Assignment 1

ITI5212
Data Analysis for Semi Structured Data

MONASH University

Student Name: Robiatul Adawiyah Al-Qosh
Student ID: 34269193

Table of Contents

Part 1: Text Classification

Introduction

The assignment utilizes a dataset sourced from arXiv.org, consisting of articles categorized under computer science, though some belong to mathematics or physics. The dataset spans from 2016 to 2024. The primary objective of Part 1 is to classify articles as Computational Linguistics (CL) or not using a binary classification approach. Two different text classification methods are employed: a statistical model (Logistic Regression) and a deep learning model (RNN). Additionally, this study evaluates the impact of dataset size (1000 samples vs. full dataset) and text length (title vs. abstract) on classification performance. As a result, a total of 8 different configurations are tested. These configurations are compared based on accuracy, precision, recall, and F1-score, with performance visualization using precision-recall curves.

Justification

1. *Library Selection*: The versions of Torch (2.3.0+cu121) and Torchtext (0.18) used in this study are not the most recent. This choice was made because Torchtext 0.18 contains specific functionalities required for building vocabulary, which were removed or modified in later versions.
2. *Handling Duplicate Data*: During data preprocessing, duplicate entries were found in the dataset. These duplicates were removed to prevent redundancy and potential bias in training, ensuring a more reliable model evaluation.
3. *Statistical Method:*
   - NLTK was used for data pre-processing in terms of tokenization and lemmatization to standardize word forms and reduce vocabulary size, improving model generalization.
   - Logistic Regression was selected due to its simplicity, efficiency, and strong performance on high-dimensional sparse data.
   - For the 1000 sample dataset, the model was trained on the first 700 rows from the training set and 300 from the development set. This decision was based on the fact that for the full dataset, both train and dev sets were merged for training, assuming that combining both sets would provide better learning, as no cross-validation was performed.
4. *RNN Method*:
   - In the data pre-processing stage, the tokenizer from Torchtext was used to ensure compatibility with PyTorch's data pipeline.
   - The model used a single layer nn.RNN, which was chosen to maintain simplicity and avoid overfitting on small datasets.
   - For the 1000 sample dataset, 300 samples were reserved for validation to help monitor training performance and prevent overfitting, given that deep learning models require careful tuning.
   - The number of epochs was set to 5 to balance between underfitting and overfitting.
5. *Testing Strategy*: For all configurations, both 1000 sample and full dataset models were tested on the entire test set, both for title or abstract based training for consistency in evaluation.

Analysis of Results

1. Comparison of Methods: Logistic Regression vs. RNN

   Logistic Regression consistently outperformed RNN in all configurations. This could be due to TF-IDF vectorization, which transforms text into meaningful numerical representations that work well with linear models. On the other hand, RNN struggled to generalize effectively, especially with smaller datasets, because it lacked the ability to capture long-term dependencies

with a single layer architecture. Even though RNN showed some improvement in the full dataset, it still could not outperform Logistic Regression. This suggests that more advanced architectures (e.g., LSTM or GRU) might be needed for RNN to fully utilize larger datasets. Moreover, Logistic Regression requires fewer computational resources and trains faster, making it a more practical choice for this classification task.

2. Impact of Dataset Size: 1000 vs. Full Dataset

   Both models performed significantly better with the full dataset. For example, Logistic Regression's accuracy increased from ~83% to ~97% when moving from 1000 samples to the full dataset. This aligns with expectations, as larger datasets provide more diverse examples for the model to learn from, reducing overfitting and improving generalization. The RNN model also saw improvements with the larger dataset in terms of accuracy metrics, but the gains were very small and not as pronounced as in Logistic Regression. This indicates that RNN needs even more data to effectively capture text patterns.

3. Comparison: Title vs. Abstract-Based Training

   Across all configurations, Logistic Regression models trained on abstracts significantly outperformed those trained on titles. Abstracts contain more informative and contextually rich data, making classification easier, while titles are often too short and ambiguous, leading to lower accuracy. However, in the RNN models, the difference was unclear. Surprisingly, for the 1000 sample dataset, the model trained on titles achieved higher precision, recall, and F1-score compared to the abstract based RNN, despite having lower accuracy. Meanwhile, for the full dataset, both title and abstract based RNNs yielded nearly identical results, suggesting that RNNs struggle to leverage additional information from abstracts as effectively as Logistic Regression models do. This could be due to the nature of RNNs, which rely on sequential dependencies and might not efficiently extract key features from longer texts like abstracts, especially with a simple single layer architecture. Additionally, the shorter length of titles may allow RNNs to perform relatively better due to their ability to capture short term dependencies more effectively.

