Topic Modeling

Extracted from this notebook HTML

Project Overview

- Goal: Discover latent topics in unlabeled documents
- Unsupervised NLP with probabilistic modeling
- Outputs: Topics (top words) + document-topic distribution

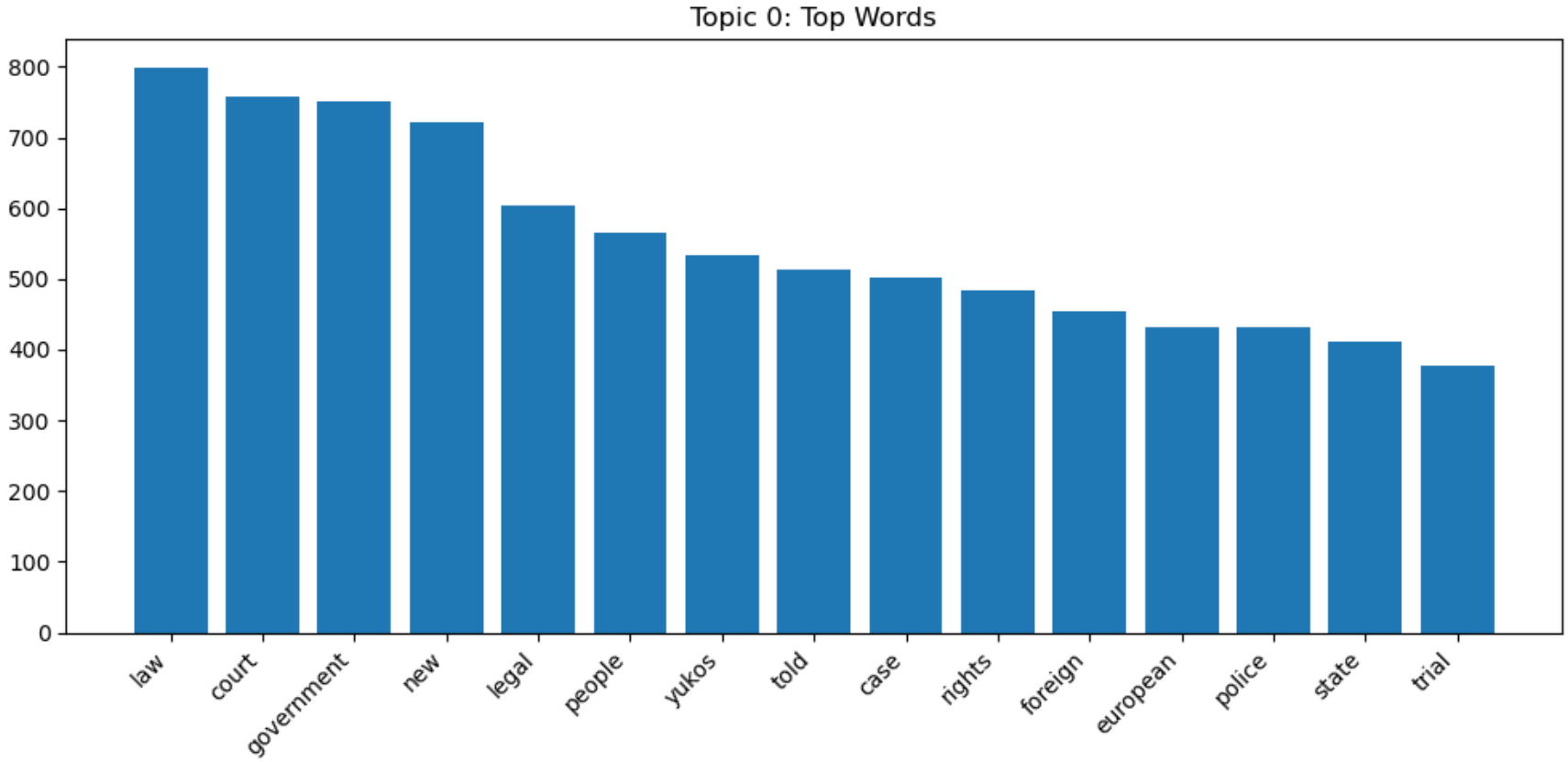
Steps

- Clean text (lowercase, remove punctuation/stopwords)
- Vectorize (Count, TF-IDF)
- Fit LDA to discover topics
- Evaluate with coherence & perplexity (if shown)
- Visualize top words + topic distribution

