Assignment 4

Part 1

Task 1

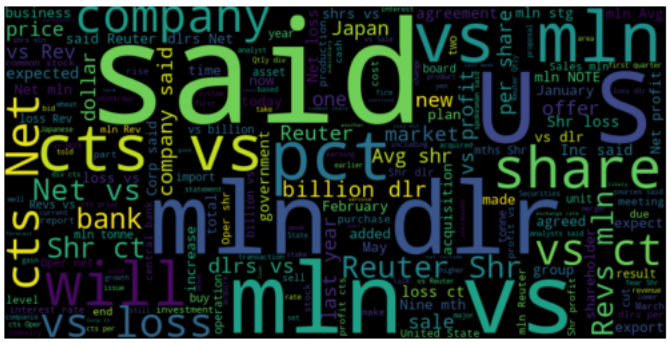
1. **Report the number of samples remaining in the dataset.**

   Number of samples after filtering: 9866

2. **Extract news related to the 10 most frequent topics:**
   a. **Identify and report the 10 topics with the highest number of samples.**
   Sample distribution across top 10 topics:
   earn: 3776
   acq: 2134
   money-fx: 605
   grain: 534
   crude: 517
   trade: 453
   interest: 391
   ship: 284
   wheat: 265
   corn: 209
   b. **Report the number of samples remaining in the dataset and the number of samples across the 10 topics.**
   Number of samples after retaining top 10 topics: 8284

Task 2

1. **Word Cloud**
   a. **Attach a screenshot in your write-up. What do you observe from the word cloud?**

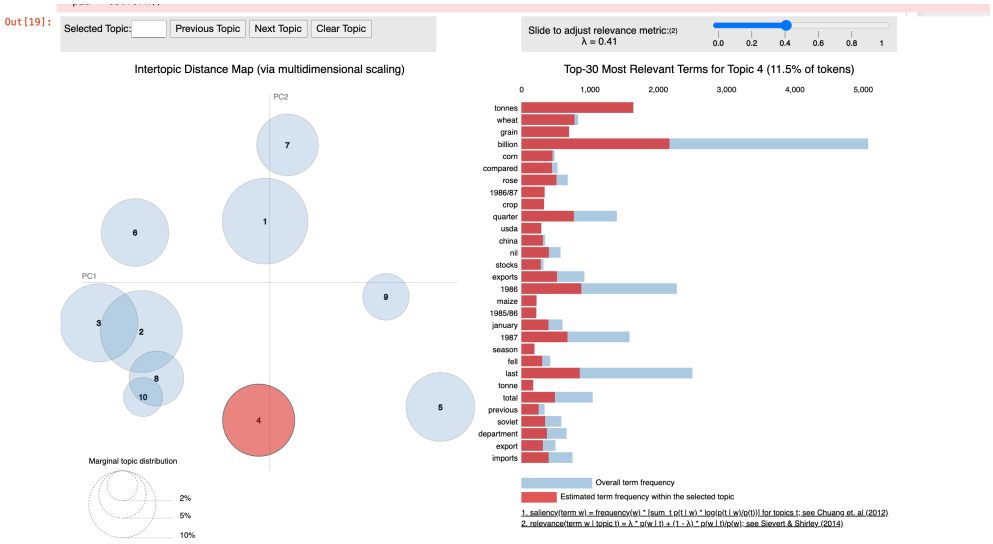


The Word Cloud shows that finance-related terms like CTS, Share, and Dlr(dollors) are the most prominent. The word "mln" appears frequently across the top 10 topics, likely as a shorthand for "million." The most common word overall is "said," which makes sense given that the dataset consists of news reports where quoting sources is common.

Task 2

Change the number of topics (e.g., 5) and compare the results. Which number of topics generate the most insightful results?

