News Topic Modeling Documentation

Team Number: 31

ID	Name
2022170617	محمد طارق محمد نبهان
2022170426	مصطفى محمد مصطفى ابراهيم الشرقاوي
2022170557	حسام عبدالرحمن محمد عبدالتواب
2022170480	هادي أحمد عثمان عطية الكفراوي
2022170506	يوسف احمد محمد على

Preprocessing

Raw text was cleaned and normalized through several steps to prepare it for topic modeling and clustering:

Named Entity Recognition (NER)

- Used spaCy's `en_core_web_sm` model to detect named entities (e.g., `South Korea`) and replaced spaces with underscores (`South_Korea`) to preserve multi-word entities as single semantic units.

Contraction Expansion

- Applied the `contractions` library to expand contractions (e.g., `don't` → `do not`) to ensure uniformity in tokenization.

Tokenization & Cleaning

- Converted text to lowercase and tokenized using `nltk.word_tokenize`.
- Removed stopwords and non-alphabetic characters to reduce noise in the corpus.

Language Detection

- Filtered out non-English articles using the `langdetect` library to maintain language consistency.

Lemmatization

- Reduced words to their base forms using spaCy's lemmatizer (e.g., `running` → `run`) to unify variants of the same word.

Null Handling

- Dropped the `url` column due to 100% null values.

- Retained only articles with more than 50 words after preprocessing to ensure sufficient context.

Feature Extraction

After preprocessing, the text data was transformed into a numerical format suitable for modeling:

Vectorization (TF-IDF / Count Vectorizer)

- Created a document-term matrix using either:
  - **TF-IDF Vectorizer**: Weighs each term by its importance across the corpus.
  - **Count Vectorizer**: Simply counts term occurrences per document.

Vocabulary Limitation

- Limited the vocabulary to the top `n` frequent words (e.g., `max_features=1000`) to reduce sparsity.
- Filtered out words that were either too frequent (`max_df`) or too rare (`min_df`) to avoid noisy or irrelevant tokens.

Modeling

Topic Modeling

Topic modeling uncovers hidden thematic structures in text. We applied the following three algorithms:

NMF (Non-negative Matrix Factorization)

NMF factorizes the document-term matrix into two lower-dimensional matrices with non-negative values. These matrices represent:

- The document-topic distribution.
- The topic-word distribution.

NMF is especially useful when interpretability is key, as it naturally leads to parts-based representations. In this notebook, NMF is used to extract coherent topics where each topic is defined by a few high-weighted terms.

LDA (Latent Dirichlet Allocation)

- A generative probabilistic model assuming:

    - Each document is a mixture of latent topics.
    - Each topic is a mixture of words drawn from a Dirichlet distribution.
- Suitable for interpreting hidden themes in large text corpora.
- Per-topic keywords and document-topic proportions aid interpretability.

LSA (Latent Semantic Analysis)

- Applies Singular Value Decomposition (SVD) to reduce dimensionality of the term-document matrix.
- Captures latent semantic structure through linear projections.
- Unlike LDA/NMF, LSA is not probabilistic and does not directly output topic distributions.

Clustering

Clustering groups similar documents based on their content or topic representations.

KMeans Clustering

- A centroid-based algorithm that divides the data into `K` clusters by minimizing intra-cluster variance.
- Operates on vectorized representations (e.g., TF-IDF or topic embeddings).
- Ideal for grouping articles into distinct categories.
- The elbow method or silhouette score may be used to select the optimal `K`.

Summary of Evaluation Metrics

Model	Coherence (c_v)	Diversity	Exclusivity	Perplexity	Silhouette Score
NMF	0.686	89.5%	0.952	N/A	N/A
LDA	0.590	94.0%	0.976	32,869.8	N/A
LSA	0.384	34.5%	0.589	N/A	N/A
KMeans	0.642	77.0%	0.932	N/A	0.008

Notes:

1. **Coherence (c_v):** Measures interpretability of topics using word co-occurrence statistics.
    - Best: **NMF** (0.686), followed by **KMeans** and **LDA**.

- Lowest: **LSA** (0.384).

2. **Diversity:** Percent of unique top-n words across all topics.
   - **LDA** leads with 94.0%, indicating highly varied topics.
3. **Exclusivity:** Measures how unique a word is to a single topic (closer to 1 = better).
   - **LDA** scores highest (0.976).
4. **Perplexity:** Evaluates how well a probabilistic model (like LDA) predicts unseen data (lower = better).
   - **LDA** shows high perplexity (32,869.8), suggesting overfitting or suboptimal tuning.
5. **Silhouette Score:** Used in clustering to measure how well-separated clusters are (closer to 1 = better).
   - **KMeans** scored poorly (0.008), indicating overlapping or weakly defined clusters.

Key Observations

- **NMF** provided the best balance of coherence and interpretability.
- **LDA** offered the most distinct and diverse topics but struggled with coherence and perplexity.
- **KMeans** gave moderate topic coherence but poor clustering separation.
- **LSA** underperformed across all evaluated metrics.