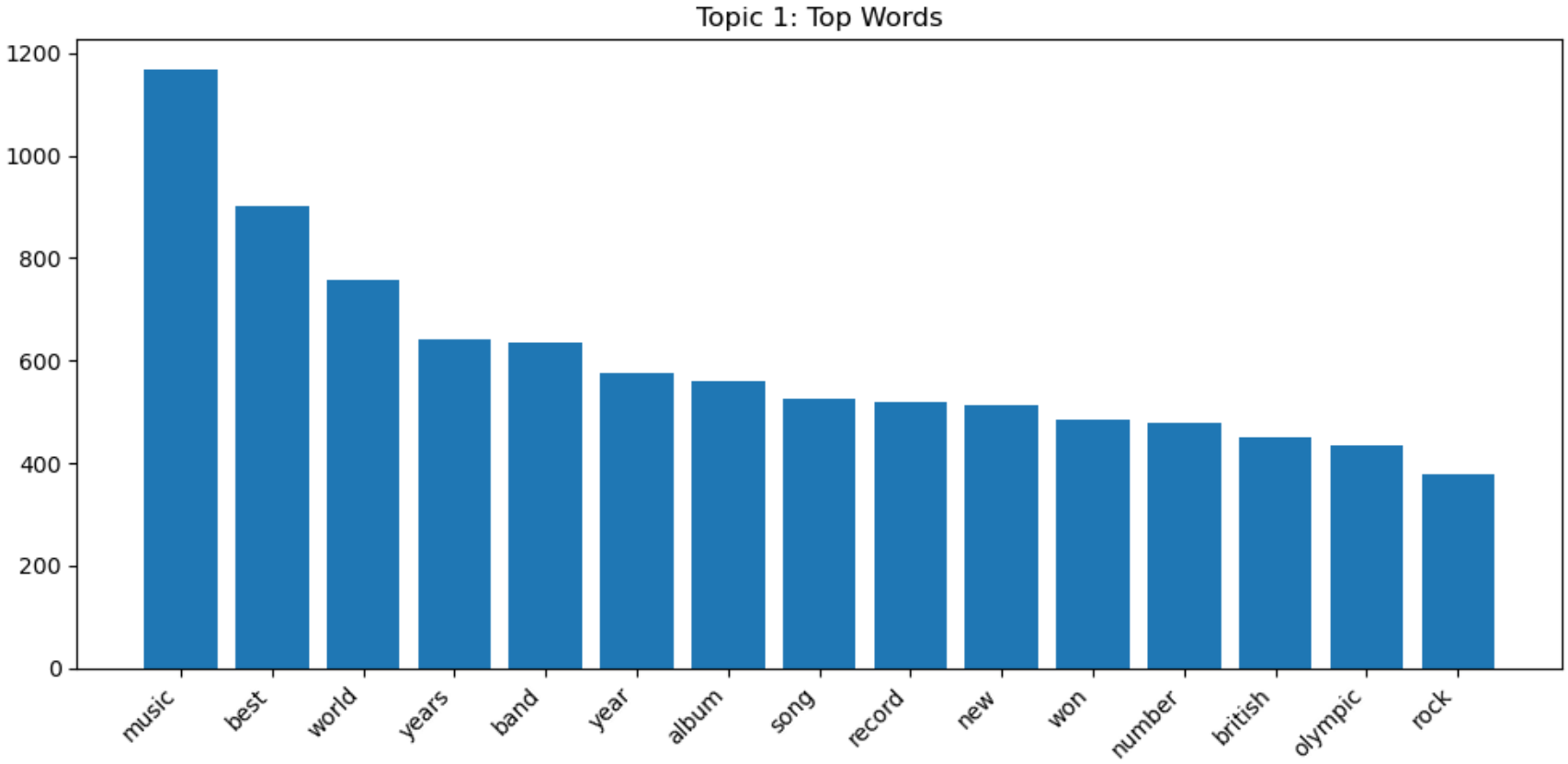
Method

- Model: Latent Dirichlet Allocation (LDA)
- Features: CountVectorizer and TF-IDF
- Preprocessing: lowercasing, digits/punctuation removal, stopwords

