**Welcome to report #4**

**Hala Grefat**

**ID:207921008**

**Part 1:**

In this part we were testing how word2vec calculates similarities between words, and "guessing" words based on models understanding of relation between words, it was fun testing out all sorts of words, and trying to figure out how it's working.

1. Similarity:

   The list of words = [[sea, ocean], [up, down], [night, sundaynight],

   [berry, strawberry], [strong, tough], [alot, many],

   [nothing, naught], [pink, blue], [pin, paper], [barry, bob]]

   My hypothesis for the distances:

   For both models, distances between words that are similar or are used together often (sea, ocean)(up, down) were close to 1, and for words that aren't used together, or have no relation meaning wise (up, down), (night, sundaynight) distance was close to 0, which are the results you'd expect, in addition to that I noticed that for words that aren't Similar But are related (are used in the same context often) such as [pin, paper] have smaller distance.

   In conclusion I think that both models consider context, as well as spelling of the word, the results are similar, I think model 300 is more accurate, it "understands" the meaning of the word better.

2. Most similar:

The list of Quartets:

['hate', 'love'], ['up', 'down']

['eat', 'food'], ['water', 'drink']

['daughter', 'mother'], ['son', 'father']

['books', 'read'], ['write', 'book']

['flower', 'lily'], ['fruit', 'strawberry']

The results were different each time, I didn't see any big difference between the models, many times both models gave same result.

**Part 2:**

For this part and on, I decided to work with model 300, even though both models gave fairly similar results, I noticed that this model gives more information about the word, and "understands" which context the word might be used in.

Results discussion:

Honestly, the results are better than I expected, some of the sentences could have been Biden's tweet.

example: "If we act now on the American job Plan, in 50 years, people will look back and say this was the moment that America won the future."

"I will need the faith that you have placed in me."

Other sentences, didn't make sense, and some were coherent but with incorrect grammar.

My personal favorite: "That's how we'll 3-2 this pandemic."

3-2 = beat.

Questions answers:

1. Not all sentences made sense, words that weren't replaced well:
   - **me** was replaced with **you**. "I'm honored that you have chosen you to lead our great country"
   - **Matters** was replaced with **decisions** "It decisions whether you wash your hands."
   - **Help** was replaced with **give** "…and can give save lives"
   - **Vaccine** was replaced with **measles** "— take the measles when it's your turn and available."
   - **Get** was replaced with '**ll** "Let's 'll to work."
   - **Impeach** was replaced with **acquit**

   And more, the reason why these word's weren't replaced correctly, is because even though words they were replaced with have similar meaning, their meaning is different within the context.

2. Some tweets are coherent for example "We can save 60,000-100,000 lives in the years and months ahead if we step up together. Wear another mask. staying socially distanced. Avoid large outdoor gatherings. Each of us has a duty to do what we can to protecting ourselves, our families, and our fellow Americans."

Others are not "America is so much need than what we're seeing today."

Some tweets\ sentence have different meaning "It matters whether you continue to culturally distance."

3. Most of the tweets that do still make sense, have grammar mistakes, there are a few that don't. "If we act now on the

American job Plan, in 50 years, people will look back and say this was the moment that America won the future."

**Part 3:**

In this part we embedded songs by there average.

For model, I used the 300 model for the same reason explained previously.

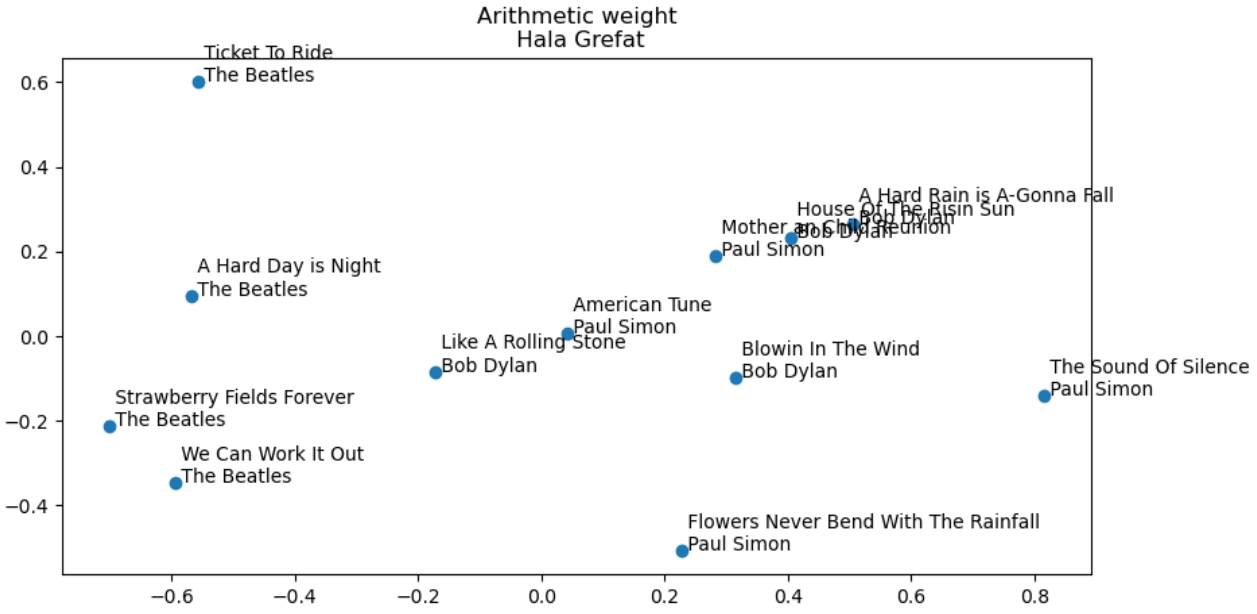
My weight function:

