General Report of Results Practical Assignment 1

Students: Osadici Darius, Rădulescu Adelina

1. Comparison TF-IDF vs LDA for “Food & Drink” description articles

We implemented both TF-IDF and LDA models to analyze similarity between Food & Drink articles in the news dataset. The table below summarizes our experimental results:

Method	Ratio Quality	Model Creation Time	Comparison Time	Total Goods
TF-IDF	59.92%	29.13s	5.02s	731/1220
LDA	38.77%	39.68s	0.27s	473/1220

1.1 Ratio Quality Comparison

Our results show that TF-IDF performed significantly better, achieving 59.92% accuracy compared to LDA's 38.77%. This means TF-IDF correctly identified Food & Drink articles in the top-10 similar documents about 21% more often than LDA did.

Why TF-IDF outperformed LDA:

The main reason is vocabulary specificity. Food & Drink articles contain highly specific terminology like ingredient names, cooking methods, and restaurant-related words that appear consistently within this category. TF-IDF excels here because it directly matches these terms - so articles mentioning "burgers," "chicken," or "pizza" get matched together naturally.
LDA struggled because with only 30 topics distributed across the entire dataset, Food & Drink content got spread across several topics instead of being concentrated in dedicated ones. This dilutes the model's ability to recognize this specific category. Additionally, while TF-IDF uses 161,607 features capturing fine-grained word distinctions, LDA compresses everything down to just 30 dimensions, losing a lot of the specific vocabulary signals that distinguish food-related content.

2.2. Execution Time Analysis

LDA took longer to build (39.68s vs 29.13s, about 36% slower) because it requires iterative sampling to discover latent topics, whereas TF-IDF only needs to compute term frequencies and inverse document frequencies. However, once built, LDA is dramatically faster for comparisons - 0.27 seconds versus 5.02 seconds for TF-IDF. That's about 18.6 times faster. This happens because LDA operates in a compressed 30-dimensional space while TF-IDF works with all 161,607 dimensions.

3.3 Trade-offs Between Methods

Each approach has strengths depending on what matters most for your application.

TF-IDF works better when categories have clear, specific vocabulary - like Food & Drink with its ingredient names and cooking terms. It's also straightforward to understand because you can see exactly which words drive the similarity scores, and there's no need to tune parameters. The downside is that comparisons take longer since you're working with massive vectors. Also, TF-IDF can't tell that "bicycle" and "helmet" are related - it just sees different words. Same problem with synonyms like "car" and "automobile."

LDA is much faster once built and can discover semantic relationships through word co-occurrence patterns. The dimensionality reduction helps computational efficiency too. But our results show it's less accurate for identifying specific categories. You also need to experiment with settings like topic count and passes to get decent results. Plus, the topics LDA creates don't necessarily align with our predefined categories, which makes interpretation trickier.

2. “Sports” worse quality results compared to “Food & Drink”

We repeated the analysis for Sports articles and compared them to Food & Drink. The results revealed significant performance differences:

Category	Method	Articles	Ratio Quality	Model Creation Time	Comparison Time	Total Goods
Food & Drink	TF-IDF	122	59.92%	29.13s	5.02s	731/1220
Food & Drink	LDA	122	38.77%	39.68s	0.27s	473/1220
Sports	TF-IDF	110	38.73%	28.75s	4.42s	426/1100
Sports	LDA	110	5.91%	14.15s	0.30s	65/1100

2.1 Performance Drop Analysis

The drop in quality for Sports was substantial. TF-IDF achieved only 38.73% compared to Food & Drink's 59.92% - a 35.4% relative decrease. This represents 305 fewer correct matches, confirming that Sports articles are much harder to cluster using lexical similarity

LDA's performance was catastrophic for Sports, achieving just 5.91% compared to 38.77% for Food & Drink - an 84.7% drop. Only 65 out of 1,100 possible matches were correct, suggesting the 30-topic model completely failed to capture Sports category coherence.

2.2 Why Sports Performed Worse

There are several reasons why Sports showed significantly worse results than Food & Drink.

First, Sports has much more vocabulary diversity. Most of articles offered talk about different sports like basketball, soccer, gymnastics. These share almost no words, so TF-IDF can't find matches even though they're both sports content. Compare this to Food & Drink where terms like "burgers," "pizza," and "chicken" appear consistently across different articles.

The LDA failure (5.91%) is particularly revealing. With only 30 topics for the entire document dataset, different sports got scattered into separate topic clusters instead of grouping together. Basketball, gymnastics, and soccer each likely ended up in their own topics, completely fragmenting the Sports category. Food & Drink didn't have this problem because culinary vocabulary tends to co-occur naturally - recipes mention ingredients, cooking techniques appear alongside food names.

Finally, Sports is just a broader category overall. Food & Drink focuses on eating, cooking, and restaurants - all closely related domains. Sports covers team sports, individual athletes, competitions, business deals, and athlete profiles, each with completely different vocabularies.

Interestingly, LDA model creation for Sports (14.15s) was significantly faster than for Food & Drink (39.68s). This likely happened because the algorithm stopped early when topics couldn't stabilize properly - actually a warning sign of poor model fit rather than an advantage.

3.3 Execution Time Insights

Interestingly, LDA model creation for Sports (14.15s) was significantly faster than for Food & Drink (39.68s), likely due to stopping convergence earlier when topics couldn't stabilize properly. This faster training time is actually a warning sign of poor model fit.

3. Tag resorting methods discussion for "Food & Drink" description.

Finally, we explored using article tags instead of full text descriptions to calculate similarity for Food & Drink articles.

3.1 Results of methods:

We tested two tag-based approaches:

- Jaccard Tag Similarity: 61.88%
- Weighted Tag Similarity: 59.91%

Both performed well, with simple Jaccard actually achieving the best result of all methods we tested (slightly better than TF-IDF's 59.92%).

3.2 Advantages of Tag-Based Similarity:

Tags offer several advantages over content-based approaches. They represent human-curated or algorithmically extracted key concepts, reducing dimensionality compared to analyzing full text. Tags explicitly capture what an article is about rather than relying on word frequency statistics. They also filter out common words and focus on meaningful concepts, which reduces noise.

Computationally, comparing tag sets is much faster than computing TF-IDF similarities across 161,607 features or running LDA models. Tags can also be combined with content-based methods in a hybrid approach for potentially even better results.

3.3 Jaccard vs Weighted Approach

Interestingly, simple Jaccard similarity (61.88%) slightly outperformed our weighted approach (59.91%). We implemented the weighted method to give rarer tags more importance, similar to IDF weighting in TF-IDF. However, this suggests that for Food & Drink, common tags like "Taste," "bruger," or "baking" are actually valuable discriminators rather than noise. The weighting may have over-emphasized niche tags that appear in only a few articles, reducing the ability to find similar mainstream content.

3.4 Conclusion:

Tag-based similarity proved to be the best-performing method overall for Food & Drink categorization. While the improvement over TF-IDF was modest (about 2 percentage points), tags offer significant computational advantages and would likely show even greater benefits for categories like Sports where content-based methods struggled. For future work, a weighted approach might still be valuable, but the weights should be calibrated based on the specific category characteristics rather than blindly following TF-IDF's IDF weighting scheme.