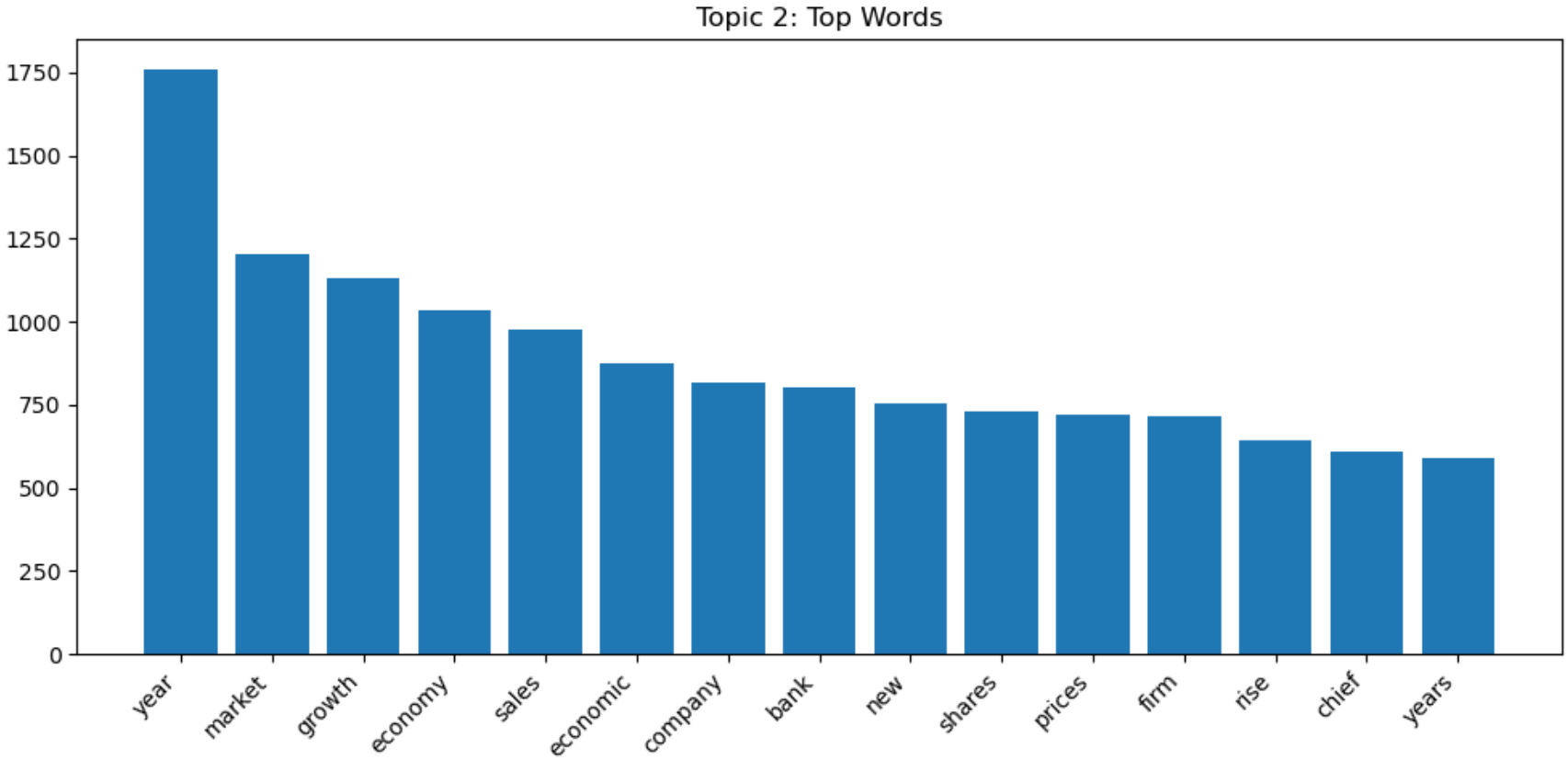
Visualization



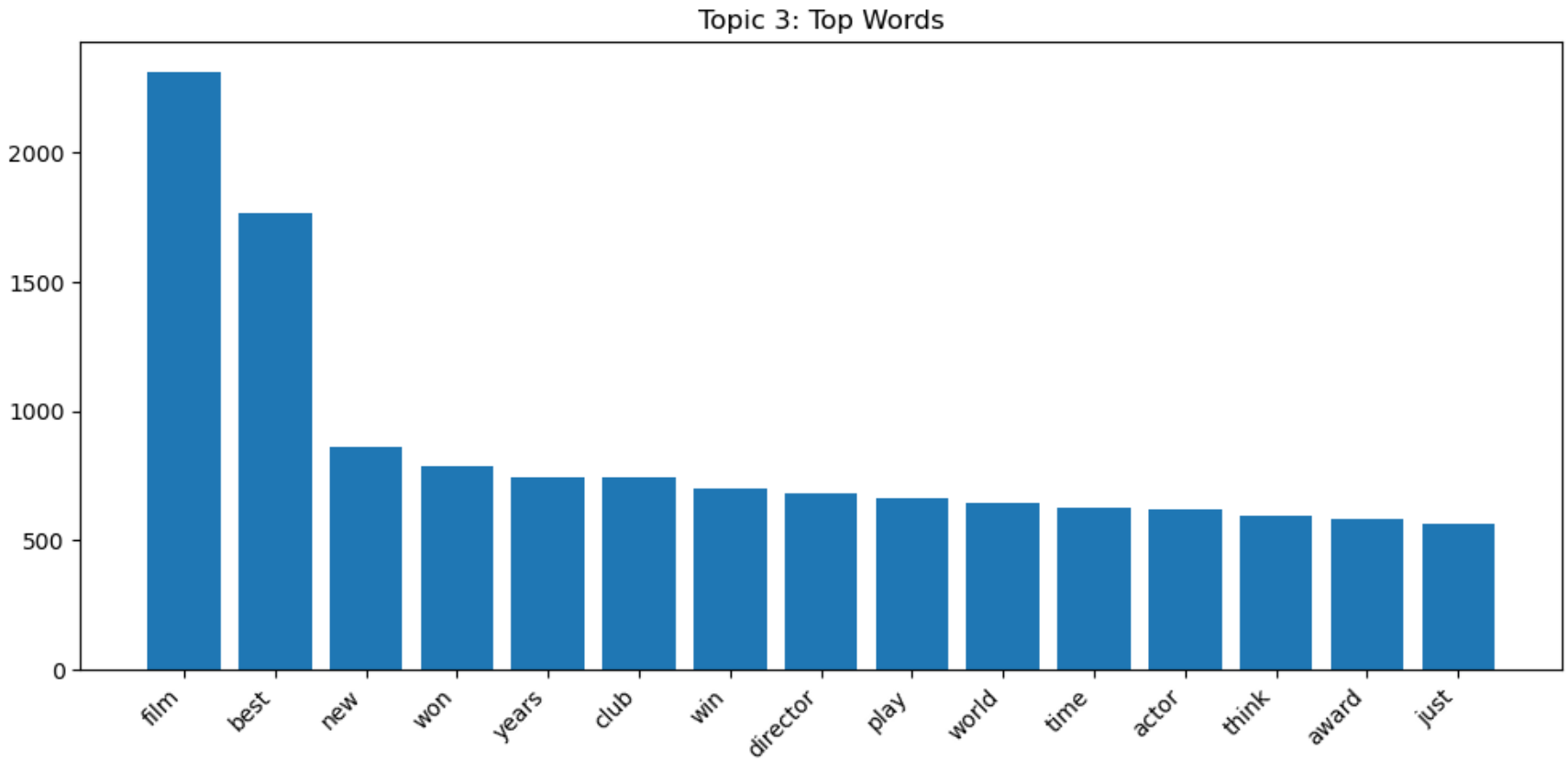
