DAT470 Assignment 6

Olof Lindberg
Rikard Roos

March 2025

Problem 1

```
Time to normalize the dataset
    /data/courses/2025_dat470_dit066/glove/glove.840B.300d.txt:
2.99219012260437 seconds.
```

Problem 2

Query	Closest word	Second closest word	Third closest word
priest	priests	bishop	Priest
fork	forks	spoon	Fork
horse	horses	pony	Horse
beef	pork	meat	chicken
daoist	taoist	daoism	confucian
polish	nail	polishes	nails
vehicle	vehicles	car	automobile
crepe	Crepe	chiffon	crêpe
daytime	nighttime	day-time	night-time
scotland	wales	glasgow	scottish

Table 1: Closest word matches for each query term

```
matrix multiplication took 0.5483620166778564 s
sorting took 0.5947718620300293 s
total time 1.1431338787078857 s
```

Problem 3

```
Time to transform the dataset
    /data/courses/2025_dat470_dit066/glove/glove.840B.300d.txt
with D=50 hyperplanes:
6.331609010696411 seconds.
```

Problem 4

Locality-Sensitive Hashing

```
Hyperparameters: D = 50, k = 20, L = 10
fit took 84.96329164505005 seconds.
query took 0.07381701469421387 seconds.
```

Hyperparameter search

We evaluate the hyperparameter settings by changing one hyperparameter at a time. The results from $D = 50, k = 20, l = 10$ (baseline) are listed below for comparison. Compared to the "ground truth" in table 1, the results are quite good with these settings. While the closest words are not identical, the semantic meaning of the words are mostly the same. The most potential of improvement lie in the interpretation of daoist.

```
Hyperparameters: D = 50, k = 20, L = 10
priest: deacon shaman congregation
fork: forks spoons wishbone
horse: sled breeder llama
beef: meats roast burgers
daoist: hellenistic invlolved two-handers
polish: nail stain pedicure
vehicle: off-road roadway roads
crepe: brunch flour cooked
daytime: hour-long drowsy dizziness
scotland: scottish britain manchester
```

With $D = 100$ (Appendix, Listing 1), we increase the embedding into the Hamming space. However, we received weird results such as `fork: claws failed foon` and `daoist: skold 01:25:53 disembedded`. These results indicates that we need to take into account noise of the the data (embeddings), and have less precision in our calculations (i.e., increasing D leads to overfitting). Lowering D ($D = 20$) yielded good results(Appendix, Listing 2), in particular it improves the interpretation of daoist in comparison to the baseline (`daoist: taichi arboreous zoroastrian`).

The higher the k, the fewer collisions we get, and the stronger dependency of the correctness of the embeddings we get in the results. If our hypothesis of noisy data is correct, we would not want a high k. Decreasing k compared to the baseline (we tried $k = 10$) (Appendix, Listing 3) yielded better results in comparison to the baseline, in particular for daoist and daytime (`daoist: confucianism atropa buddist, priest: Priest deacon vicar`). Increasing k yields nonsensical results, as expected (Appendix, Listing 4). Several words have no neighbors, and the only words with neighbors with some semantic relation are beef and vehicle.

Increasing L increases the probability of success, and since we have lowered the accuracy of the data in D and k, we probably want to increase L. Setting $L = 20$ (Appendix, Listing 5) yielded similar but somewhat better results compared to the baseline (although while doubling the time it took to fit the data). In particular, the semantic interpretation is closer to the query. For example, priest yielded christian words (archbishop and church), horse yielded gelding, foal and colt. Furthermore, crepe yielded the fabrics gingham and corduroy, more similar to the interpretation of the ground truth counterpart. For $L = 5$ (Appendix, Listing 6), we decrease the accuracy of some of the results. For example, shaman is exchanged for celebrant for priest, and spoons is replaced for chromoly in fork. This is expected, as we decreased the number of buckets we look at when approximating the closest neighbor.

Combining the results from each parameter, we should seemingly decrease D and k, and increase L. For $D = 20, k = 10, L = 20$ (Appendix, Listing 7), we get excellent results. All the queries have the same semantic meaning of the closest neighbors, and fork, horse, beef, polish, and crepe yields identical results as the ground truth. This indicates that we could most probably fine tune the hyperparameters further to get the exact same top three results.

A Appendix

Listing 1: Query results with hyperparameters D=100, k=20, L=10

```
priest: Subdeacon minstrel kayaker
fork: claws failed foon
horse: car Hanoverian victorian
beef: burgers sandwiches chefs
daoist: skold 01:25:53 disembedded
polish: sticky cookware solids
vehicle: police under puncture
crepe: dessert buttercream desserts
daytime: sunrises late-season spurts
scotland: tyne atlantic dover
```

Listing 2: Query results with hyperparameters D=20, k=20, L=10

```
priest: Priest deacon pastor
fork: axle crankset unscrew
horse: Horse gelding foal
beef: lamb venison onions
daoist: taichi arboreous zoroastrian
polish: polishes lacquer glitter
vehicle: vehicles speeding pickup
crepe: satin pancake pastry
daytime: late-night restful t.v.
scotland: ireland halifax manchester
```

Listing 3: Query results with hyperparameters D=50, k=10, L=10

```
priest: Priest deacon vicar
fork: forks skewer tongs
horse: horses equine donkey
beef: meat steak meats
daoist: confucianism atropa buddist
polish: polishes manicure Polish
vehicle: vehicles automobile cars
crepe: chiffon blouse skirt
daytime: Daytime primetime naps
scotland: wales scottish edinburgh
```

Listing 4: Query results with hyperparameters D=50, k=30, L=10

```
priest: Diocese Uganda Faqir
fork: MichiganSite
horse:
beef: vegetables herb varieties
daoist: RayGun altenative rockchip
polish: lacquer  ON  Pieni
vehicle: roadways world-class ultra-competitive
crepe:
daytime: visions sonorous Lec-Lab-Credit
scotland: cremation ALCOHOLISM VELLA
```

Listing 5: Query results with hyperparameters D=50, k=20, L=20

```
priest: deacon archbishop church
fork: forks spoons wishbone
horse: gelding foal colt
beef: meats roast burgers
daoist: hellenistic invlolved two-handers
polish: nail gloss lacquer
vehicle: vehicles off-road roadway
crepe: omelette gingham corduroy
daytime: hour-long drowsy bedtimes
scotland: scottish ireland aberdeen
```

Listing 6: Query results with hyperparameters D=50, k=20, L=5

```
priest: deacon congregation celebrant
fork: forks wishbone chromoly
horse: sled shoe swim
beef: meats roast burgers
daoist: hellenistic invlolved cherr
polish: stain sanded coated
vehicle: roadway roads roadways
crepe: brunch flour cooked
daytime: dizziness wetting sensation
scotland: scottish norfolk essex
```

Listing 7: Query results with hyperparameters D=20, k=10, L=20

```
priest: bishop Priest ordained
fork: forks spoon Fork
horse: horses pony Horse
beef: pork meat chicken
daoist: taoist confucian budhist
polish: nail polishes nails
vehicle: vehicles automobile cars
crepe: Crepe chiffon crêpe
daytime: nighttime day-time nightime
scotland: wales scottish ireland
```