CSC 296

Introduction to Artificial Intelligence

Instructors:

Eduardo Blanco

Kobus Barnard

TA:

Hojae Lee

Assignment #7

Jose Salinas Meza

12/07/2025

1. What are the Most Frequent Words in Each Dataset?

1.1 TED

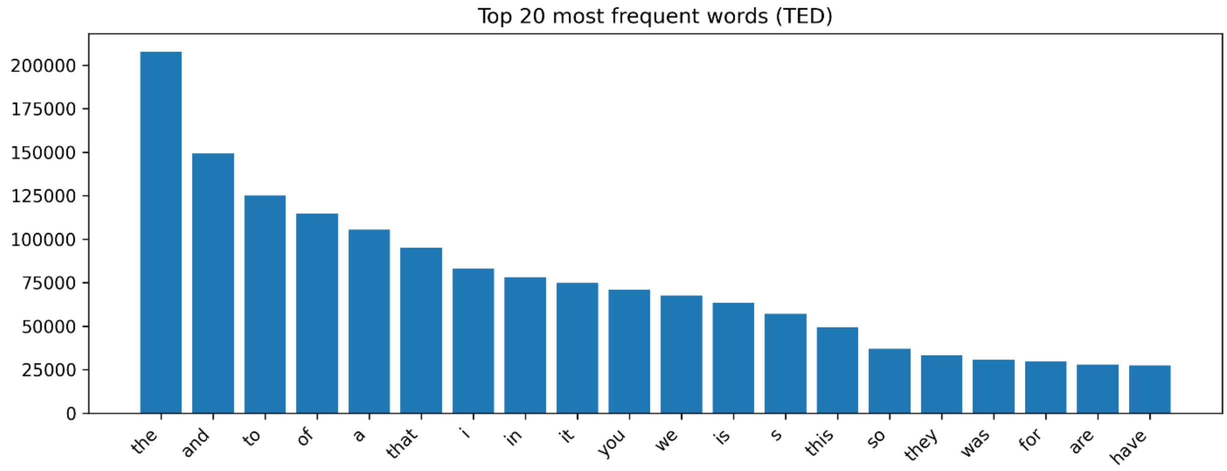

Figure 1: "TED – Dataset Top 20 Most Frequent Words."

1.2 WIKI

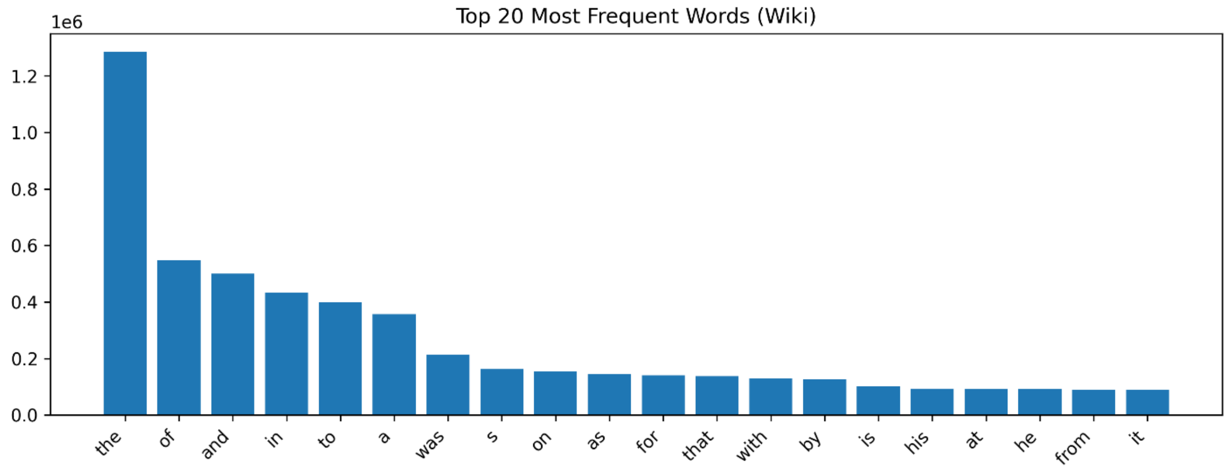

Figure 2: "WIKI – Dataset Top 20 Most Frequent Words."