Visualization 2



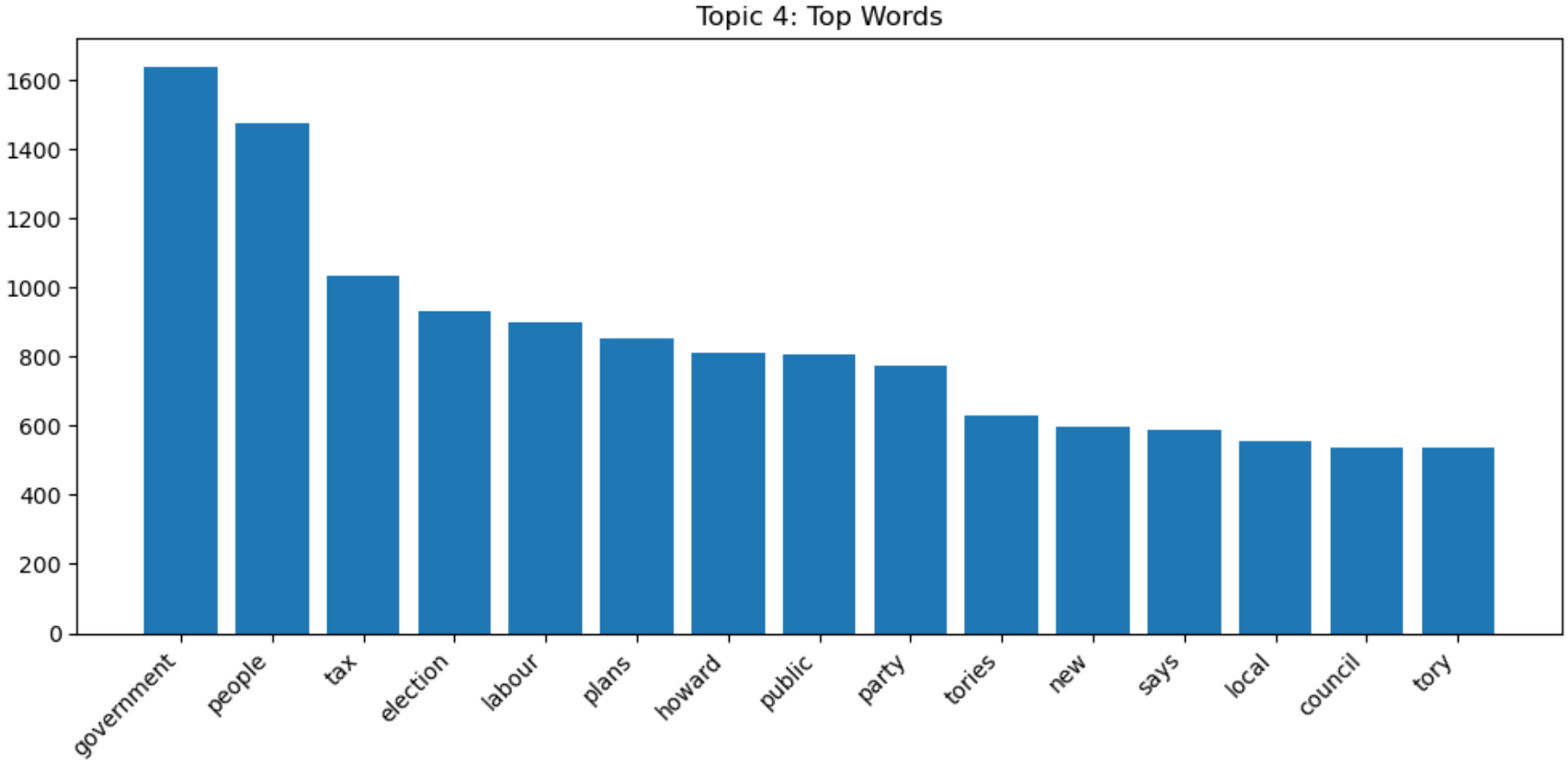
Visualization 3