	Configuration	Test Accuracy	Test Precision	Test Recall	Test F1
0	Logistic Regression 1000 Title Data	0.828720	0.923527	0.378746	0.537187
1	Logistic Regression All Title Data	0.952081	0.940000	0.873152	0.905344
2	Logistic Regression 1000 Abstract Data	0.834696	0.969114	0.382341	0.548346
3	Logistic Regression All Abstract Data	0.974887	0.963737	0.939672	0.951553
4	RNN 1000 Title Data	0.737548	0.368774	0.500000	0.424476
5	RNN All Title Data	0.737548	0.368774	0.500000	0.424476
6	RNN 1000 Abstract Data	0.737234	0.368733	0.499787	0.424372
7	RNN All Abstract Data	0.737548	0.368774	0.500000	0.424476

Figure 1. Metrics Comparison

4. Evaluation of Metrics: Accuracy, Precision, Recall, and F1-score
   - *Accuracy*: Logistic Regression with full abstracts provided the highest accuracy (~97%), proving that both data richness and dataset size matter.

- *Precision & Recall*: Logistic Regression models maintained higher precision, meaning they made fewer false positive errors. RNN models, while achieving decent recall, had lower precision, indicating more false positives.
- *F1-score*: Since F1-score balances precision and recall, Logistic Regression's full abstract model had the highest value (~0.95%), reinforcing its overall effectiveness.

5. Precision-Recall Curve Analysis

Precision-recall curves illustrate the stability of Logistic Regression compared to RNN. The Logistic Regression PR curve is smooth and well-distributed, while RNN's PR curve is more erratic, indicating variability in predictions. The best PR-AUC score (~0.9863) was achieved by Logistic Regression on full abstracts, highlighting its reliability in distinguishing between classes.

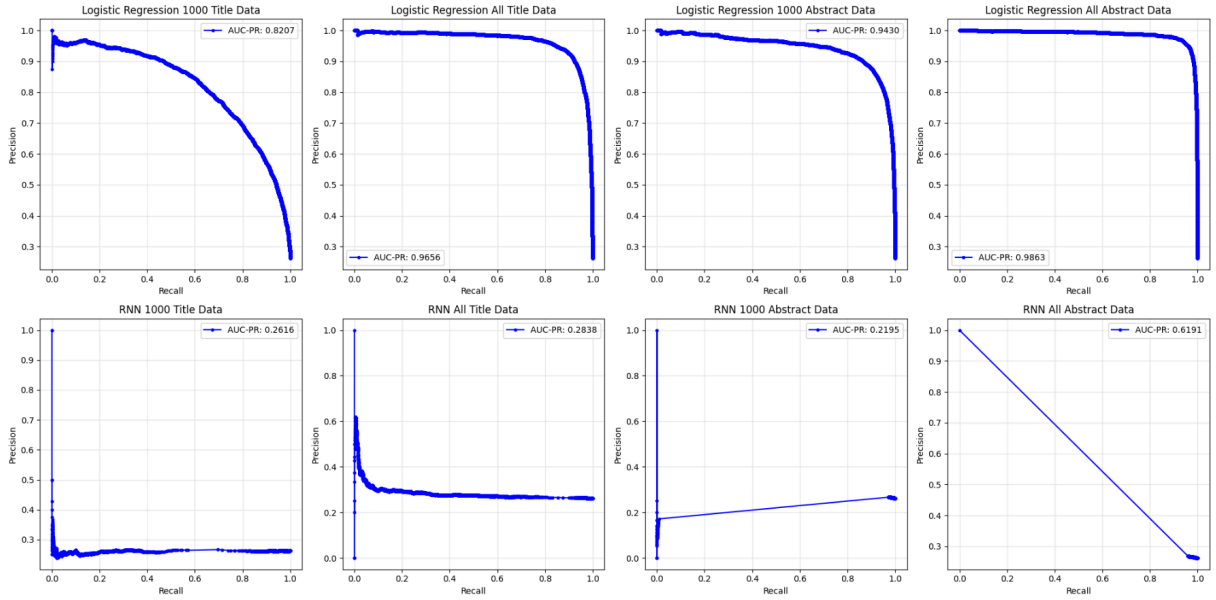

Figure 2. Precision-Recall Curve Charts

Conclusion

Logistic Regression demonstrated superior performance across all configurations, particularly when trained on full abstracts. Dataset size played a crucial role, as both models improved with more data, though RNN required significantly more training samples to be competitive. The choice of text source also influenced performance, with abstracts providing more meaningful context than titles. Overall, Logistic Regression remains the best choice for this task, whereas RNN would require more advanced architectures and larger datasets to compete effectively.

