CSE 422 Wi26 sample solutions

January 2026

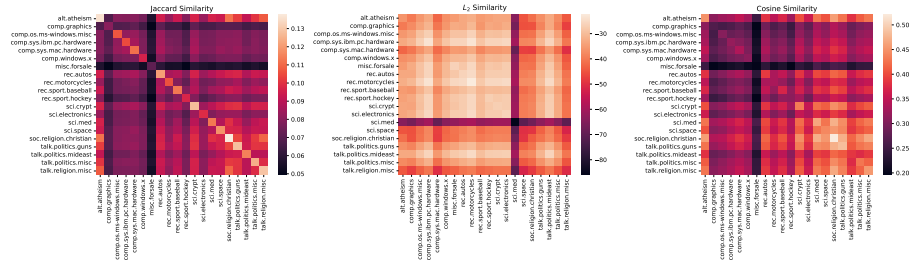
1 Homework 2

Problem 1

(By Eric Ye)

Part a

Here are the heatmaps



Part b

Based on the heatmaps, it seems that Cosine similarity is the most appropriate measure. Here's my observations across each of the similarity measures:

- **Jaccard Similarity:** Most of the values seem to be very small and sparse, with weak relationships across most groups (making the difference between the min and max large). This suggests that articles in the dataset have limited word overlap, making the denominator much larger than the numerator.
- **L_2 Similarity:** L_2 is strongly affected by the relative magnitudes of the counts of each word. Therefore, many of the entries in the heatmap are affected more by the relative length and word counts of the documents themselves rather than by any similarity.
- **Cosine Similarity:** By normalizing for document length, the cosine similarity depends on the direction of the feature vectors instead of the magnitudes, resulting in a more appropriate measure.

I see relatively high similarity scores between soc.religion.christian, talk.religion.misc, and alt.atheism. This might be because the vocabulary used by these religion related newsgroups have some overlap, such as "God".

Part c

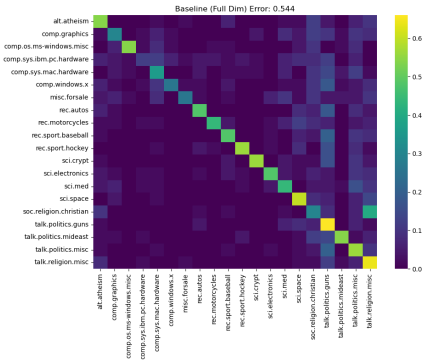
By injecting a hidden 21$^{\text{st}}$ newsgroup, every observed group would contain a mixture of $\frac{20}{21}$ of the original group, and $\frac{1}{21}$ of the hidden group. This should increase the between-group similarity, and decrease the similarity on the diagonals. This is what would happen to each similarity measure:

- **Jaccard Similarity:** Jaccard similarity measures token overlap. Since the original Jaccard similarity values are so small, even a small number of shared tokens across many documents can noticeably inflate the similarity, meaning that if the new newsgroup has a high concentration of common-vocabulary documents, it will cause strong global brightening.
- **L_2 similarity:** L_2 measures the magnitude and variance, in addition to just direction. This means that if the hidden newsgroup has high-norm or long documents, it can cause strong global brightening.
- **Cosine Similarity:** Cosine similarity normalizes vectors, so the small hidden group contributes a small shared direction, but unless it is highly aligned many groups, it is unlikely to affect the overall similarity measures, as cosine similarity preserves relative angular structure.

Therefore, Jaccard Similarity and L_2 similarity are likely to get noticeably affected if the hidden newsgroup has documents of a specific structure, but Cosine Similarity should not get affect as much.

Problem 2

Part a (by TA)



The average classification error rate is 0.544.

Part b (by Matthew Deng)

For each pair of articles, we compute the cosine similarity of two N dimensional vectors, which is $O(N)$. Iterating all articles while keeping track of the article of highest similarity makes the time complexity $O(nN)$.

Part c (by Matthew Deng and TA)

