Handout 1: Introduction to Topic Modeling (Basic Level)

Machine Learning for Smarter Innovation

1 Handout 1: Introduction to Topic Modeling (Basic Level)

1.1 What is Topic Modeling?

Topic modeling is like having a smart assistant that reads thousands of documents and tells you: "These documents are mainly about these themes." It automatically discovers hidden patterns in text collections.

1.1.1 Real-World Analogy

Imagine sorting a huge pile of letters into groups without reading each one completely. Topic modeling does this digitally - it finds common themes across documents automatically.

1.2 Key Concepts

1.2.1 1. Topics

- A topic is a collection of words that frequently appear together
- Example: {battery, charging, power, drain} = "Power Issues" topic
- Documents can contain multiple topics

1.2.2 2. Documents

- Any text: reviews, emails, reports, social media posts
- Each document is a mixture of topics
- Example: A product review might be 60% about quality, 30% about price, 10% about delivery

1.2.3 3. Words

- Building blocks of topics
- Same word can belong to different topics
- Context determines meaning

1.3 Why Use Topic Modeling?

1.3.1 Business Benefits

- **Scale**: Analyze thousands of documents in minutes
- **Objectivity**: Remove human bias from categorization
- **Discovery**: Find patterns humans might miss
- **Speed**: 70% faster than manual analysis

1.3.2 Common Applications

1. **Customer Feedback Analysis**
   - What are customers talking about?
   - Which issues are most common?
2. **Innovation Mining**
   - What ideas keep appearing?
   - Where are the opportunity gaps?
3. **Content Organization**
   - Auto-categorize articles
   - Create navigation structures

1.4 Simple Example with TextBlob

```
from textblob import TextBlob
import pandas as pd

# Sample reviews
reviews = [
    "The battery life is amazing and charges quickly",
    "Great battery, lasts all day long",
    "Beautiful design, looks premium",
    "Sleek design and comfortable to hold"
]

# Simple word frequency approach
word_counts = {}
for review in reviews:
    blob = TextBlob(review.lower())
    for word in blob.words:
        if word not in ['the', 'is', 'and', 'to']:
            word_counts[word] = word_counts.get(word, 0) + 1

# Find top words (simple topics)
top_words = sorted(word_counts.items(), key=lambda x: x[1], reverse=True)[:5]
print("Top theme words:", [word for word, count in top_words])
```

1.5 Practice Exercise

1.5.1 Task: Manual Topic Discovery

Given these 5 product reviews, identify 2-3 main topics:

1. “Easy to set up, installation was quick, great instructions”
2. “Simple installation process, anyone can do it”
3. “Expensive but worth it, high quality materials”
4. “Price is high but the quality justifies it”
5. “Customer service was helpful and responsive”

Your answers: - Topic 1: _____ (Words: _____) - Topic 2: _____ (Words: _____) - Topic 3: _____ (Words: _____)

1.5.2 Solution Guide

- Topic 1: **Installation** (setup, installation, easy, simple, quick)
- Topic 2: **Value** (expensive, price, worth, quality, high)
- Topic 3: **Service** (customer, service, helpful, responsive)

1.6 Key Takeaways

1. **Topics are word patterns** - Groups of words that appear together
2. **Documents mix topics** - Real text contains multiple themes
3. **Scale matters** - Manual works for 10 documents, not 10,000
4. **Objectivity helps** - Removes personal interpretation bias
5. **Speed is valuable** - Get insights in minutes, not days

1.7 Tools to Explore

1.7.1 Beginner-Friendly

- **Orange3**: Visual programming for text mining
- **MonkeyLearn**: No-code topic modeling
- **Google Cloud Natural Language**: Pre-built API

1.7.2 Next Level

- **Gensim**: Python library with tutorials
- **scikit-learn**: Integrated with data science tools

1.8 Questions for Reflection

1. What text data does your organization have that could benefit from topic modeling?
2. What patterns might you discover in customer feedback?
3. How could topic modeling speed up your current processes?

1.9 Further Reading

- "Topic Modeling for Humans" - Gensim Tutorial
- "A Friendly Introduction to Topic Modeling" - Medium article
- "LDA Explained Simply" - Towards Data Science

Remember: Start simple, focus on understanding the concepts before diving into complex algorithms.