Part 2: Topic Modelling

Introduction

This part 2 explores unsupervised topic modeling using Latent Dirichlet Allocation (LDA) to uncover hidden themes within research articles from arXiv.org. Unlike Part 1, which focused on classification, this part aims to extract latent topics without predefined labels. The analysis investigates how topics differ based on dataset size (1000 vs. 20,000 articles) and text representation (Unigram vs. Bigram). This results in four different configurations, each assessing the impact of data

volume and phrase structure on topic coherence and distribution. Finally, the findings are analyzed using LDA visualizations.

Justification

1. *Data Used*: Given that Part 1 showed abstracts performed better than titles in text classification, this study exclusively utilizes abstracts for topic modeling. Abstracts contain richer and more descriptive content, making them more suitable for discovering latent topics.
2. *Data Preprocessing*: Spacy is used for tokenizing & lemmatizing to normalize text, reduce variations in word forms, and efficiently handle large-scale NLP tasks due to its optimized pipeline.
3. *Unigram vs. Bigram Representation*: Both unigram (single words) and bigram (two-word phrases) representations are used to compare their effects on topic quality. Unigrams capture individual word distributions, while bigrams help identify more meaningful phrase-level topics (e.g., "neural network" instead of separate words "neural" and "network").
4. *Filtering Extreme Words*: Words appearing in less than 20 documents or in more than 50% of the corpus are removed to balance rare and dominant terms. Filtering low-frequency words helps remove noise that could introduce irrelevant topics, while eliminating highly frequent words prevents overly generic topics, improving topic coherence and ensuring that extracted topics are more meaningful.

Analysis of Results

1. Effect of Dataset Size (1000 vs. 20,000 Articles)

   The larger dataset (20,000 articles) produced more distinct and coherent topics compared to the smaller dataset (1000 articles). In the 1000-article dataset, some topics overlapped significantly, likely due to insufficient data for LDA to properly differentiate between topics. In contrast, the 20,000-article dataset resulted in clearer topic separation, with more specialized clusters emerging. This suggests that LDA benefits from larger datasets, as it can better identify topic structures when given more text.

2. Effect of Text Representation (Unigram vs. Bigram)

   The bigram representation provided more interpretable topics compared to unigram. This is because bigrams capture phrase-level meaning, reducing ambiguity. For instance, while unigram topics contained words like "network", "deep", and "model", the bigram approach revealed clearer terms such as "neural network" and "deep learning". However, some bigram topics also contained redundant or less meaningful word pairs, which may require additional filtering.

3. Topic Coherence and Overlapping Topics
   - In the 1000-article unigram model, there are overlapping topics, particularly those discussing machine learning and NLP. A clear example is Topic 2 (Medical AI & Computer Vision) and Topic 5 (Autoencoders, Clustering & Representation Learning). Topic 2 contained articles related to biomedical benchmarks, cytology diagnosis, and medical image classification using CNNs, while Topic 5 focused on autoencoder architectures, subspace clustering, and pre-trained language models. The overlap occurred because terms like "graph" or "image" were relevant to both topics, causing some articles, such as those discussing pre-trained models in medical AI, to be classified into both topics. This highlights how unigram representations may fail to capture distinct phrase-level contexts, leading to topic ambiguities.