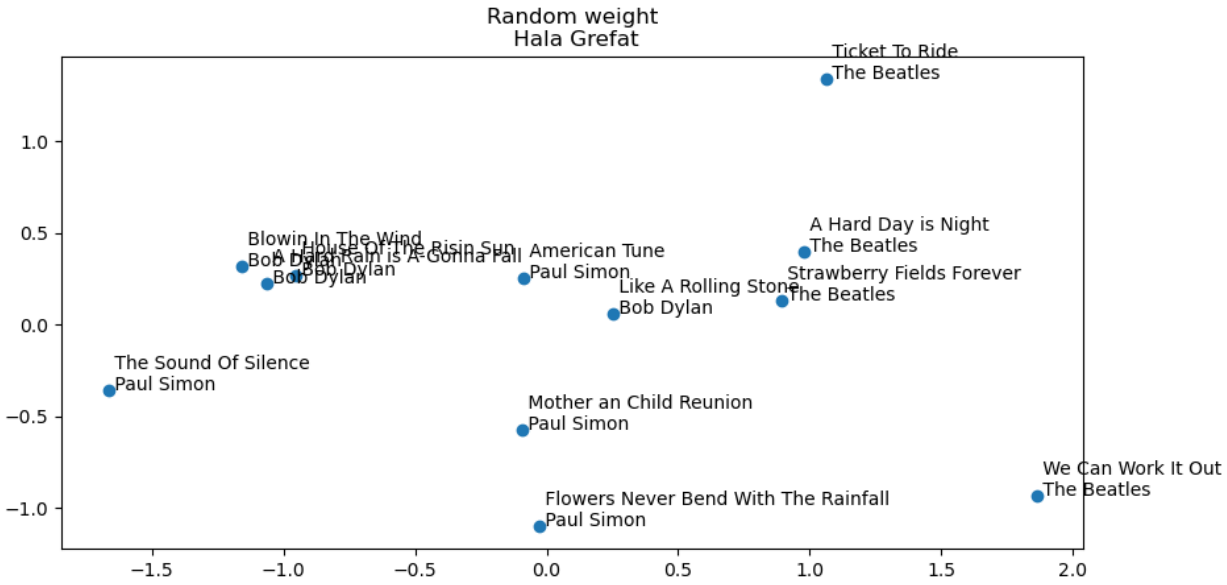
I weight based on count, for each song I  chose a threshold to be the mean of count for all words  ((max_count – min_count)/2)

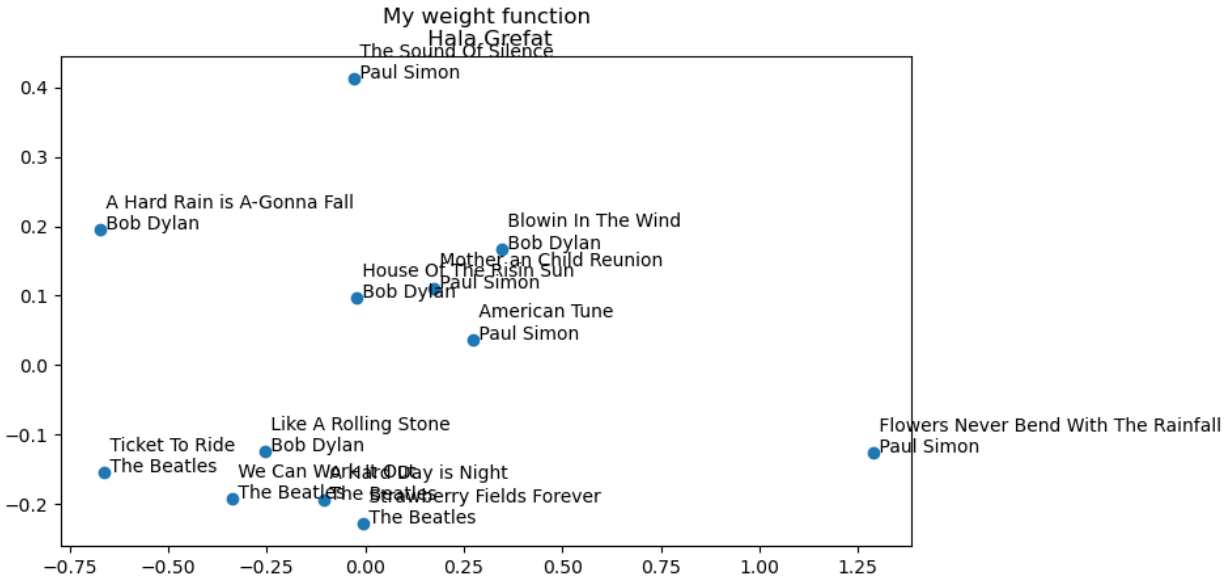
For words that appear in the song for a number higher than mean they got weight 1, and if less they got weight 4.

I chose this weight function so that "unique" words, that are rare get higher weight, because they're rare that means they'd be rare for other songs, which would help distinguish between songs.

Results:






Answers:

1. Non of the weight function gave an obvious separation.

2.* In my weight function there's a posable separation with one mistake (Bob Dylan – Blow in the wind).

* In Random weight function there's a posable separation with two mistakes (Bob Dylan – Like a rolling stone, Paul Simon-The sound of silence).

3. The Beatles were well separated in all graphs, which aligns with fact that they're a rock band who are English.

While Bob Dylan and Paul Simon are both American and were most popular in the same era (the 1960's) which would explain why their songs are similar.

Being from different countries, means using different words, or slangs, the difference would be visible when embedding the song.

4.Seeing that the differences between the Beatles and Bob and Simon were verry visible, as well as the similarities between Bob and Simon, that means that using these models for classification seems like a good idea.

Thank you for reading.

Have a nice day 😊!