tf_idf

September 14, 2021

[ ]:
```
from sklearn.feature_extraction.text import TfidfVectorizer
import numpy as np
import wikipedia
from kmeans import KMeans
```

[ ]:
```
titles = [
    'Linear algebra',
    'Data Science',
    'Artificial intelligence',
    'European Central Bank',
    'Financial technology',
    'International Monetary Fund',
    'Basketball',
    'Swimming',
    'Cricket'
]
```

[ ]:
```
def load_data():
    articles = [wikipedia.page(
        title, preload=True).content for title in titles]
    vectorizer = TfidfVectorizer(stop_words={'english'})
    x_train = vectorizer.fit_transform(articles).toarray()
    y_train = np.arange(len(titles))

    return (x_train, y_train), vectorizer
```

[ ]:
```
(x_train, y_train), vectorizer = load_data()
```

[ ]:
```
def main():
    print("Data loaded, Finding Clusters ...")
    k = [4, 8]
    losses = []
    for num_clusters in k:
        kmeans = KMeans(x_train, y_train, num_clusters=num_clusters,
                        seed='cluster', tol=1e-9, max_iter=200)
        kmeans.fit(verbose=False)
        print("Clusters found, printing results ...")
```

```
        losses.append(kmeans.calc_loss())
        print(kmeans.cluster_labels)
```

[ ]:
```
main()
```

```
Data loaded, Finding Clusters …
Total Iterations: 1, Loss: 0.22279518598223452
Clusters found, printing results …
[3 0 2 1 1 1 1 1 1]
Total Iterations: 1, Loss: 0.028789961445557916
Clusters found, printing results …
[4 7 0 6 1 6 5 3 2]
```

[ ]:
```

```