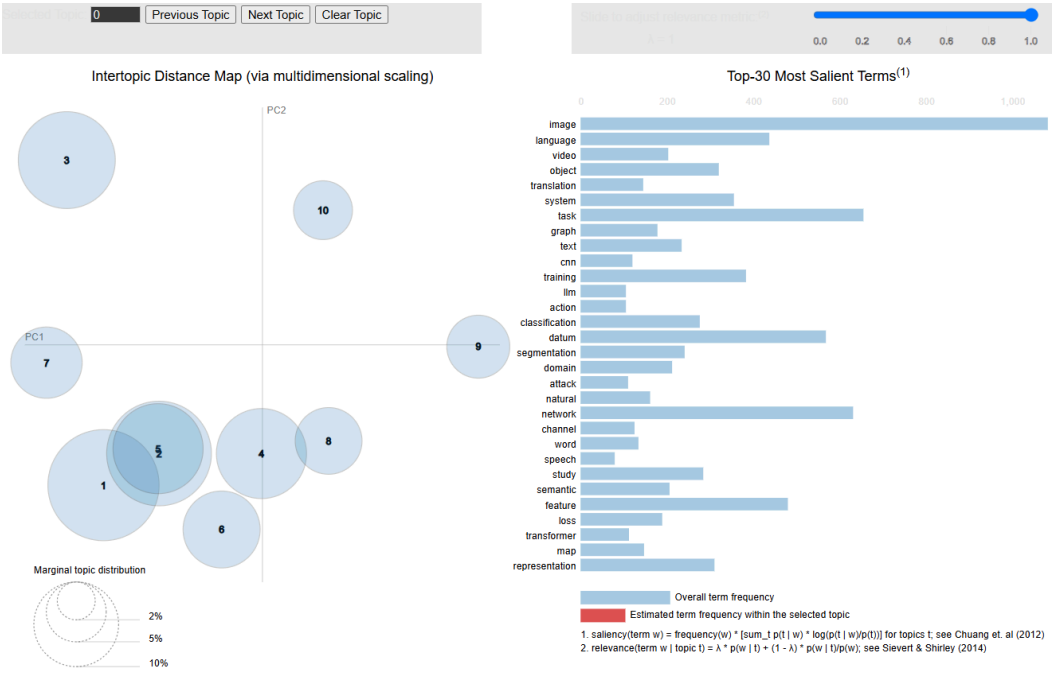

Figure 3. pyLDAvis LDA Visualization 1000-Article Unigram

```
# Topic 2
show_exemples(lda_1k_unigram, corpus_1k_unigram, dict_1k_unigram, article_1k['abstract'], 1)
```

```
Example 1:   Biomedical language understanding benchmarks are the driving forces for
artificial intelligence applications with large language model (LLM) back-ends.
However, most current benchmarks: (a) are limited to English which makes it
challenging to replicate many of the successes in English for other la...

Example 2:   Cytology is a low-cost and non-invasive diagnostic procedure employed to
support the diagnosis of a broad range of pathologies. Computer Vision
technologies, by automatically generating quantitative and objective
descriptions of examinations' contents, can help minimize the chances of
misdiagnoses...

Example 3:   Convolutional Neural Network (CNN) has been successfully applied on
classification of both natural images and medical images but not yet been
applied to differentiating patients with schizophrenia from healthy controls.
Given the subtle, mixed, and sparsely distributed brain atrophy patterns of
sc...
```

```
# Topic 5
show_exemples(lda_1k_unigram, corpus_1k_unigram, dict_1k_unigram, article_1k['abstract'], 4)
```

```
Example 1:   We present a new autoencoder-type architecture that is trainable in an
unsupervised mode, sustains both generation and inference, and has the quality
of conditional and unconditional samples boosted by adversarial learning.
Unlike previous hybrids of autoencoders and adversarial networks, the
adve...

Example 2:   Subspace clustering is the classical problem of clustering a collection of
data samples that approximately lie around several low-dimensional subspaces.
The current state-of-the-art approaches for this problem are based on the
self-expressive model which represents the samples as linear combinatio...

Example 3:   In light of the success of the pre-trained language models (PLMs), continual
pre-training of generic PLMs has been the paradigm of domain adaption. In this
paper, we propose QUERT, A Continual Pre-trained Language Model for QUERy
Understanding in Travel Domain Search. QUERT is jointly trained on f...
```

- In the 1000-article bigram model, the separation between topics might be improved, but new overlaps emerged. Specifically, Topic 5 moved away from Topic 2 but became more closely related to Topic 10 (Signal Processing & Adaptive Networks), which contained research on distributed recursive estimation, adaptive LMS algorithms, and signal processing. The shift occurred because bigram representation grouped related phrases, but numerical and optimization techniques linked Topic 5 (representation learning) with Topic 10 (adaptive modeling).

