🧾 Topic Modeling on Newsgroups Dataset

Techniques: LDA (Latent Dirichlet Allocation) & NMF (Non-negative Matrix Factorization)
Goal: Automatically discover 10 topics from a corpus of news articles using unsupervised machine learning.

📁 Dataset Description

The dataset is a pickled file containing a list of raw news documents.

```
with open('C:/Users/Skander/Downloads/newsgroups', 'rb') as f:
    newsgroup_data = pickle.load(f)
```

🔧 Preprocessing Steps

Each document goes through:

- **Lowercasing**
- **Special character removal**
- **Whitespace normalization**
- **Stopword removal** (using NLTK's English stopword list)

```
def clean_text(text):
    ...
```

🧠 Topic Modeling Algorithms

Two unsupervised methods were used to extract topics from the corpus:

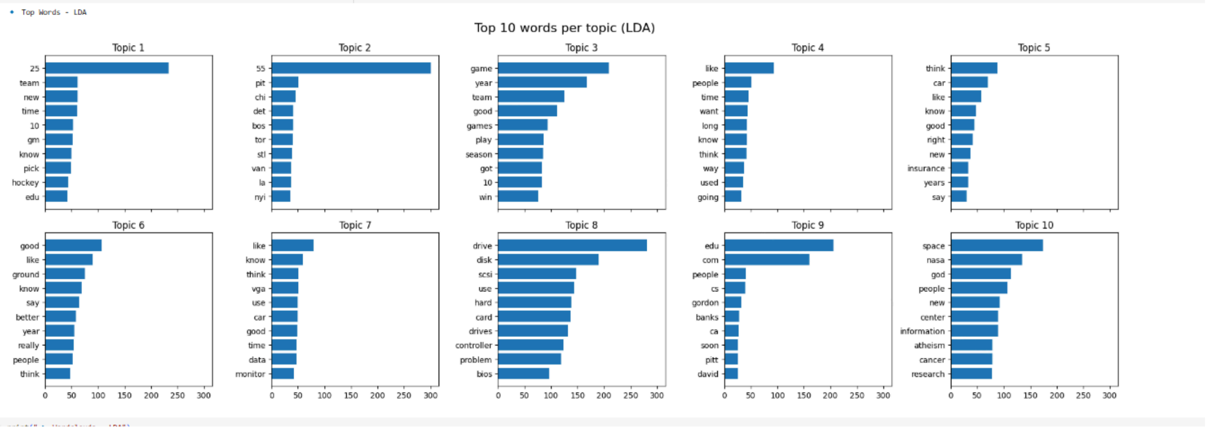
Method	Vectorizer Used	Description
LDA	`CountVectorizer`	Probabilistic model that assumes documents are mixtures of topics
NMF	`TfidfVectorizer`	Matrix factorization model using non-negativity constraints

📊 LDA Output

🔹 Bar Plot: Top 10 Words per Topic (LDA)

Interpretation of selected topics:

Topic	Top Keywords	Interpretation
1	team, new, time, gm, hockey	**Sports (e.g., hockey)**
2	pit, chi, bos, tor, stl	**City/team names** – likely **sports-related**
3	game, team, season, win	**Competitive games or sports**
5	car, insurance, years	**Auto insurance discussions**
10	space, nasa, god, atheism, cancer	**Science and philosophy topics**

