CSCI 720 - Project 2: Tweet Label Distributions and K-Means

Jimenez Guzman, Alejandrina
aj7354@rit.edu

Santurkar, Vaibhav
vs4503@rit.edu

December 9, 2021

1 Task 4

1.1 Graphs generated by pldl.py

Observe Figures 1 and 2.

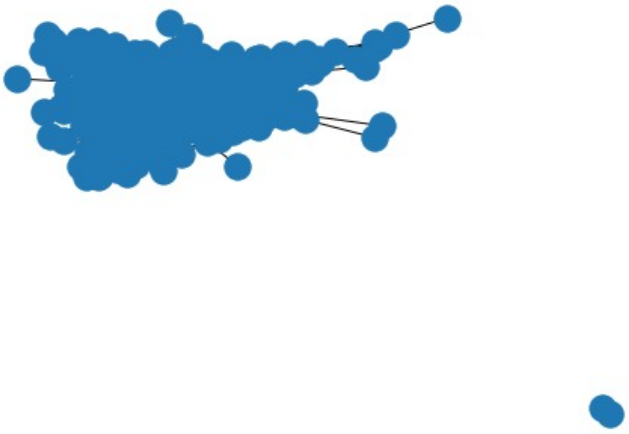

Figure 1: Label Distributions Graph (Initial Data set)- 1st Run

2 Task 5 and 6

2.1 Graph Data and Degree Histogram - 1st Run

- Number of nodes in graph: 757

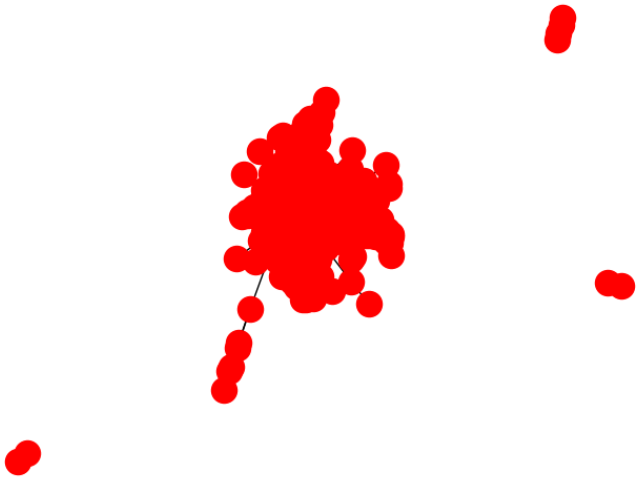

Figure 2: Label Distributions Graph (Enhanced Data Set)- 2nd Run

- Number of edges in graph: 12462
- Number of connected edges in graph = 2
- Density of graph = 0.04355119414564593

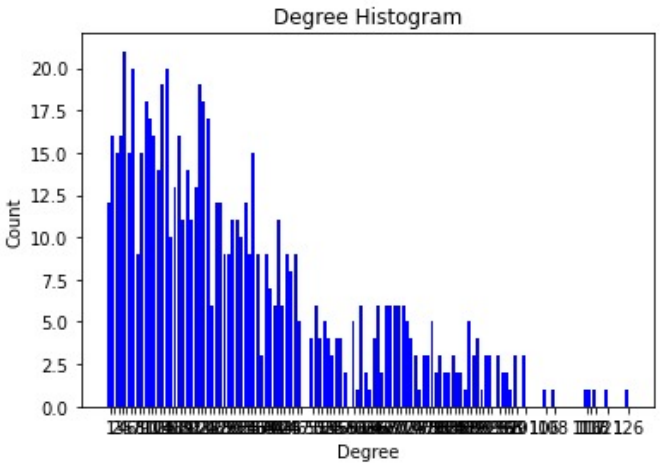

Figure 3: Degree Histogram (Initial Data set)

