1-e) Run your linear predictor with feature extractor extractCharacterFeatures. Experiment with different values of n to see which one produces the smallest test error. You should observe that this error is nearly as small as that produced by word features. Why is this the case?Construct a review (one sentence max) in which character n-grams probably outperform word features, and briefly explain why this is so. Note: There is a function in submission.py that will allow you to test different values of n. See the Docstring of testValuesOfN(n) how to run it. Remember to write your final written solution.

The lowest validation error is achieved for n=7, at 0.27068. This might be because this n-gram length manages to catch specific patterns in the reviews. On the flip side, character n-grams outperform word features in this review:

“Men in Black is a amezing movie! It's abot these guys who are secrit agents fighting against alien invesions in New York. The chracters are so cool, especially Will Smoth and Tommy Li Jones. The soryline is gr8 and the spects and effets are mind blowing! It's defenetly one of my all time favrits.”

This is because we have many spelling mistakes and poorly written things in it, so the n-grams might capture more detailed information in the review.

2a) Suupose we have a feature extractor $\emptyset$ that produces 2-dimensional feature vectors, and a toy dataset $\boldsymbol{D_{train}} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}, \boldsymbol{x_4}\}$ with

- $\emptyset(\boldsymbol{x_1}) = [\mathbf{10}, \mathbf{0}]$
- $\emptyset(\boldsymbol{x_2}) = [\mathbf{30}, \mathbf{0}]$
- $\emptyset(\boldsymbol{x_3}) = [\mathbf{10}, \mathbf{20}]$
- $\emptyset(\boldsymbol{x_4}) = [\mathbf{20}, \mathbf{20}]$

Run 2-means on this dataset until convergence. Please show your work. What are the final cluster assignments $\boldsymbol{z}$ and cluster centers $\boldsymbol{\mu}$? Run this algorithm twice with the following initial centers:

i) $\boldsymbol{\mu_1} = [\mathbf{20}, \mathbf{30}]$ and $\boldsymbol{\mu_2} = [\mathbf{20}, -\mathbf{10}]$

for $k = 1$ and μ_1

distances

$$d_1 = \sqrt{(10-20)^2 + (0-30)^2} = 31.62$$

$$d_2 = \sqrt{(30-20)^2 + (0-30)^2} = 31.62$$

$$d_3 = \sqrt{(10-20)^2 + (20-30)^2} = 14.14$$

$$d_4 = \sqrt{(20-20)^2 + (20-30)^2} = 10$$

for $k = 1$ and μ_2

$$d_1 = \sqrt{(10-20)^2 + (0+10)^2} = 14.14$$

$$d_2 = \sqrt{(30-20)^2 + (0+10)^2} = 14.14$$

$$d_3 = \sqrt{(10-20)^2 + (20+10)^2} = 31.62$$

$$d_4 = \sqrt{(20-20)^2 + (20+10)^2} = 30$$

Assignments z

$$\mu_1: x_4$$

$$\mu_2: x_1, x_2$$

new centers

$$\mu_1 = (20{,}20)$$

$$\mu_2: \left(\frac{10+30}{2}, \frac{0+0}{2}\right) = (20{,}0)$$

for $k = 1$ and μ_1

distances

$$d_1 = \sqrt{(10-20)^2 + (0+20)^2} = 22.36$$

$$d_2 = \sqrt{(30-20)^2 + (0+20)^2} = 22.36$$

$$d_3 = \sqrt{(10-20)^2 + (20+20)^2} = 10$$

$$d_4 = \sqrt{(20-20)^2 + (20+20)^2} = 0$$

for $k = 2$ and μ_2

$$d_1 = \sqrt{(10-20)^2 + (0+0)^2} = 10$$

$$d_2 = \sqrt{(30-20)^2 + (0+0)^2} = 10$$

$$d_3 = \sqrt{(10-20)^2 + (20+0)^2} = 22.36$$

$$d_4 = \sqrt{(20-20)^2 + (20+0)^2} = 20$$

Assignments z

$$\mu_1 : x_4$$

$$\mu_2 : x_1, x_2$$

new centers

$$\mu_1 = (20{,}20)$$

$$\mu_2 : \left(\frac{10+30}{2}, \frac{0+0}{2}\right) = (20{,}0)$$

i) $\mu_1 = [0{,}10]$ and $\mu_2 = [30{,}20]$

for $k = 1$ and μ_1

distances

$$d_1 = \sqrt{(10-0)^2 + (0-10)^2} = 14.14$$

$$d_2 = \sqrt{(30-0)^2 + (0-10)^2} = 31.62$$

$$d_3 = \sqrt{(10-0)^2 + (20-10)^2} = 14.14$$

$$d_4 = \sqrt{(20-0)^2 + (20-10)^2} = 22.36$$

for $k = 1$ and μ_2

$$d_1 = \sqrt{(10-30)^2 + (0+20)^2} = 20$$

$$d_2 = \sqrt{(30-30)^2 + (0+20)^2} = 0$$

$$d_3 = \sqrt{(10-30)^2 + (20+20)^2} = 28.28$$

$$d_4 = \sqrt{(20-30)^2 + (20+20)^2} = 22.36$$

Assignments z

$$\mu_1 : x_1, x_3$$

$$\mu_2 : x_2$$

new centers

$$\mu_1 = \left(\frac{10+10}{2}, \frac{0+20}{2}\right) = (10{,}10)$$

$$\mu_2 = (30{,}0)$$

for $k = 1$ and μ_1

distances

$$d_1 = \sqrt{(10-10)^2 + (0+10)^2} = 10$$

$$d_2 = \sqrt{(30-10)^2 + (0+10)^2} = 22.36$$

$$d_3 = \sqrt{(10-10)^2 + (20+10)^2} = 10$$

$$d_4 = \sqrt{(20-10)^2 + (20+10)^2} = 14.14$$

for $k = 2$ and μ_2

$$d_1 = \sqrt{(10-30)^2 + (0+0)^2} = 20$$

$$d_2 = \sqrt{(30-30)^2 + (0+0)^2} = 0$$

$$d_3 = \sqrt{(10-30)^2 + (20+0)^2} = 28.28$$

$$d_4 = \sqrt{(20-30)^2 + (20+0)^2} = 22.36$$

Assignments z

$$\mu_1: x_1, x_3$$

$$\mu_2: x_2$$

New centers

$$\mu_1 = \left(\frac{10 + 10}{2}, \frac{0 + 20}{2}\right) = (10,10)$$

$$\mu_2 = (30,0)$$