text data. — 0 / 1 point

- (•) True

False

Incorrect

Incorrect. The high dimensionality usually associated with sparse data means that the distance of any two objects will likely be a quadratic mean of their lengths. Please review Task 3 of the hands-on component of this project. The Manhattan distance metric (L1 norm) is consistently preferable than the Euclidean distance metric (L2 norm) for high-dimensional data mining.

5. What is the straight-line distance between 2 points in Euclidean (metric) space? **1 / 1 point**

- (•) Euclidean distance
- ( ) Cosine distance (similarity)
- ( ) Manhattan distance

Correct

Correct!

```
tsne = TSNEVisualizer(metric="euclidean")
tsne.fit(docs, labels)
tsne.poof()
```