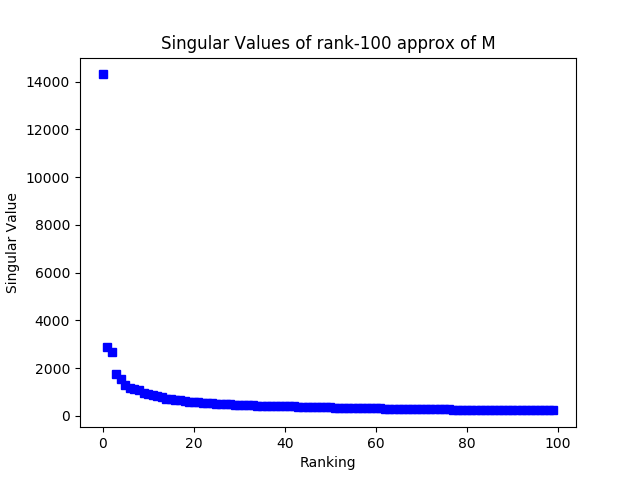
CS168 Project 5

1.

B.



Yes, M seems to be close to a low rank matrix because only the first few singular values are notably higher than the rest.

C.

The first and second singular vectors represent word frequency, the third represents males, the fourth represents music and tv show terminology and common lyrics, and the fifth represents social systems.

1st:

Top 10: the, and, of, in, to, for, as, is, was, by, with

Bottom 10: interred, midst, gf, peakposition, iucn, islander, increment, gmina, householder, insee

2nd:

Top 10: the, and, of, in, to, for, with, by, as, is, was

Bottom 10: rating, achievement, nba, positively, Olympics, brick, Methodist, happening, drugs, neighborhoods

3rd:

Top 10: born, john, james, jr, Richard, henry, brother, scott, steve, brian, William

Bottom 10: these, storage, provide, can, its, level, any, data, distribution, specific

4th:

Top 10: you, your, album, episode, love, girl, my, song, me, ep, can

Bottom 10: department, airport, council, government, union, united, regional, district, county, national

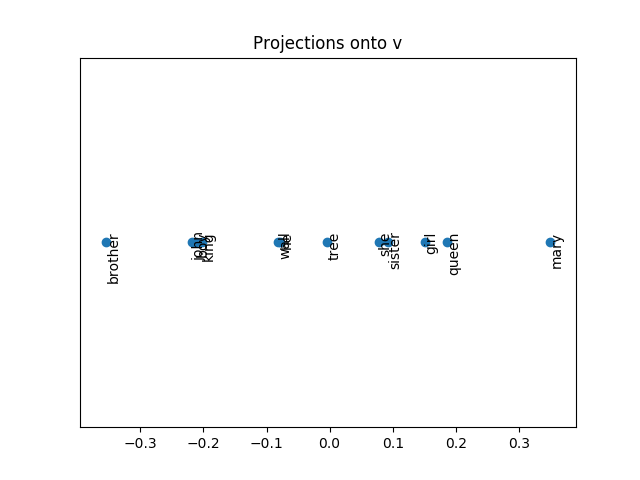
5th:

Top 10: political, government, social, policy, law, rights, minister, religion, religious, legal, committee

Bottom 10: blue, station, km, near, route, ft, bus, yellow, located, jpg

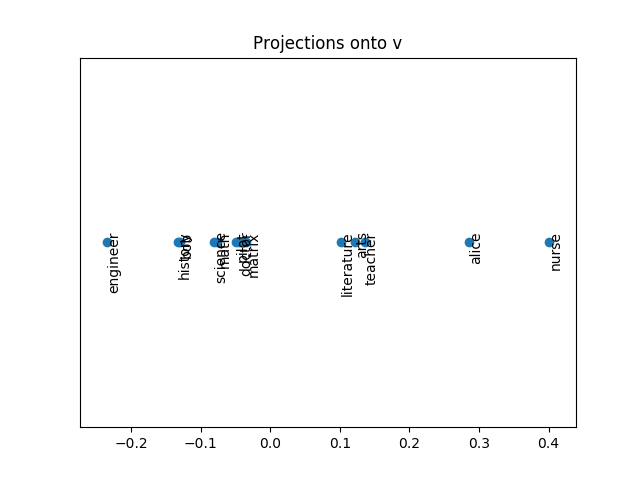
Not all of the singular vectors have vectors have easy-to-interpret semantics because most of the vectors have a dramatically smaller singular value than the top 5 vectors, meaning that the “concept” that those vectors represent are less strong and thus, not as visible in the data and not as easy to interpret. It could be the case that there is not really a concept to represent and it is much closer to randomness than a concrete concept. Additionally, the number of singular values is much smaller than the size of the corpus.