1.3 Conclusion

As shown in the graphs of the most common words in both datasets, "the" is the clear winner. After that, from 2nd to 20th position, words are in different positions, and some words are at the top that the other dataset doesn't have, but the words at the top are in both datasets.

For word counts, there is a significant difference; it could be due to one dataset being larger than the other. However, if we roughly estimate percentages for each word based on the word counts in each dataset, they should be on a similar scale.

2. Similar Single Words

2.1 TED - Man

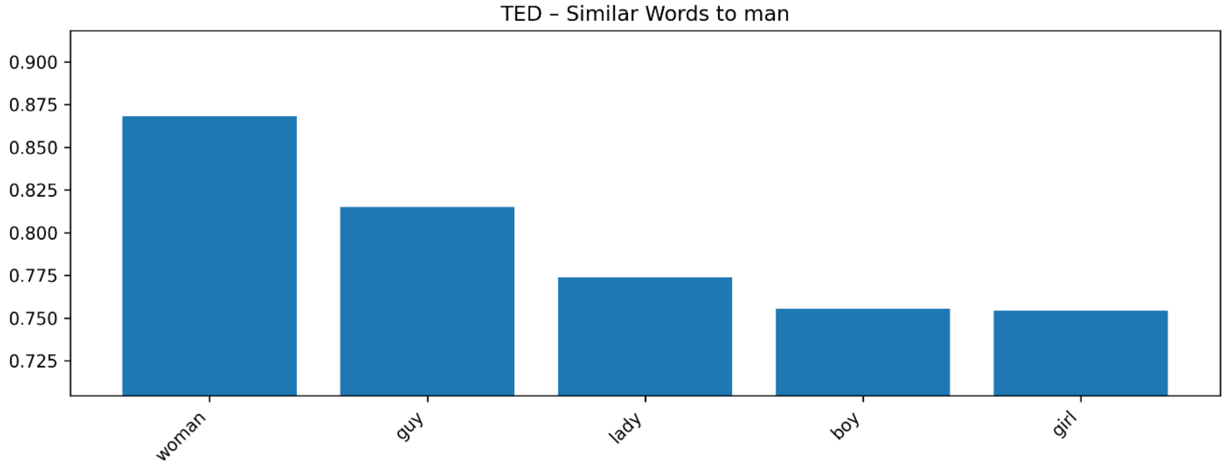

Figure 3: “TED – Dataset Similar Words to Man.”

2.2 TED - Woman

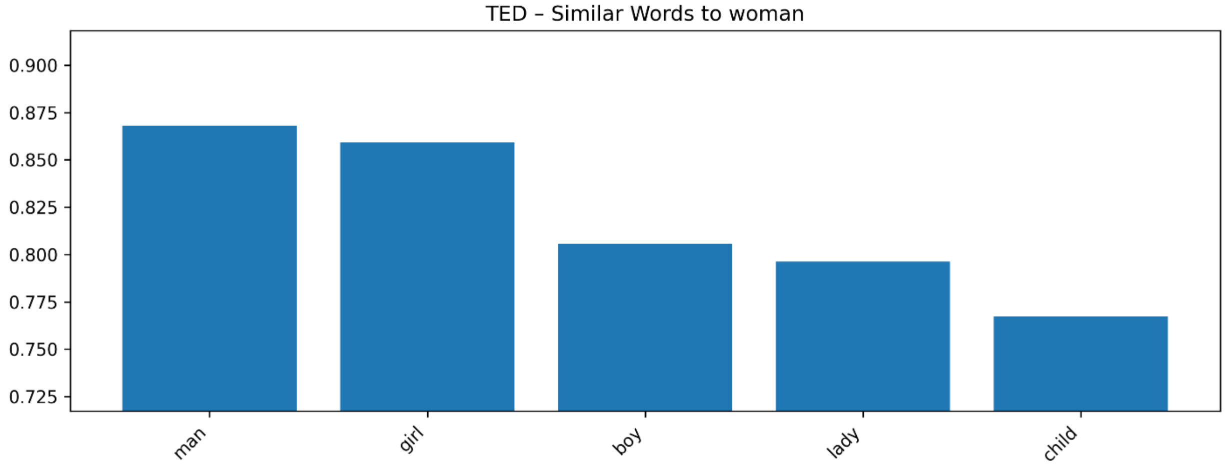

Figure 4: “TED – Dataset Similar Words to Woman.”