Top 10 Topics:



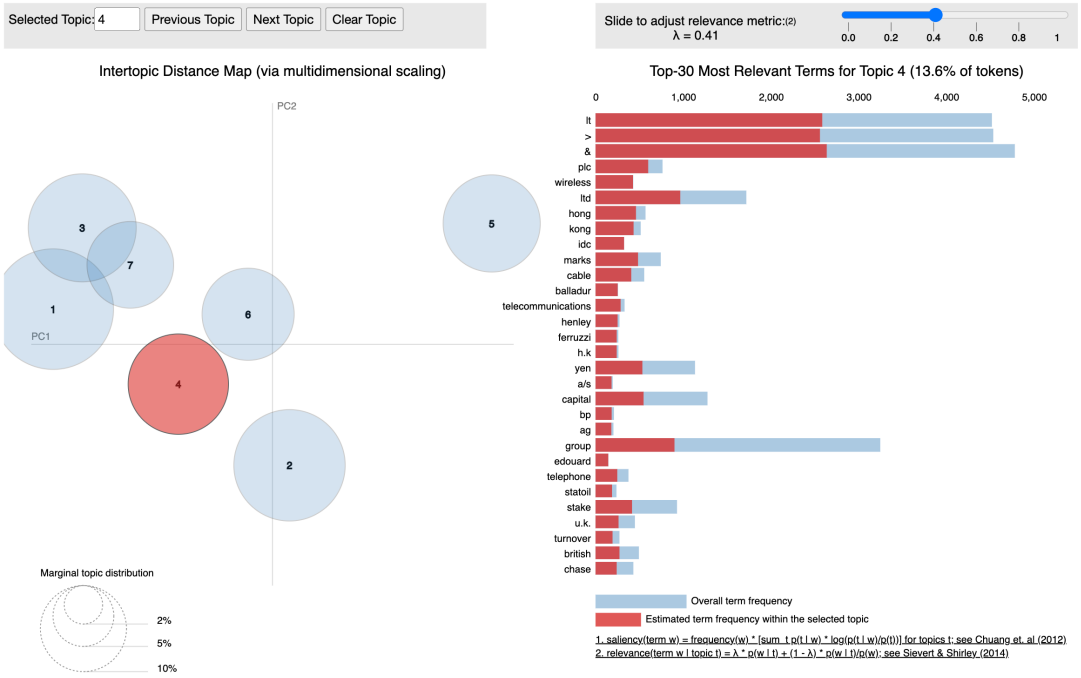
Coherence Score

```
Num Topics: 2, Coherence Score: 0.418945659981411
Num Topics: 3, Coherence Score: 0.4472296435282117
Num Topics: 4, Coherence Score: 0.413957618486714
Num Topics: 5, Coherence Score: 0.4475847275055832
Num Topics: 6, Coherence Score: 0.46709942783958275
Num Topics: 7, Coherence Score: 0.5192699232298065
Num Topics: 8, Coherence Score: 0.497068536053739
Num Topics: 9, Coherence Score: 0.46342465569469915
Num Topics: 10, Coherence Score: 0.45858620692958796

Best number of topics: 7
```

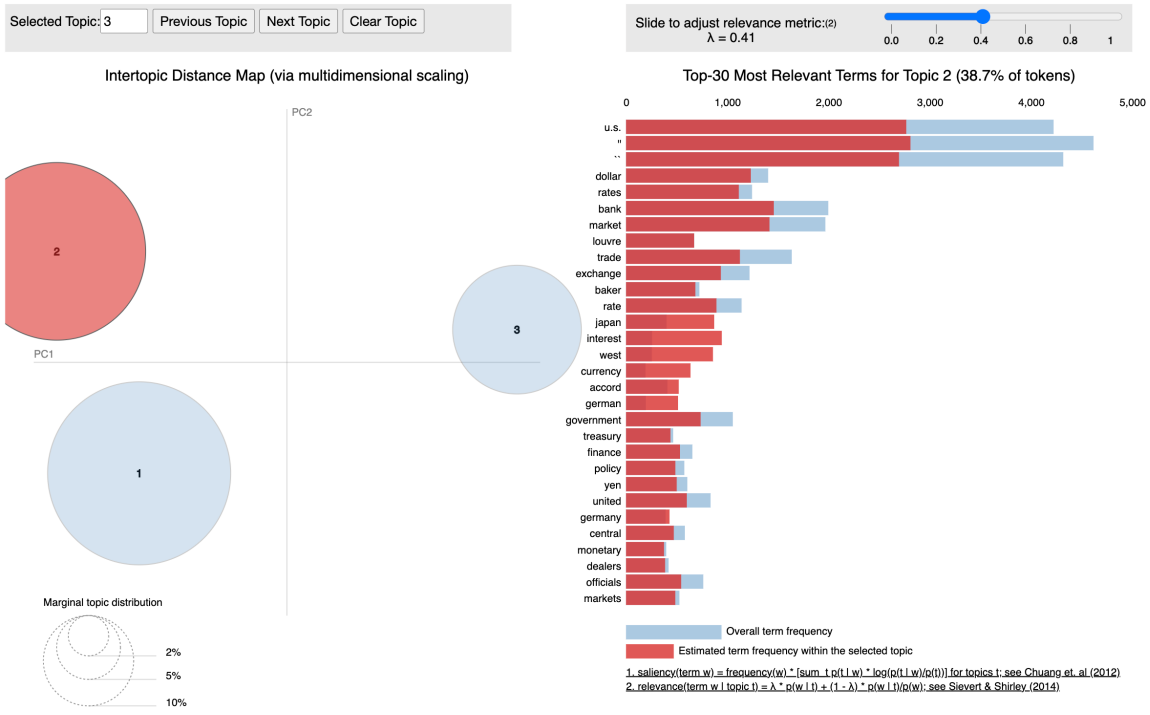
Jensen-Shannon Divergence

```
Num Topics: 2, Avg JSD Score: 0.079179558244582
Num Topics: 3, Avg JSD Score: 0.07234417263563742
Num Topics: 4, Avg JSD Score: 0.07559421442340651
Num Topics: 5, Avg JSD Score: 0.08200052453476792
Num Topics: 6, Avg JSD Score: 0.10835359186857968
Num Topics: 7, Avg JSD Score: 0.11066026773436077
Num Topics: 8, Avg JSD Score: 0.10524018162219094
Num Topics: 9, Avg JSD Score: 0.10780470617256122
Num Topics: 10, Avg JSD Score: 0.10906753711008087

Best number of topics based on JSD: 7
```