2.2 Graph Data and Degree Histogram - 2nd Run

- Number of nodes in graph: 1034

- Number of edges in graph: 14246
- Number of connected edges in graph = 4
- Density of graph = 0.026674855494035324

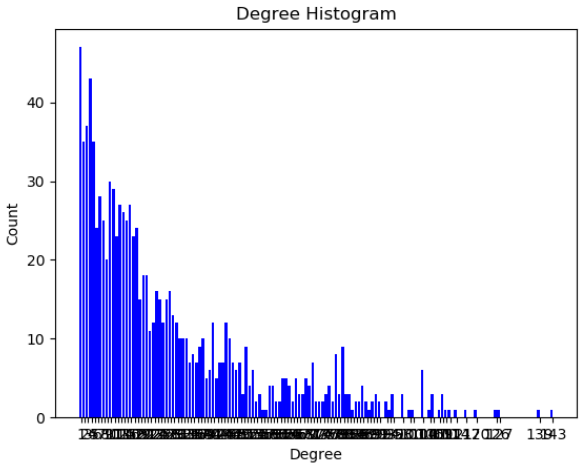

Figure 4: Degree Histogram (Enhanced Data set)

3 Task 7

Cohen's kappa coefficient (k) is a statistic to measure the reliability between annotators for qualitative (categorical) items. It is a more robust measure than simple percent agreement calculations, as k takes into account the possibility of the agreement occurring by chance. It is a pairwise reliability measure between two annotators. In this project, we have several labels that can be assigned to a same message, but Cohen's Kappa is not designed for multiple choices for a single item by two annotators. Therefore, we decided to calculate Cohen Kappa for each of the labels separately and then obtained the average. These are the results:

- Label 0: 0.5268
- Label 1: 0.5769
- Label 2: 0.271
- Label 3: 0.5689
- Label 4: 0.6751

- Label 5: 0.5684
- Label 6: 0.623
- Label 7: -0.0114
- Label 8: 0.3272
- Label 9: 0.5513
- Label 10: 0.0281
- Label 11: 0.3017
- **Final average:** 0.41725

This result shows that reached a moderate agreement in our annotations. We can also show the result of our evaluations in the confusion matrices presented in Figure 5.

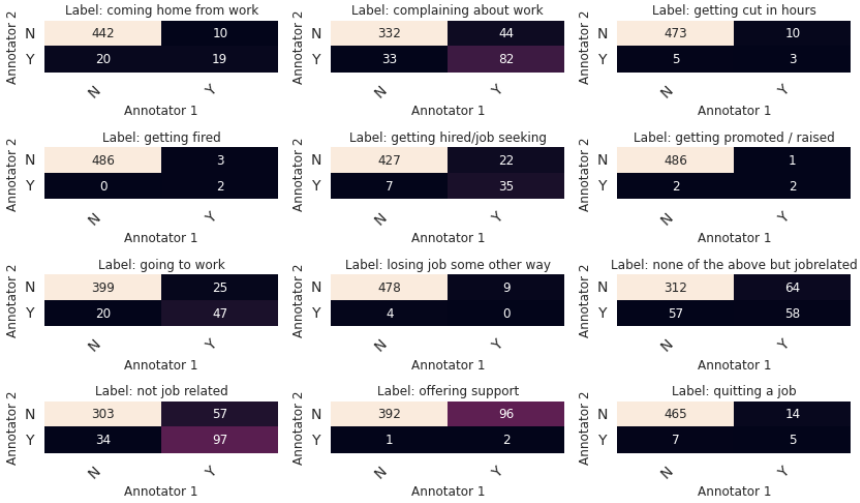

Figure 5: Confusion Matrices for the annotations.

4 Task 8

From our results of Taks 7, we were able to see that we had reached a low level of agreement for tweets under label 8, thus we focused our attention in 10 of those mismatched items.