2.3 WIKI - Man

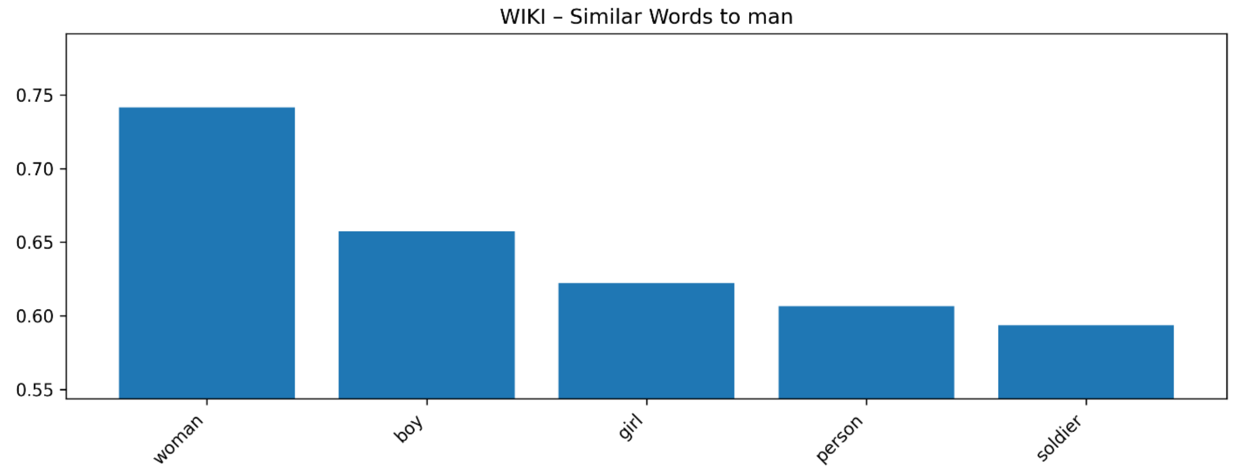

Figure 5: “WIKI – Dataset Similar Words to Man.”

2.4 WIKI - Woman

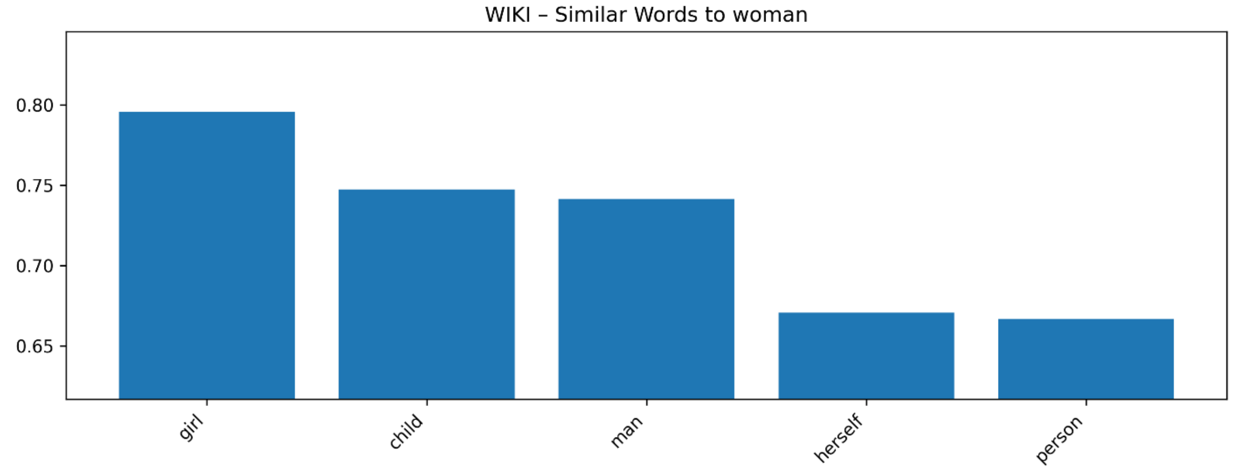

Figure 6: “WIKI – Dataset Similar Words to Woman.”

2.5 Conclusion

In both cases, we see that the words differ across datasets. It could be due to context: some words in one dataset might not appear in the other, so it replaces them with another word.