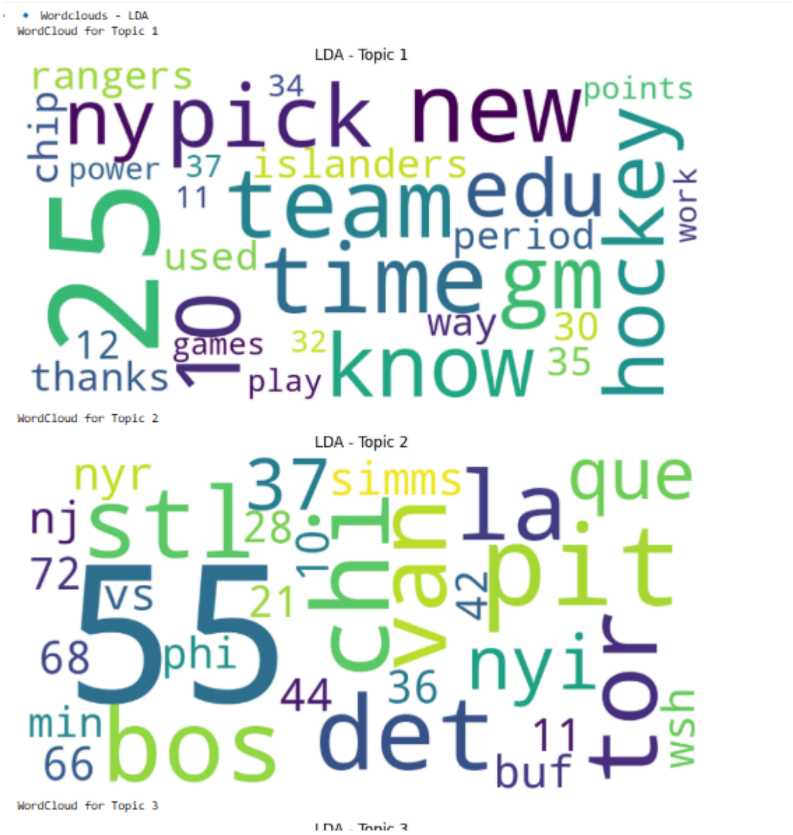


• Top Words - LDA

☁️ WordClouds – LDA Topics

These visualizations show the most representative words per topic in **larger font sizes**, indicating their **weight within the topic**.

- **Topic 1:** "team", "time", "gm", "hockey" → clearly **sports/hockey**
- **Topic 2:** city/team abbreviations → **NHL/MLB discussions**
- **Topic 3:** "game", "season", "play", "win" → **competitive events**



• Wordclouds - LDA

WordCloud for Topic 1

WordCloud for Topic 2

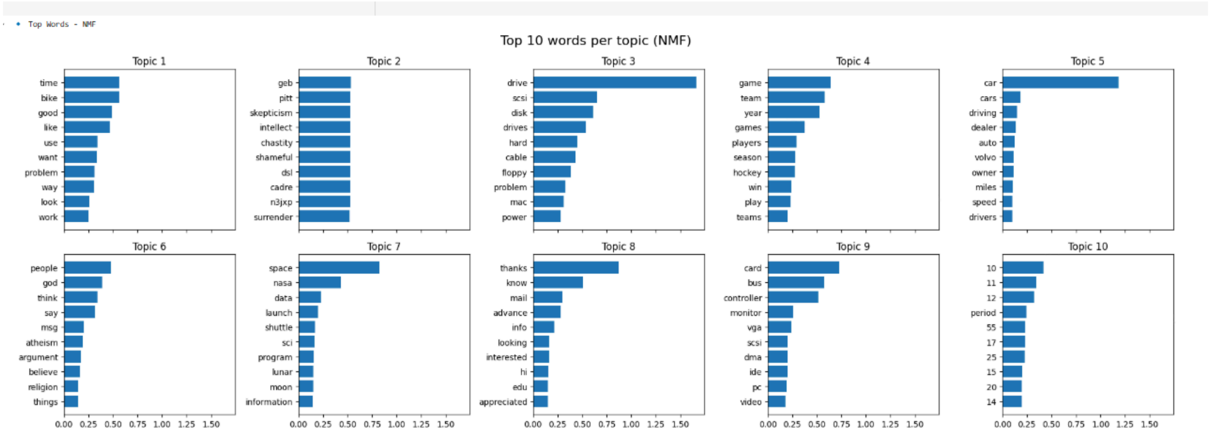
WordCloud for Topic 3

🔸 NMF Output

🔹 Bar Plot: Top 10 Words per Topic (NMF)