Message	losing job some other way (a - Vaibhav)	losing job some other way (a - Alejandrina)
Is it bad that the only thoughts going thru my head are how i want to burn down mcdonalds so i dont have to go to work?	1	
I really should of went to work		1
@SOMEONE your job and thanks for not texting back	1	
I'm the only one to get the job done iont know no bitch that can cover for me !	1	
"@SOMEONE A month off...need a job		1
I should of went to work today		1
@SOMEONE no idea i was at work😩	1	
i had your job before i would never behave that way	1	
looks like rg3 lost his starting job ahahaha	1	
Prosecutor in LDC case leaving for new job http://LINK	1	
"@SOMEONE @SOMEONE lol no it's not gon work u might as well quit Monday comin up lol" 😂😂😂😂😩 we'll see I'm still live though lol	1	

Figure 6: Tweet Disagreement Table.

4.1 Vaibhav's Tweet Description of tweets disagreed with Alejandrina with respect to label 8.

1. The person wants to burn down his/her place of work, which is a way of losing their job
2. This person is talking about going to work and regretting having not gone
3. The tweet implies that the person lost a job and was not texted about it
4. The person is angry that no one else can do her job, indicating she lost her job already
5. This person is looking for a job and not about losing his/her job
6. This person is also regretting having not gone to work
7. The text implied that no one in higher position than the tweeter was aware he was in the office even after being laid off
8. The tweet indicates the person has already lost the job and is warning the next employee
9. The words 'lost his starting job' clearly indicate that the person lost his job
10. The words 'leaving his job' indicates the prosecutor lost his job
11. The person is complaining about the work conditions and advising future employees from applying, which could be a cause for losing their job.

4.2 Alejandrina's Tweet Description of tweets disagreed with Vaibhav with respect to label 8.

1. The person wants to burn down his/her place of work, which means that she is complaining about this job but will not necessarily act upon these wishes of burning the place down.
2. This tweet implies that the person might have lost their job because they did not go to their job establishment.
3. Here there is not a clear relationship to a specific job, this seems more like a complaint to someone who did not show up to something.
4. The person is angry that no one else can do her job, therefore she cannot rest or be absent so is complaining that no other person is as qualified to do the job and it is tiring.
5. This person might have lost their job for some reason a month ago, therefore they need to find a new one.
6. Very similar to the second tweet.
7. This person did not know about some event because she was working.
8. This person is either given advice or criticizing a person who is currently working in a position that the person previously worked.
9. Not job related
10. The prosecutor quit their job in order to start in a better position in a different job.
11. Complaining about work.

We did not decide to relabel our data. This is because from our Cohen Kappa's values we noticed that we already agreed on approximately 75% of the tweets we labelled. Therefore, to preserve the diversity of the data set we chose not to relabel the data.

5 Task 9

Clustering is a technique that allows us to find groups of similar objects, objects that are more related to each other than to objects in other groups. For this task we decided to use the K-means clustering algorithm employing the label distributions as features. The term "k-means" was first used by James MacQueen[1], and is a method of vector quantization, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest cluster centroid, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells.

We decided to use this algorithm because it is extremely easy to implement and is also computationally very efficient compared to other clustering algorithms. While k-means is very good at identifying clusters with a spherical shape, one of the drawbacks of this clustering algorithm is that we have to specify the number of clusters, k, in advance. An inappropriate choice for k can result in poor clustering performance. We were able to set k to 12 because we had the information about the number of labels present in the dataset, therefore the performance of the algorithm was good. We saved the clustered data in a .cvs file that is generated by the script that performs tasks 9 and 10. A visualization of the data with only two labels is present in a following section.