As we can see, for the TED dataset, similar words to “man” are “woman, guy, lady, boy, girl.” Similar words to “woman” are: “man, girl, boy, lady, child.”

The relationship between each word for this dataset: “man – woman” more than sharing letters they are related as one is used to refer to males and the other to females under same context. “man – guy” relationship here is that both words can bee used to refer to a male person depending on context same can be said for “boy”. “man – lady” same case as “man – woman” as lady is another word to refer to “woman” same case for “girl.” In the case of similar words to “woman” share similar similarities to “man” the only different word is “child” which makes sense as “woman” usually carry childs with them so we can see the relationship there.

For the WIKI dataset, it says that similar words to “man” are: “woman, boy, girl, person, soldier.” Similar words to “woman” are: “girl, child, man, herself, person.”

In this dataset, we can see that some words similar to “man”, such as “soldier”, could imply that the dataset covers military topics and that “man” is related to this word, as when men join the military, they become soldiers. In the case of “woman,” we have two different words in the other dataset, “herself” and “person,” which are related to “woman” depending on context.

3. TED – Similar Words

3.1 Nice + Man

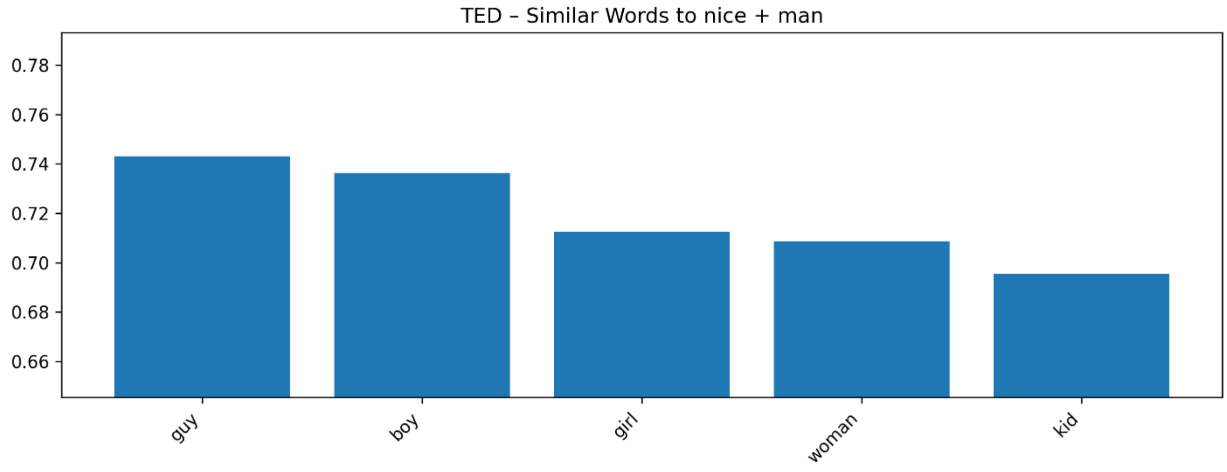

Figure 7: “TED – Dataset Similar Words to Nice + Man.”