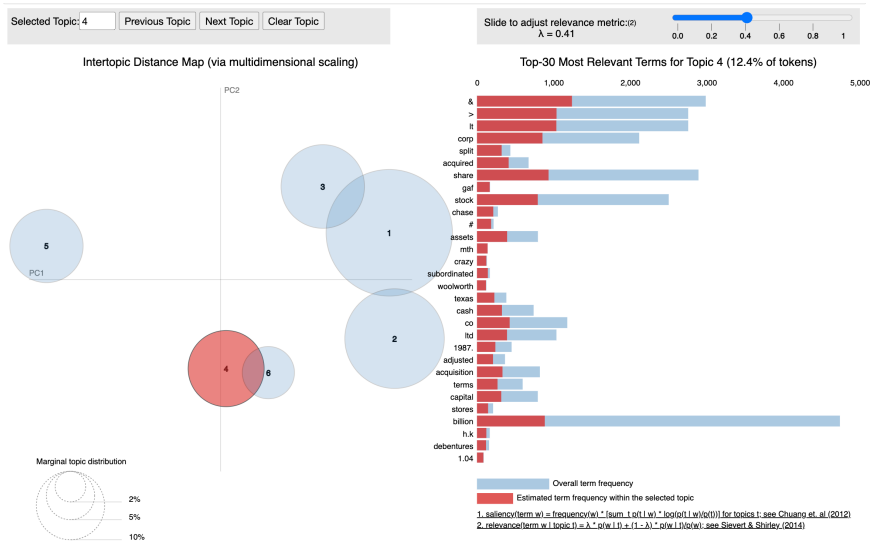
Top 7 Topics:



Top 3 Topics



6 Topics



Through my experiments, I found that setting the number of topics to three produced distinct, non-overlapping topics. However, it lacked granularity and generalized various words into same category. After evaluating the Coherence score and Jensen-Shannon Divergence, I determined that seven topics yielded the best results. Manually reviewing the top words for each topic showed that some shared common themes. Num of Topics equals to seven effectively distinguished different categories, keeping trade and geopolitical topics (1, 3, 7) closely related while defining financial topics more clearly than other configurations.