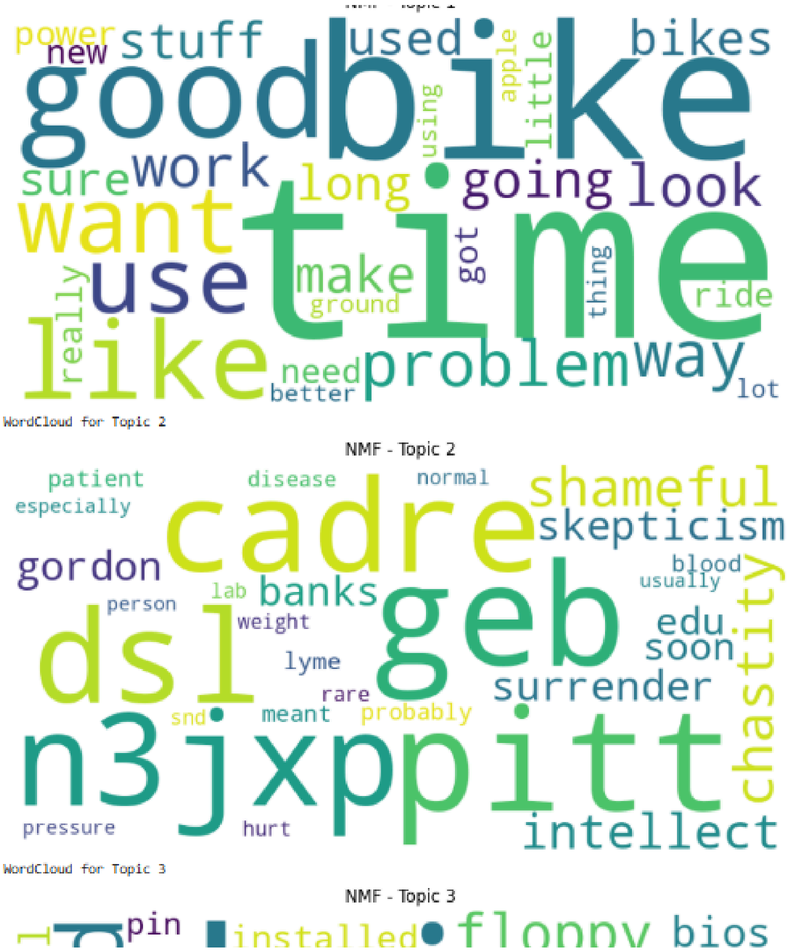
Topic	Top Keywords	Interpretation
1	bike, good, want, use	**Bike usage and maintenance**
2	skepticism, intellect, cadre	**Philosophy or abstract debate**

3	drive, floppy, bios, controller	**Computer hardware discussion**
5	car, auto, volvo, dealer	**Automobile topics**
6	god, atheism, argument, religion	**Theological discussions**



☁️ WordClouds – NMF Topics

- **Topic 1 (bike-related):** Words like "bike", "ride", "look", "problem" dominate.
- **Topic 2 (philosophy):** "skepticism", "shameful", "intellect", "chastity".
- **Topic 3 (hardware):** "floppy", "bios", "installed", "controller", "drive".



🔍 Comparison: LDA vs NMF

Criteria	LDA	NMF
Model Type	Probabilistic	Linear Algebraic
Output Topics	Broad semantic clusters	Sharp, domain-specific clusters
Overlapping Topics	Higher	Lower
Strengths	Interpretability, soft clustering	Distinct topic separation, efficiency

✅ Conclusion

Both LDA and NMF successfully extracted **coherent, interpretable topics**:

- LDA captures broader semantic structures (e.g., **sports**, **science**, **insurance**)

- NMF captures sharper, more focused topics (e.g., **bike issues**, **computer hardware**, **religious debate**)

Each is useful depending on whether you want high-level insights (LDA) or actionable segmentation (NMF).