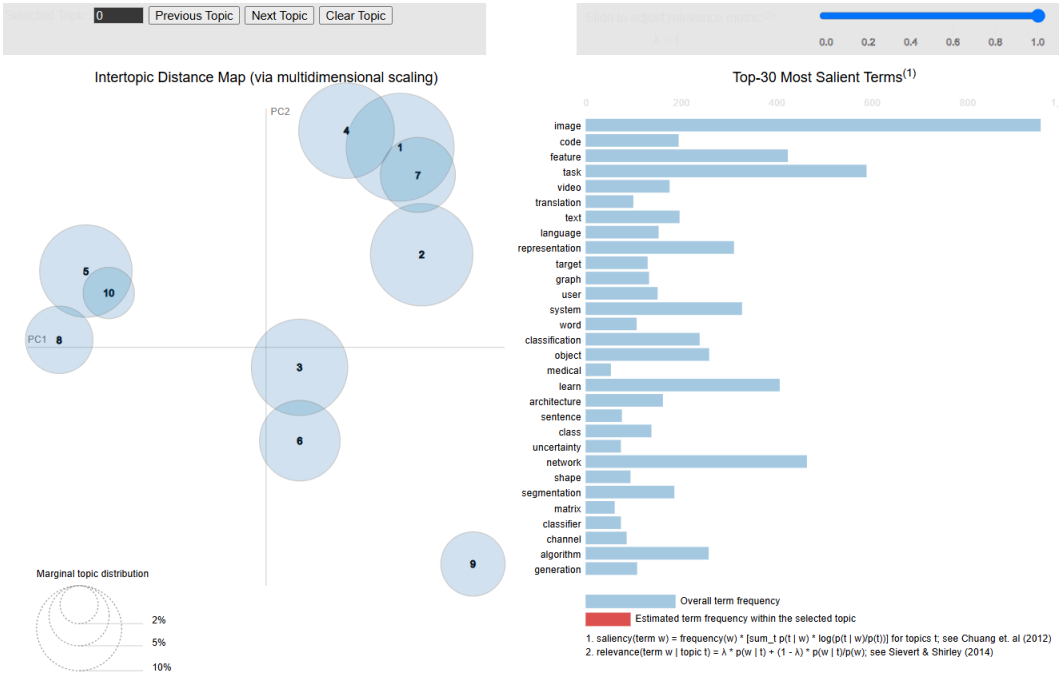

Figure 4. pyLDAvis LDA Visualization 1000-Article Bigram

```
# Topic 10
show_exemples(lda_1k_bigram, corpus_1k_bigram, dict_1k_bigram, article_1k['abstract'], 9)
```

```
Example 1:   We consider distributed recursive estimation of consensus+innovations type in
the presence of heavy-tailed sensing and communication noises. We allow that
the sensing and communication noises are mutually correlated while independent
identically distributed (i.i.d.) in time, and that they may both...

Example 2:   The performance of short polar codes under successive cancellation (SC) and
SC list (SCL) decoding is analyzed for the case where the decoder messages are
coarsely quantized. This setting is of particular interest for applications
requiring low-complexity energy-efficient transceivers (e.g.,
inter...

Example 3:   In this paper we focus on the tracking performance of incremental adaptive
LMS algorithm in an adaptive network. For this reason we consider the unknown
weight vector to be a time varying sequence. First we analyze the performance
of network in tracking a time varying weight vector and then we exp...
```

- In the 20,000-article unigram model, topic overlaps were still present, particularly between Topic 4 and Topic 6, which were closely related to machine learning applications and text analysis. Topic 4 (Computer Vision & Text Processing) contains research on road detection, Transformer models for text mining, and spatial-temporal action detection in videos. Meanwhile, Topic 6 (Model Explanations & NLP Applications) covers saliency maps for explainable AI, sentiment analysis, and generative approaches in mental health analysis. These topics overlapped significantly, as many machine learning techniques used in computer vision and text processing share similar methodological foundations. Additionally, Topic 4 and Topic 6 also had some interaction with Topic 2, likely due to shared methodologies in deep learning.