Task 3

Compare the type pre-assigned to the documents from the dataset and the dominant topic assigned by LDA for the documents retrieved in the previous step. Discuss your observations

```
Topic 0 (Pre-assigned: ['money-fx', 'interest'])
Topic 1 (Pre-assigned: ['trade'])
Topic 2 (Pre-assigned: ['acq'])
Topic 3 (Pre-assigned: ['earn'])
Topic 4 (Pre-assigned: ['grain', 'wheat', 'corn'])
Topic 5 (Pre-assigned: ['acq'])
Topic 6 (Pre-assigned: ['crude'])
Topic 7 (Pre-assigned: ['trade'])
Topic 8 (Pre-assigned: ['earn'])
Topic 9 (Pre-assigned: ['acq'])
```

The LDA did mostly a decent job assigning topics. It captured topics related to acquisitions (acq), earnings (earn), trade (trade), and crude oil (crude) reasonably well.

For *Topic 0 (money-fx, interest)*, LDA likely grouped it under a broader financial theme but may not have distinguished between "money markets" and "interest rates," as both.

Topic 4 contains all the crops, which I expected, as the general terms surrounding these wo uld have mostly been the same. I was a little surprised though, that it did not include trade with this as it was also part of some of the text I observed around these terms.

I also feel that the model could have done better job merging similar LDA topics: topics like 1 & 7 (trade-related) and 5 & 9 (acquisitions) could potentially be merged for better clarity.

Part 2

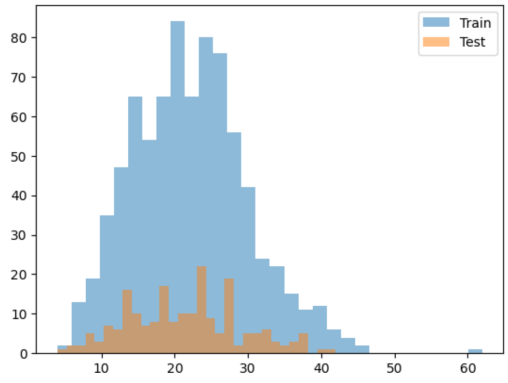
Task 1:

List down the columns and the number of instances in the training and test set.

```
Training set shape: (17806, 5)
Test set shape: (4271, 5)

Columns: ['Unnamed: 0', 'Sentence #', 'Word', 'POS', 'Tag']
```

Present the sentence maximum length, minimum length, and length distribution for training and test set.
Present the distribution of the named entity tags for training and test set in descending order of frequency of occurrence.

```
Training set – Max length: 62 , Min length: 4 , No. of training sentences: 800
Test set – Max length: 42 , Min length: 4 , No. of testing sentences: 200
```



List the most frequent 5 entities in the training and test sets.

```
Top 5 Train Tags:
 Tag
O        15246
B-gpe      472
B-geo      453
B-org      319
I-per      314
Name: count, dtype: int64

Top 5 Test Tags:
 Tag
O        3598
B-geo     130
B-org      94
I-per      86
B-tim      85
Name: count, dtype: int64
```

Task 3:

	precision	recall	f1-score	support
B-art	1.00	0.20	0.33	5
B-eve	0.00	0.00	0.00	7
B-geo	0.73	0.77	0.75	130
B-gpe	0.51	0.80	0.62	71
B-org	0.67	0.45	0.54	94
B-per	0.80	0.75	0.78	76
B-tim	0.95	0.66	0.78	85
I-art	0.00	0.00	0.00	2
I-eve	0.00	0.00	0.00	4
I-geo	0.74	0.52	0.61	33
I-gpe	0.00	0.00	0.00	0
I-org	0.52	0.69	0.59	45
I-per	0.78	0.90	0.83	86
I-tim	1.00	0.23	0.37	35
O	0.98	0.99	0.99	3598
accuracy			0.94	4271
macro avg	0.58	0.46	0.48	4271
weighted avg	0.94	0.94	0.94	4271

```
Span-Level Exact Match Evaluation:
Precision: 0.6847
Recall:    0.6496
F1-Score:  0.6667

Span-Level Approximate Match Evaluation:
Precision: 0.7271
Recall:    0.6944
F1-Score:  0.7104
```

For Token-level: -
High accuracy (0.94) but poor performance on rare tags (B-art: F1=0.33). Reveals specific weaknesses (e.g., I-tim recall=0.23)

Span-Level Exact Match: -
Significant drop from token-level (F1=0.67 vs 0.94). Highlights boundary detection issues.

Span-Level Approximate Match: -
6.6% improvement over exact match (F1=0.71 vs 0.67). More forgiving of minor boundary errors, because of which the increase from Span-level exact match.

The drop signify that the predicted values contained boundary detection issues.