6 Task 10

For this task we concatenated the tweets in each cluster into a giant "meta tweet", and were able to get the following results after some processing:

6.1 Most Frequent Words - 1st Run

- Cluster 1: [('work', 58), ('job', 17), ('someone', 15), ('day', 10), ('today', 10), ('hate', 7), ('go', 7), ('lt', 6), ('come', 6), ('tonight', 5)]
- Cluster 2: [('work', 63), ('someone', 31), ('job', 25), ('today', 10), ('come', 7), ('visit', 5), ('manager', 5), ('gon', 4), ('kid', 4), ('always', 4)]
- Cluster 3: [('someone', 89), ('work', 30), ('job', 14), ('go', 8), ('ue106', 8), ('one', 7), ('day', 6), ('best', 6), ('love', 6), ('good', 6)]
- Cluster 4: [('job', 33), ('someone', 17), ('new', 7), ('call', 6), ('interview', 6), ('really', 6), ('back', 5), ('hire', 5), ('start', 4), ('hope', 4)]
- Cluster 5: [('work', 48), ('someone', 36), ('job', 33), ('go', 10), ('new', 7), ('today', 6), ('time', 6), ('day', 6), ('im', 5), ('shift', 5)]
- Cluster 6: [('work', 49), ('job', 22), ('someone', 20), ('go', 11), ('day', 9), ('today', 7), ('bitch', 5), ('time', 5), ('fuck', 5), ('school', 5)]
- Cluster 7: [('work', 67), ('someone', 28), ('go', 18), ('job', 12), ('day', 10), ('today', 9), ('30', 7), ('tomorrow', 6), ('come', 6), ('lt', 6)]
- Cluster 8: [('work', 53), ('someone', 17), ('home', 13), ('go', 8), ('come', 7), ('today', 6), ('gon', 4), ('fresh', 4), ('finally', 3), ('watch', 3)]
- Cluster 9: [('work', 47), ('go', 19), ('someone', 8), ('wan', 7), ('day', 7), ('job', 5), ('really', 5), ('hour', 5), ('feel', 5), ('11', 4)]

- Cluster 10: [('someone', 44), ('work', 41), ('job', 22), ('say', 9), ('go', 8), ('day', 7), ('come', 6), ('good', 6), ('look', 5), ('exit', 5)]
- Cluster 11: [('work', 65), ('go', 22), ('someone', 21), ('ready', 8), ('hour', 8), ('ta', 7), ('tomorrow', 6), ('gon', 5), ('minute', 4), ('lay', 4)]
- Cluster 12: [('work', 514), ('someone', 360), ('job', 208), ('go', 102), ('day', 69), ('today', 68), ('come', 47), ('time', 37), ('good', 33), ('really', 31)]

6.2 Most Frequent Words - 2nd Run

- Cluster 1: [('someone', 68), ('work', 55), ('job', 24), ('good', 8), ('great', 7), ('exit', 7), ('today', 6), ('bridge', 6), ('come', 5), ('go', 5)]
- Cluster 2: [('work', 50), ('job', 15), ('day', 10), ('today', 9), ('someone', 8), ('lt', 6), ('hate', 6), ('ta', 5), ('time', 5), ('tonight', 4)]
- Cluster 3: [('work', 83), ('someone', 52), ('job', 44), ('today', 14), ('come', 7), ('gon', 6), ('good', 6), ('call', 6), ('night', 6), ('manager', 6)]
- Cluster 4: [('work', 47), ('go', 21), ('someone', 8), ('wan', 6), ('day', 6), ('job', 6), ('really', 5), ('hour', 5), ('feel', 5), ('11', 4)]
- Cluster 5: [('work', 68), ('someone', 28), ('go', 18), ('job', 12), ('time', 12), ('day', 8), ('come', 6), ('lt', 6), ('tomorrow', 6), ('today', 6)]
- Cluster 6: [('someone', 76), ('work', 22), ('job', 10), ('go', 9), ('ue106', 8), ('day', 6), ('best', 5), ('amp', 5), ('break', 5), ('water', 5)]
- Cluster 7: [('work', 59), ('someone', 32), ('job', 21), ('go', 13), ('day', 9), ('fuck', 8), ('today', 7), ('amp', 5), ('come', 5), ('hour', 4)]
- Cluster 8: [('work', 68), ('go', 22), ('someone', 20), ('ready', 8), ('ta', 8), ('hour', 7), ('tomorrow', 6), ('gon', 5), ('back', 5), ('wan', 4)]
- Cluster 9: [('job', 40), ('someone', 18), ('new', 9), ('call', 7), ('interview', 6), ('really', 6), ('go', 6), ('back', 5), ('hire', 5), ('hope', 4)]
- Cluster 10: [('work', 61), ('someone', 56), ('job', 29), ('go', 10), ('say', 9), ('day', 8), ('today', 7), ('good', 7), ('come', 6), ('im', 6)]
- Cluster 11: [('work', 60), ('someone', 45), ('job', 23), ('go', 18), ('today', 9), ('time', 8), ('come', 7), ('right', 5), ('day', 5), ('new', 5)]
- Cluster 12: [('work', 582), ('someone', 380), ('job', 203), ('go', 123), ('today', 70), ('day', 68), ('come', 48), ('time', 40), ('good', 35), ('gon', 34)]

