To contextualize or to not contextualize?

> Can we define a topic model that does not rely on the BoW input but instead uses contextual information?

First, we want to check if ZeroShotTM maintains comparable performance to other topic models; if this is true, we can then explore its performance in a cross-lingual setting. Since we use only English text, in this setting we use English representations.

[ ] ↳ *35 cells hidden*

Results

```
npmi = pd.DataFrame.from_dict(NPMI, orient='index')
print("NPMI Coherences on W1 dataset")
npmi.set_axis(["t(50)", "t(100)"], axis = 1)
```

```
NPMI Coherences on W1 dataset
```

	t(50)	t(100)
ZeroShotTM	0.169154	0.135052
CombinedTM	0.179156	0.154274
Neural-ProdLDA	0.171004	0.138252
LDA	-0.061104	-0.158774

Zero-shot Cross-Lingual Topic Modeling

> Can the conxtextualized TM tackle zero-shot cross-lingual topic modeling?

The second dataset (W2) contains 100,000 English documents. We use 99,700 documents as training and consider the remaining 300 documents as the test set. We collect the 300 respective instances in Portuguese, Italian, French, and German.

First, we use SBERT to generate multilingual embeddings as the input of the model. Then we evaluate multilingual topic predictions on the multilingual abstracts in W2.

[ ] ↳ *26 cells hidden*

Results

```
metrics = pd.DataFrame.from_dict(metrics, orient='columns')
print("Match, KL, and Centroid Similarity for 25 and 50 topics on various languages on W2")
metrics
```

```
Match, KL, and Centroid Similarity for 25 and 50 topics on various languages on W2
```

	Mat25	KL25	CD25	Mat50	KL50	CD50
Italian	0.806667	0.125643	0.851479	0.68	0.173317	0.764182
French	0.776667	0.137309	0.834782	0.68	0.178904	0.761205
Portugese	0.810000	0.122204	0.849435	0.70	0.153781	0.780023
German	0.740000	0.141736	0.799654	0.65	0.180723	0.738930