Alan Chen

1. **How does the value of k affect the accuracy of your classifier? Provide a table or plot that shows how accuracy varies with numerous values of k. How do you explain these results?**

k value	accuracy
2	94%
7	94%
10	94%
20	95%
40	94%
100	93%
120	92%

The K value is highly dependent on your dataset. K should be big enough that some outliers won't affect the prediction highly and small enough that one factor will not dominate over another. Some claims that square root of the length of dataset is a good number, but for this case I tried many values of K to test the best result. Based on my test, k=20 gives the best results. When k is small (like 2) its prediction is highly correlated to its neighbors (2 in this case). This will not be an accurate estimate as we want to consider more than two neighbors to make a precise decision. We also do not want to increase the value of k since it will begins averaging having a majority voting affect.

2. Which 10 members of Congress are the most similar to Representative Jim McGovern, who represents Amherst as part of Massachusetts's 2nd Congressional District? Explain how you determined this, including your choice of k.

1. Janice "Jan" Schakowsky
2. José Serrano
3. Adriano Espaillat
4. Ro Khanna
5. Jamie Raskin
6. Alan Lowenthal
7. Bonnie Watson Coleman
8. Alcee Hastings
9. Pramila Jayapal
10. Sander Levin

I assigned all representative a "distance" (how likely they are going to vote the same) from James "Jim" McGovern. The distances are determined by similarity. For example, if there are 100 bills to vote on and 2 representatives always vote the opposite, then their distance is 100. Then, I chose the top 10 representatives with the lowest distance. I chose k to be 20 as it was giving me the best accuracy from question 1.

3. **The fourth column of the has the party affiliation of each member. Using your code for determining nearest neighbors, determine the top 3 Democrats and top 3 Republicans who are the most dissimilar from others in their party. Explain how you scored them to determine the ranking**

Steny Hoyer – Democrat
Eddie Johnson – Democrat
Niki Tsongas – Democrat

Kevin McCarthy - Republican
Lou Barletta – Republican
David Young - Republican

First, I increased the K (= 70) value to get an "averaging" effect of the two parties. Then I chose the names that have an incorrect prediction. This works because k-nn classifies a vote based on surrounding votes. If it is incorrect it means that it is in a grey area (a sweet spot between the two parties).