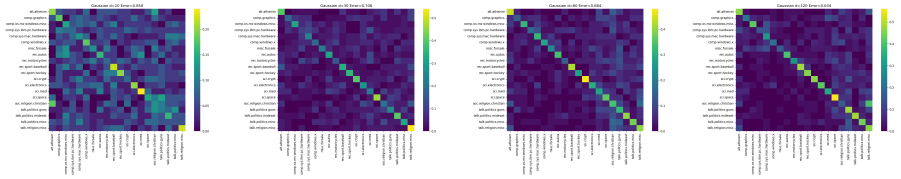


Figure 2: Gaussian

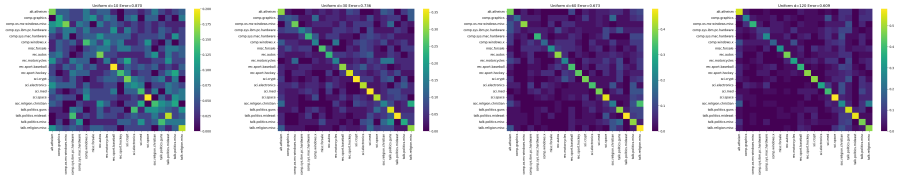


Figure 3: Uniform ± 1

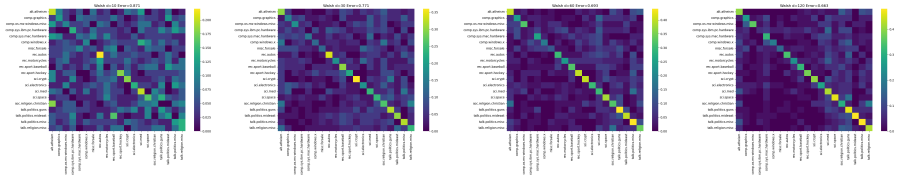


Figure 4: Walsh

	Dimension d			
Type	10	30	60	120
Gaussian	0.850	0.706	0.684	0.634
Uniform ± 1	0.870	0.736	0.673	0.609
Walsh	0.871	0.771	0.693	0.663

Table 1: Average classification error for each type of sketching and dimension

The results on dimension $d = 60, 120$ are comparable to the baseline, especially in the Gaussian and Uniform ± 1 sketching matrices.

Part d (Anonymous)

In general, for each value of d that we tested, the two randomized sketches perform similarly, while the deterministic sketch performs slightly worse.

This is because the two randomized sketching matrices give similar distributions when multiplied by the data and are more effective at retaining as much information as possible when going down into lower dimensions so their performance is similar (the Johnson-Lindenstrauss lemma gives strong guarantees for their performance). The deterministic sketch yields a much different distribution, which appears to be less effective at preserving distances than the randomized sketches.

By comparison, the deterministic matrix sketch is less effective at preserving distances due to its highly structured composition, which introduces external biases and patterns not reflected in the data. This causes the deterministic sketch to perform worse for the same value of d than the randomized sketches.

Part e (by Eric Yuxuan Ye)

Each matrix multiplication with the matrix and an article vector takes $\mathcal{O}(d \cdot N)$ time. Therefore, the preprocessing time is $\mathcal{O}(d \cdot N \cdot n + m)$ for the n articles.

For each query, we first need to preprocess the queried article, taking up $\mathcal{O}(d \cdot N)$, and for each comparison with the n articles in the database, we only need $\mathcal{O}(d)$ time to compute similarity. Therefore, the query time is $\mathcal{O}(d \cdot N + d \cdot n + m)$.

Part f (by An Yan)

Given M has rank at most d but N columns, we can say that the nullspace of M is at least $(N - d)$ dimensions.

Thus, we can use the basis vectors of nullspace of M (at least $(N - d)$ such points) and they all project to 0.

However, real data is distributed more evenly and is unlikely they are all aligned with the null space of M.

We will first prove that for all $N = 2^k$, where k is nonnegative integer, all rows in the larger $N \times N$ Walsh matrix H_N are linearly independent.

Proof. When $N = 1$, $[[1]]$ is linearly independent with rank 1. Then, we want to assume H_k is linearly independent from some k ≥ 1 and then we will prove H_{2k} has linearly independent rows and columns.

$$r_{\text{top}} = [u, u] \quad \text{and} \quad r_{\text{bot}} = [v, -v]$$

We examine the dot products of any pair of rows $r_i, r_j \in H_{2k}$:

- Top vs. Top: $[u, u] \cdot [v, v] = 2(u \cdot v) = 0$.
- Bottom vs. Bottom: $[u, -u] \cdot [v, -v] = (u \cdot v) + (-u \cdot -v) = 2(u \cdot v) = 0$.
- Top vs. Bottom: $[u, u] \cdot [v, -v] = (u \cdot v) - (u \cdot v) = 0$.

Since all distinct rows are orthogonal and contain non-zero entries (± 1), the set of N rows is linearly independent. Thus, $\text{rank}(H_N) = N$.

Thus, we can conclude by induction that matrix H has linearly independent rows and columns. □

Since all rows in H are linearly independent, we notice that the remaining $(N - d)$ rows not included in M are linearly independent to the d rows in M.

Thus, the remaining $(N - d)$ rows in H_N are a dataset such that $Mx = 0$, $\forall x \in X$.

Part g (by Vincent Chau)

Randomizing the sketch helps avoid these issues (vector projections being flattened, losing information, going to $\vec{0}$) because it randomizes the orientation of the nullspace. For any fixed dataset, it is highly unlikely many of those points fall close to the random nullspace, and equivalently, highly unlikely the random projection will "flatten" them. Distances/similarities can be preserved with high probability (by the JL-lemma). Deterministic sketches do not have this guarantee.