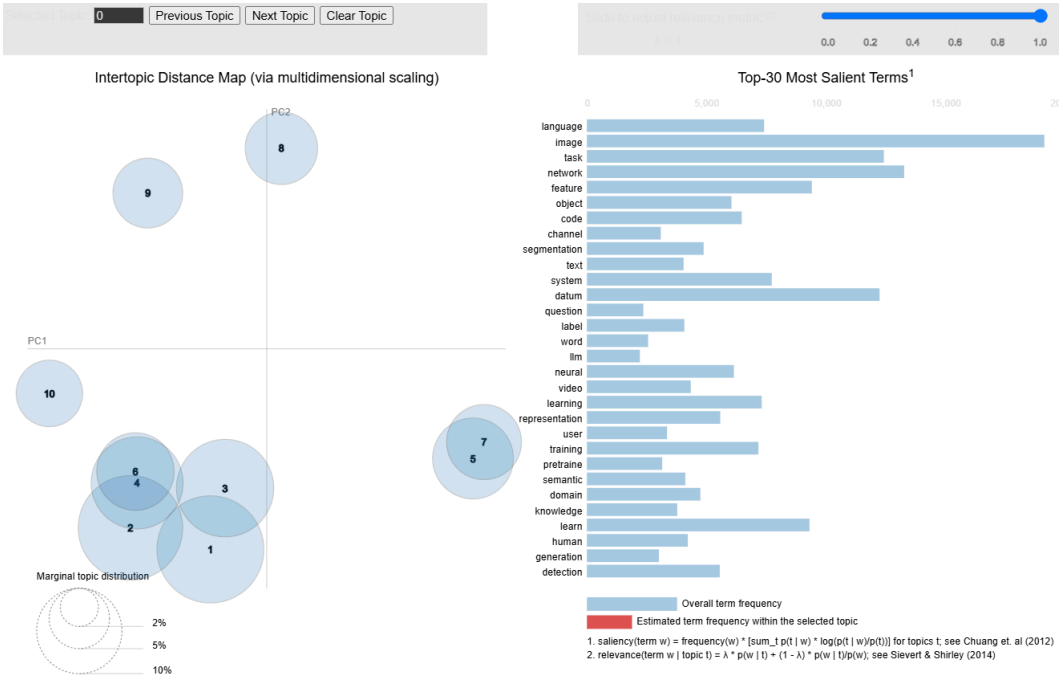

Figure 5. pyLDAvis LDA Visualization 20000-Article Unigram

```
# Topic 4
show_exemples(lda_20k_unigram, corpus_20k_unigram, dict_20k_unigram, article_20k['abstract'], 3)
```

```
Example 1:   Road detection or traversability analysis has been a key technique for a
mobile robot to traverse complex off-road scenes. The problem has been mainly
formulated in early works as a binary classification one, e.g. associating
pixels with road or non-road labels. Whereas understanding scenes with
f...

Example 2:   The recent advancement of pre-trained Transformer models has propelled the
development of effective text mining models across various biomedical tasks.
However, these models are primarily learned on the textual data and often lack
the domain knowledge of the entities to capture the context beyond ...

Example 3:   In this paper, we address the challenging problem of spatial and temporal
action detection in videos. We first develop an effective approach to localize
frame-level action regions through integrating static and kinematic information
by the early- and late-fusion detection scheme. With the intentio...
```

```
# Topic 6
show_exemples(lda_20k_unigram, corpus_20k_unigram, dict_20k_unigram, article_20k['abstract'], 5)
```

```
Example 1:   Model explanations such as saliency maps can improve user trust in AI by
highlighting important features for a prediction. However, these become
distorted and misleading when explaining predictions of images that are subject
to systematic error (bias) by perturbations and corruptions. Furthermore,...

Example 2:   This study is main goal is to provide a comparative comparison of libraries
using machine learning methods. Experts in natural language processing (NLP)
are becoming more and more interested in sentiment analysis (SA) of text
changes. The objective of employing NLP text analysis techniques is to
r...

Example 3:   Traditional discriminative approaches in mental health analysis are known for
their strong capacity but lack interpretability and demand large-scale
annotated data. On the other hand, generative approaches, such as those based
on large language models (LLMs),have the potential to get rid of heavy
...
```

- In the 20,000-article bigram model, Topic 2 (Quantum Computing and Secure Communication) emerged as a distinct topic and moved away from overlapping regions. This topic covered quantum error correction, consensus estimation in noisy environments, and security in biometric authentication. The separation of Topic 2 suggests that bigram representation helped refine topic boundaries by associating key terms more explicitly.

However, Topic 4 and Topic 6 still remained somewhat close, although they were less intertwined than in the unigram model.