3.2 Nice + Woman

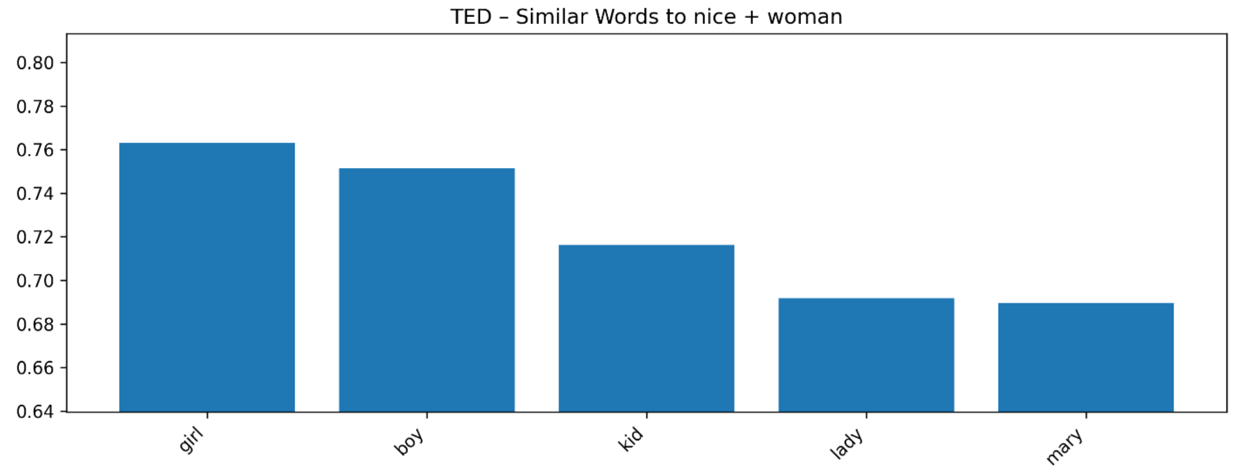

Figure 8: "TED – Dataset Similar Words to Nice + Woman."

3.3 Conclusion

In this step we checked for similar words to "nice + man/woman" and the results are pretty similar to the original without adding the extra word "nice" which I do think it makes sense as "nice" is an adjective that modify the noun "man/woman" so it puts words related to them but that have a nice meaning, taking out words like "soldier" which isn't very nice as is related to war and that stuff.

4. t-SNE Visualization

4.1 TED

For this dataset, I have marked the five spots that appear to me to be the locations where a cluster should be, as this is where nodes are most concentrated. I do believe these clusters make sense.