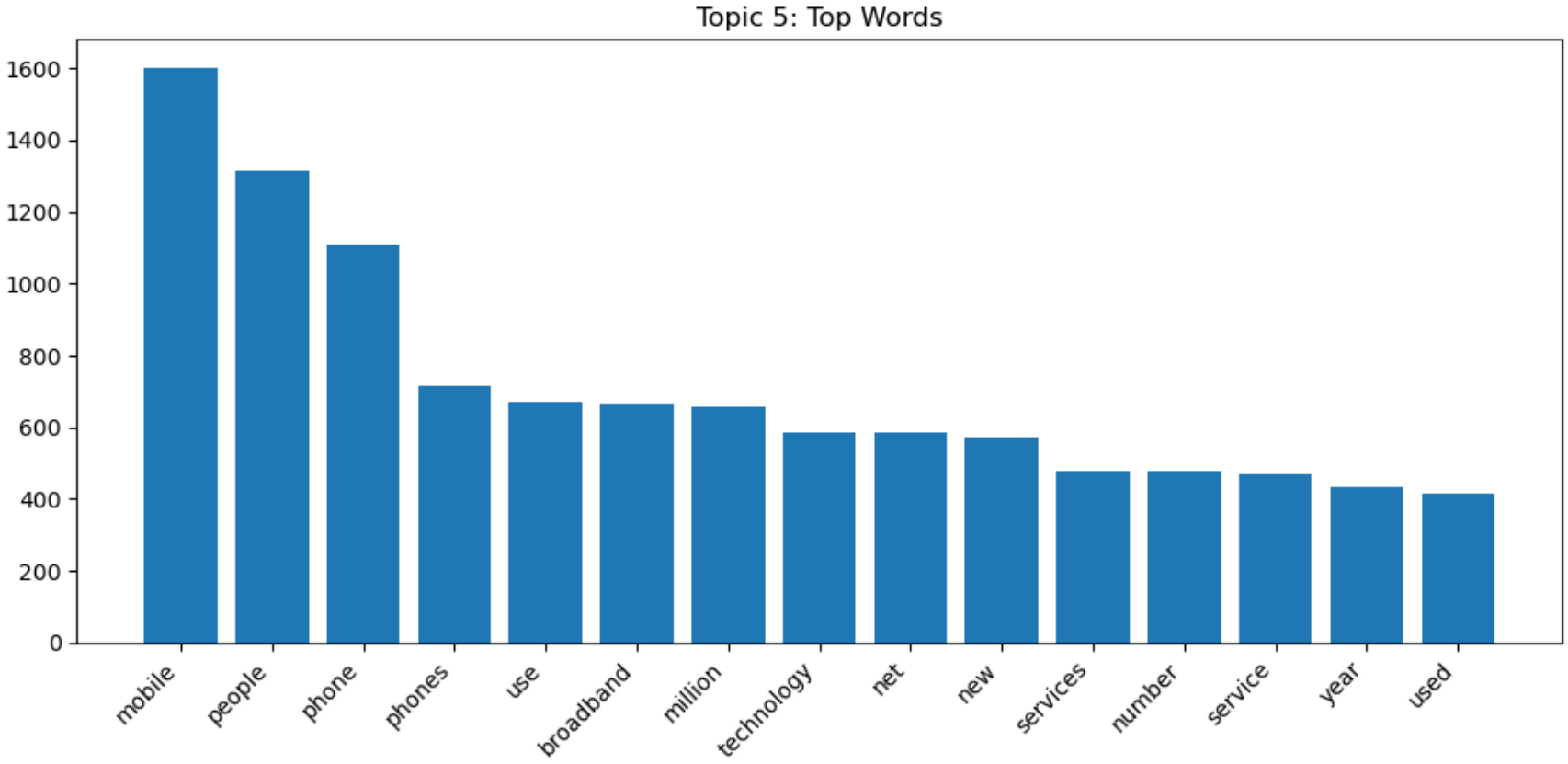
Visualization 4



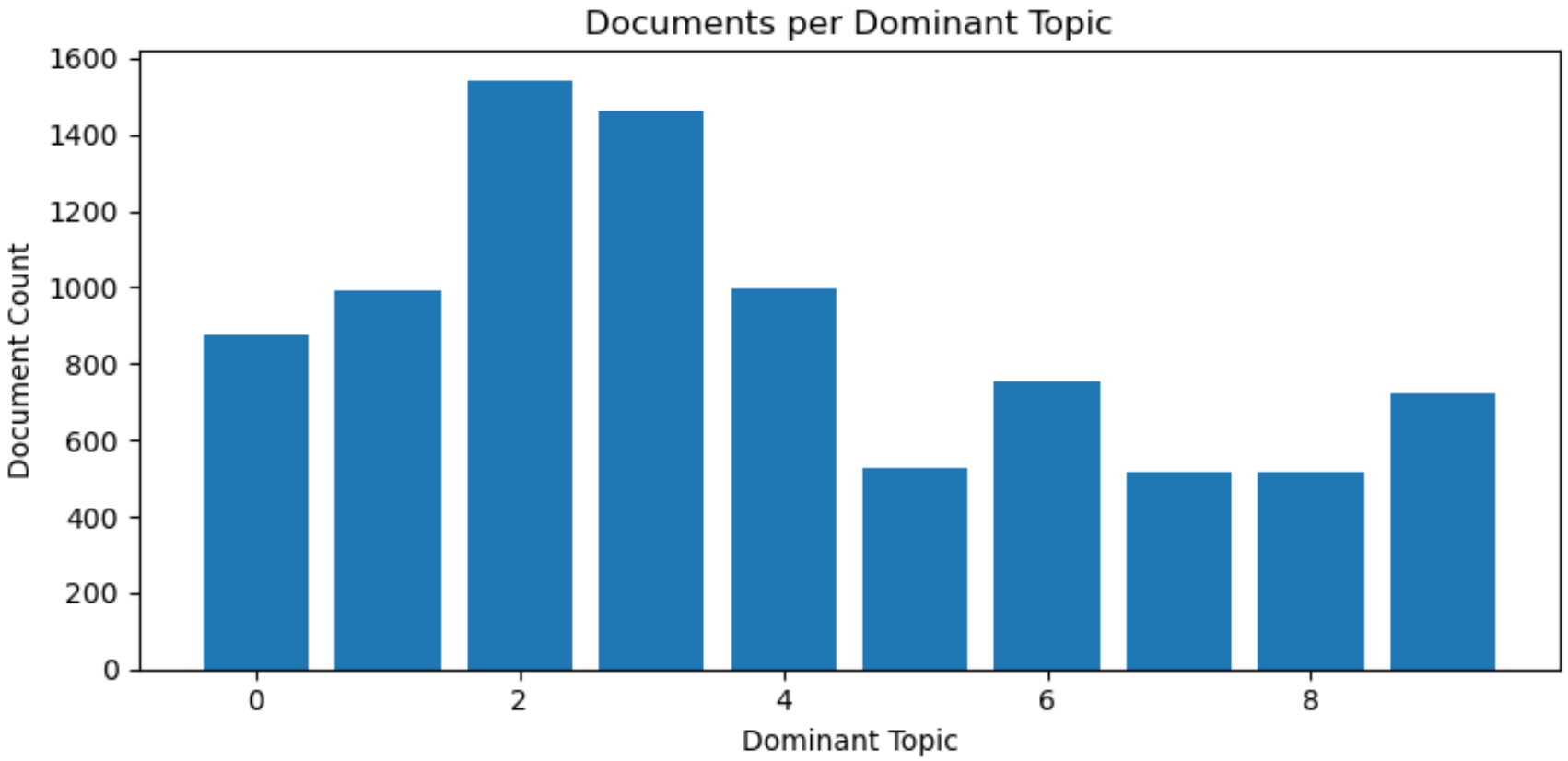
Visualization 5



Visualization 6



Visualization 7



Conclusion

- LDA reveals coherent topics via top words
- Figures reused directly from the notebook
- All numbers and visuals come from this HTML only