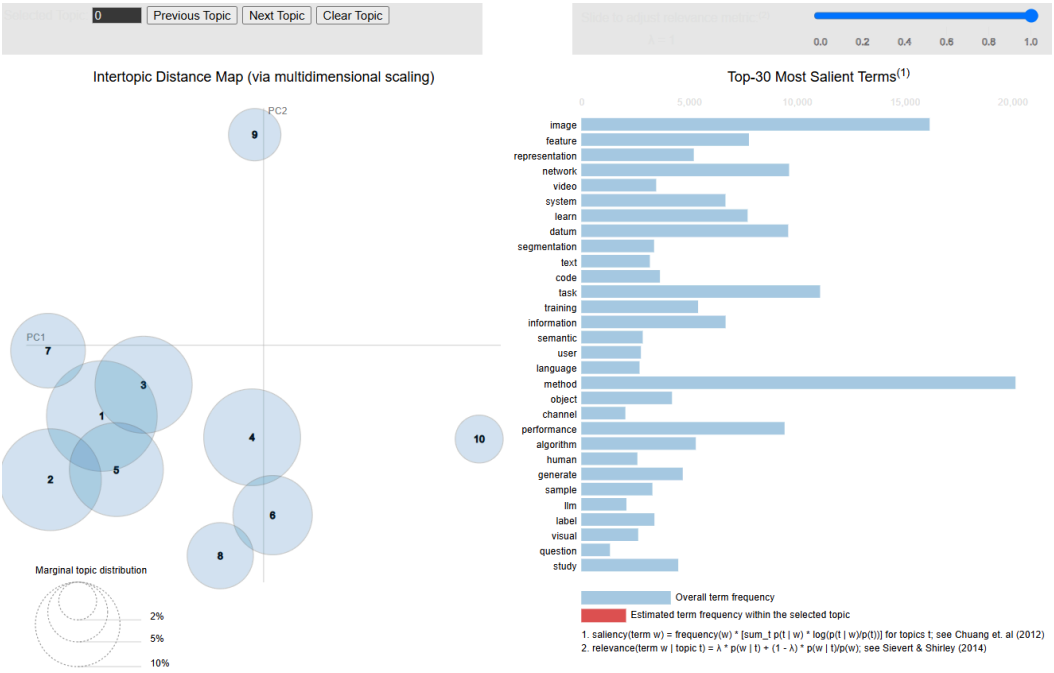

Figure 6. pyLDAvis LDA Visualization 20000-Article Bigram

```
# Topic 2
show_exemples(lda_20k_bigram, corpus_20k_bigram, dict_20k_bigram, article_20k['abstract'], 1)
```

```
Example 1:   We prove that the known formulae for computing the optimal number of
maximally entangled pairs required for entanglement-assisted quantum
error-correcting codes (EAQECCs) over the binary field hold for codes over
arbitrary finite fields as well. We also give a Gilbert-Varshamov bound for
EAQECCs a...

Example 2:   We consider distributed recursive estimation of consensus+innovations type in
the presence of heavy-tailed sensing and communication noises. We allow that
the sensing and communication noises are mutually correlated while independent
identically distributed (i.i.d.) in time, and that they may both...

Example 3:   We address security and privacy problems for digital devices and biometrics
from an information-theoretic optimality perspective, where a secret key is
generated for authentication, identification, message encryption/decryption, or
secure computations. A physical unclonable function (PUF) is a pro...
```

4. Topic Shifting Between 1000 and 20,000 Articles

An interesting observation is the shift in Topic 2 between the 1000-article and 20,000-article datasets. In the 1000-article unigram model, Topic 2 was primarily centered around biomedical applications and medical imaging, covering large language models for biomedical understanding, cytology, and neural networks applied to medical data. However, in the 20,000-article unigram model, Topic 2 transitioned into a broader discussion of quantum computing, security, and privacy concerns in digital systems. This shift suggests that topic evolution occurs as more data is introduced, allowing LDA to capture more distinct themes that may have been underrepresented in smaller datasets. The addition of bigrams further refined Topic 2's definition, making it more focused on quantum error correction and cryptography rather than a mixture of medical AI and security topics.

Conclusion

In summary, larger datasets and bigram representations improve topic coherence and clarity. The 20,000-article bigram model produced the best separation between topics, while smaller datasets and unigram models resulted in overlapping topics. The findings highlight how dataset size and text representation significantly impact LDA's effectiveness, reinforcing the importance of choosing the right configuration based on the desired level of topic granularity.

Acknowledgements

I acknowledge the use of AI assistance, specifically ChatGPT, for paraphrasing and refining grammar to enhance the clarity and coherence of this report.