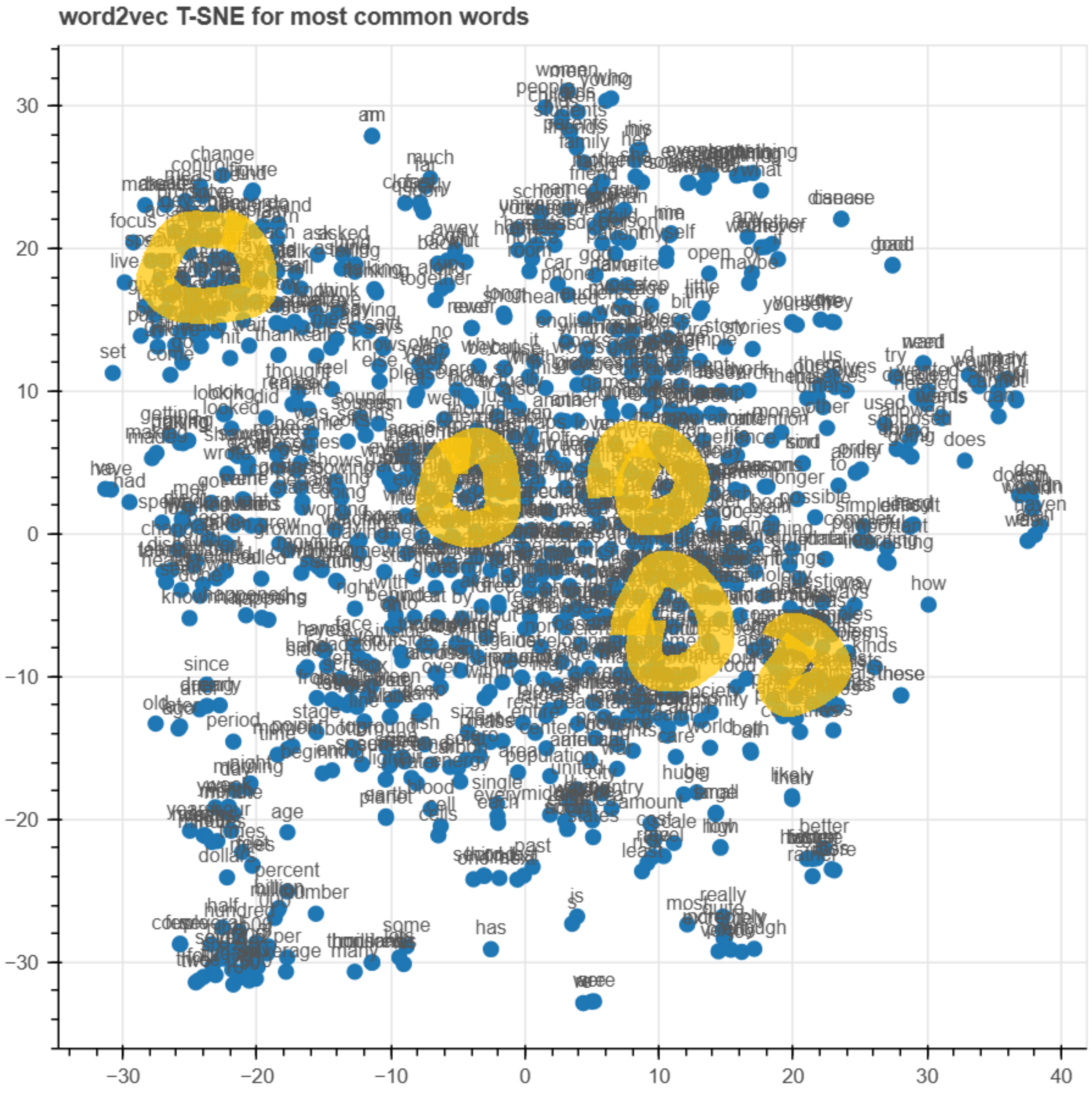

Figure 9: “TED – Dataset t-SNE Visualization plot.”

4.2 WIKI

Same case as for the TED – Dataset: here I marked five spots that appear to be the locations where nodes concentrate most, making sense as clusters.