7 Task 11

For this task we plotted our own graph and colored the clusters, since the graph in task 4 was too complicated for us to modify.

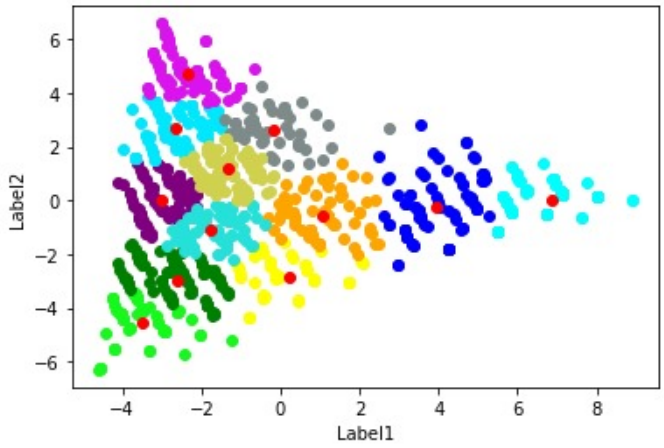

Figure 7: Label Distributions Graph (Initial Data set)- 1st Run

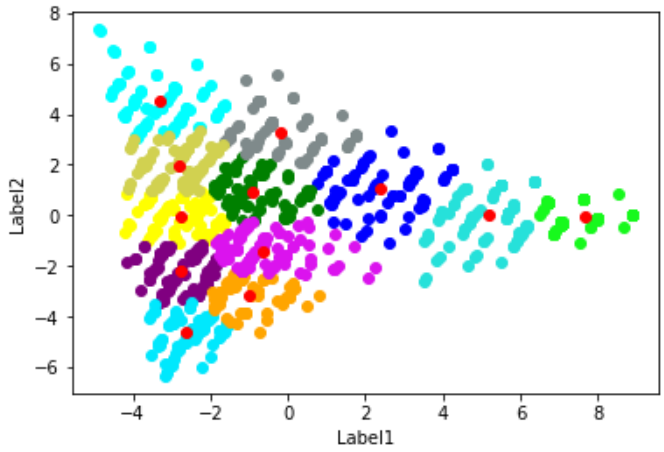

Figure 8: Label Distributions Graph (Enhanced Data set)- 2nd Run

8 Task 15

From the first graph, we observed that the data points were concentrated , i.e, the distance between individual points were very low. This implies that these data points are very

closely related in terms of their label distributions. In the second graph with the enhanced data, we noticed the points were more spread out because of the extra annotations that we used in the second run. In terms of similarities, both graphs have a similar distribution of data points. Both graphs have a vaguely triangular distribution of points. Overall the clusters retain approximately the same data points in both runs with a few minor changes in the distance.

9 Appendix

Note: By **second run** and **enhanced dataset**, we mean the results obtained after performing task 13.

References

[1] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. Le Cam and J. Neyman, editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.