D.

i. 

The highest values are female-associated words, such as “Mary”, “queen”, “girl”, “sister”, and “she”. The lowest values are male-associated words, such as “brother”, “John”, “boy”, and “king”. “Tree”, being a gender-neutral word is near a value of 0, and “wall” and “he” are tied slightly closer to the male side.

ii.



Professions typically associated with women, such as “nurse”, “teacher”, “arts”, and “literature”, have positive values while professions typically associated with men, such as “engineer”, “history”, “science”, “math”, and “doctor” have negative values; since these professions have historically been dominated by one gender or another, Wikipedia articles reflect this by containing more articles with men as doctors and women as nurses, for example. This could be harmful; for example, if LinkedIn used these embeddings to improve their “search for qualified job candidates” option, men would automatically be seen as more qualified for engineering and math positions while women would be seen as more qualified for nursing and teaching positions before any consideration for the rest of the parts of their resumes.

E.

i. The words most similar to “stanford” are other highly ranked schools; “harvard”, “cornell”, “ucla”, “yale”, “princeton”, “penn”, “auburn”, “mit”, and “berkeley”.

ii. The accuracy was 55.5%. The model seems to have a hard time accurately working with city-state-city-? analogies. Our conjecture for why this is is that there word embeddings provided do not differentiate very well between different states, and as a result, all the states are clustered very closely together. This leads to easy misidentification of the correct answer for the cosine similarity metric.

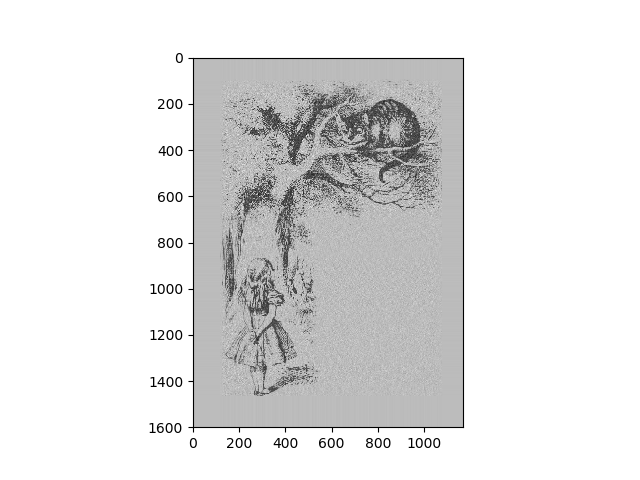
The model also appears to have a lot of errors when dealing with nationality. From examining the errors, many of the misclassifications seem to come because the model outputs an answer that is a similar nationality to the one expected. For example, when expecting “Swedish”, the model frequently gives “Danish” or “Norwegian” as an answer because these three countries are very close culturally and geographically.

On the other hand, the model performs very well with capital-country analogies and familial relationships.

2.

A. We expect that for any column or row that does not include the moon, there will be a completely black column. This is because there is no need to capture anything else besides the black pixels. For the rows and columns that do include the moon, we expect them to be varying shades of gray depending on how much of the moon they must include in the approximation (and the color of the moon itself). This is because since this is only the rank 1 approximation, the approximation is unable to capture much more than the distribution of white/black pixels in each row and column.

B.



C. We stopped at 1170 because the dimensions of V are 1170 by 1170, meaning that there are only up to rank 1170 approximations. If there were more than 1170, it would also attempt to make the original matrix more complex than it actually is and take away from the idea of image compression.

D.

With the Rank 150 approximation, it is about 4.5 times more memory efficient.

E. The background noise is the small values calculated with the current part of the matrices that are used in the current approximation. They have to be canceled out by the smaller singular values and vectors, which is why a larger k approximation leads to a decrease in the noise.