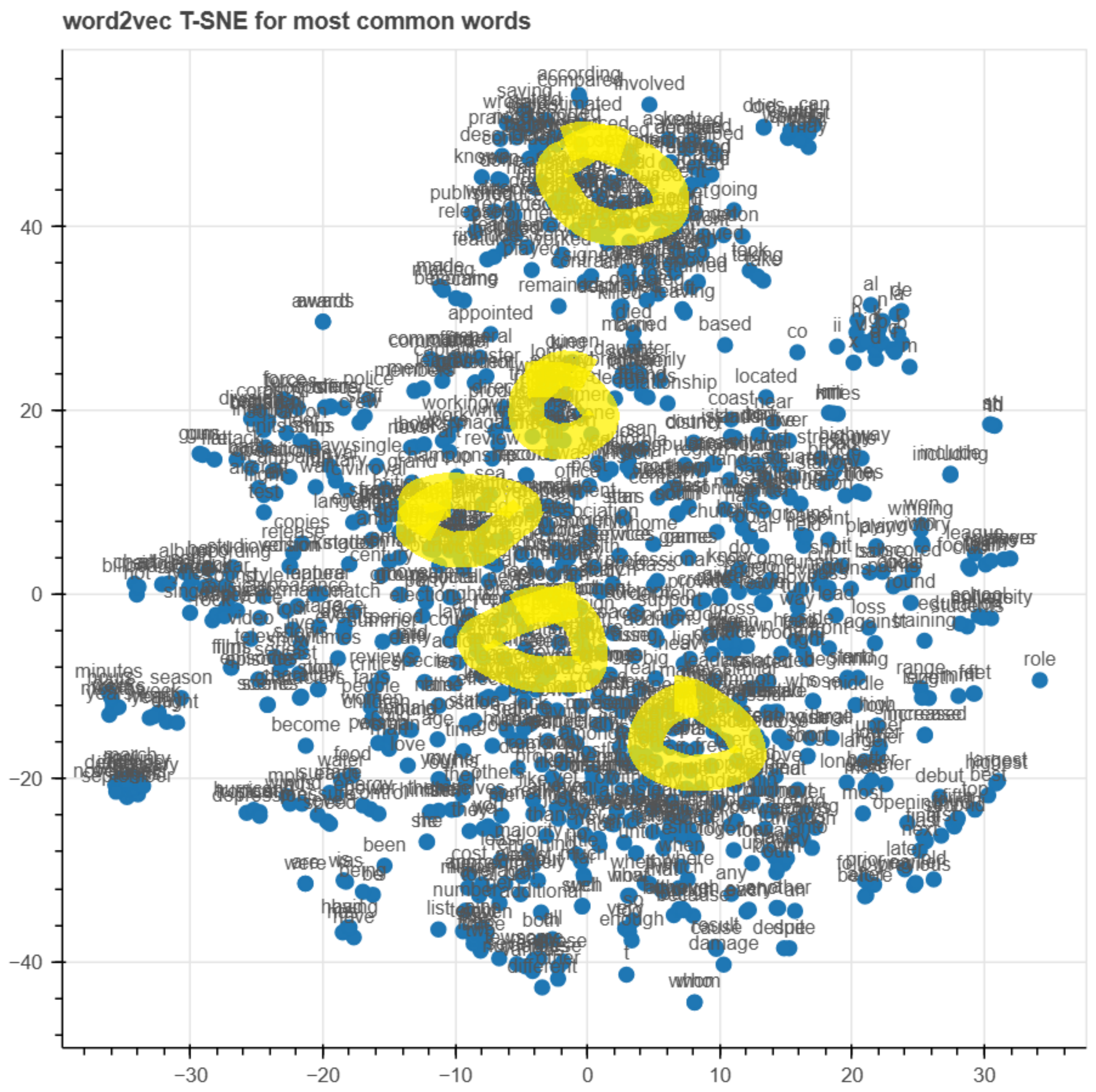

Figure 10: “WIKI – Dataset t-SNE Visualization plot.”

5. Time & Date Words

By looking at both plots from both datasets, we can clearly see that there is definitely two main clusters in them, one for words representing time and another for words representing date so, I do think both datasets differentiate from time and date words, but in my opinion, the Wiki dataset differentiates them more than ted dataset as they appear to be more spread out.

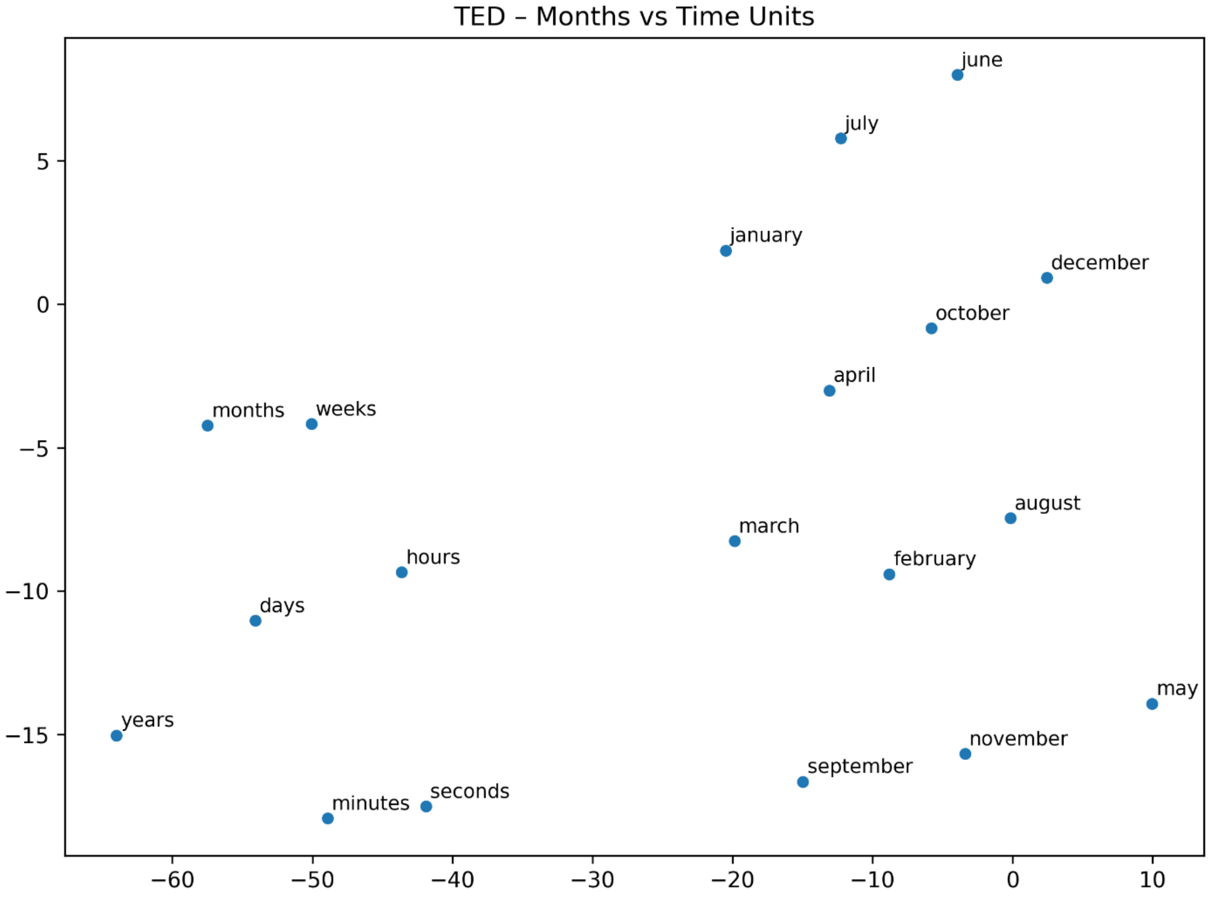

Figure 11: “TED – Dataset Time & Dates Words.”

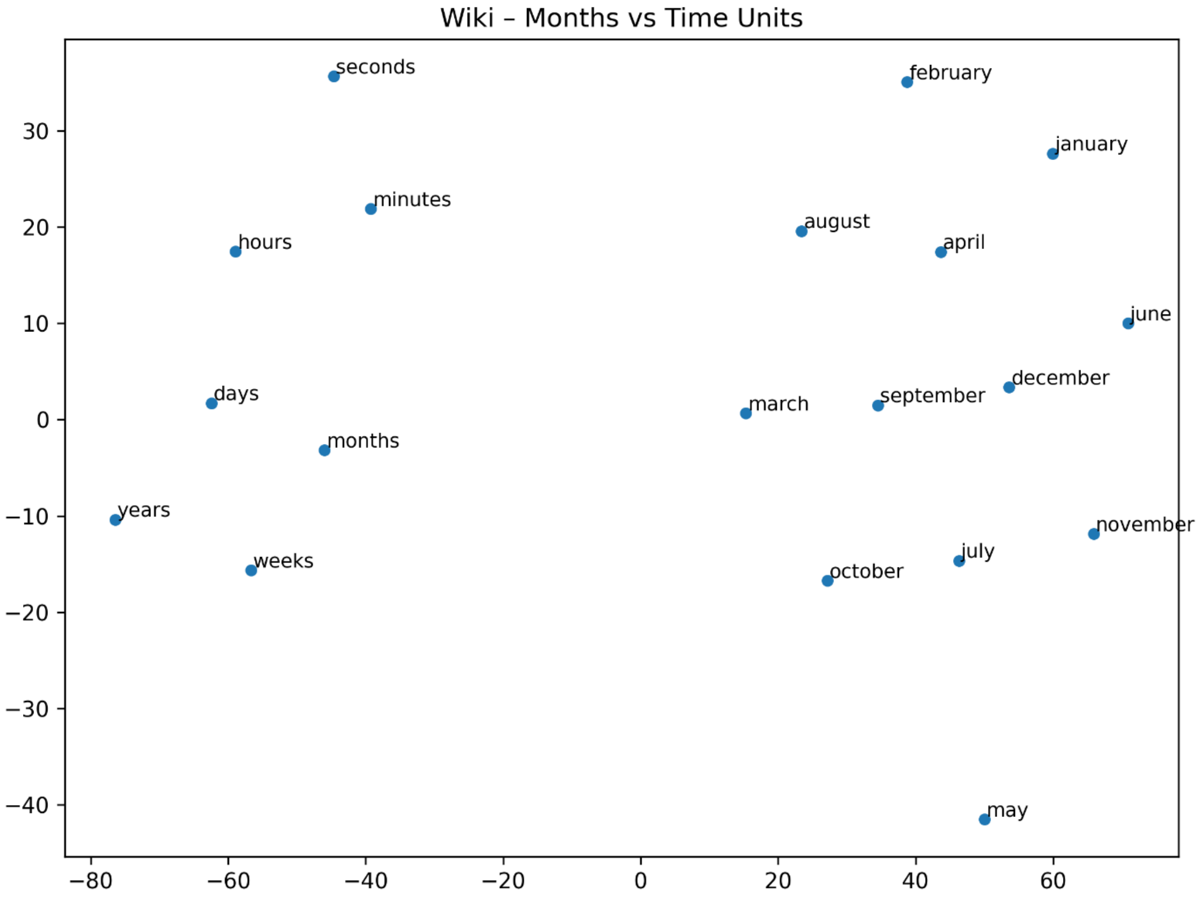

Figure 12: “WIKI – Dataset Time & Dates Words.”