Random Word Pairs and Cosine Similarity Values (Using glove.840B.300d)

1. -0.16561156515545453 for '['motorcycles !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!']'
2. -0.14379252973800494 for '['welcome Slemish']'
3. -0.04642816096465379 for '['apt-get Anti-Vaccine']'
4. 0.03212581063798954 for '['9835 Google-backed']'
5. 0.04409694404456912 for '['va. ProjectMy']'
6. 0.07269354136405105 for '['IP-65 Dala']'
7. 0.08442179405873532 for '['cover-all Tittlemouse']'
8. 0.10517627275735762 for '['Bhikaji suddenness']'
9. 0.11817569635425539 for '['nanoelectronics Corima']'
10. 0.1321270814193379 for '['Budich perron']'
11. 0.13620535795925284 for '['T-Boned deadmau5']'
12. 0.13813239826956644 for '['kinfolks for.Below']'
13. 0.14019228686379887 for '['WORLDHOTELS IvCNuB4']'
14. 0.15356087886256647 for '['Testicle Stauffers']'
15. 0.15804493461009256 for '['39,905 ballistix']'
16. 0.19193805463610222 for '['Sonderausgabe LowBest']'
17. 0.21846308278089877 for '['DRAUGHTSMAN 56.04']'
18. 0.22384644894601616 for '['Zaborski Appoigny']'
19. 0.22651017354380015 for '['Oct12 .2610']'
20. 0.24988821679029527 for '['16,842 6217']'
21. 0.2720351147742814 for '['7,574 05-13-06']'
22. 0.3003876957094562 for '['ITALIANI 9,80']'
23. 0.31085565069441234 for '['18795 ETCHINGS']'
24. 0.4519903034784326 for '['11:20:01 11/03/99']'
25. 0.7286901427271523 for '['14:53:14 18:47:18']'

These values are often meaningless as the random words don’t have very solidified relationships with one another. As such, I focused most on the extreme values that I ran across. Something very interesting to me was that the most positive similarity scores were for matching number sets. This makes sense, as one would imagine numbers to live in the same vector space as other numbers. I got two times, and those were the highest similarity score I ran across. The lowest I ever got was the above example with motorcycle and the exclamation points. This may be because that many exclamation points are associated with high excitement and emotion and maybe motorcycles aren’t or could have something to do with punctuation not being associated with vehicles. The closest to zero that I got was '9835 Google-backed' which is fair because I can’t see any connection among those words either.

Cosine Sentence Scores

|  |  |  |
| --- | --- | --- |
| Cosine Similarity Score | Sentence 1 | Sentence 2 |
| 0 | Maybe instead, he should be in a movie about a guy who doesn't know anything about Secretariat. | No more tomfoolery. |
| 0 | If I don't interact with this app every 80 seconds, it charges my credit card five dollars. | Hey, champ |
| 0 | That sound means it's time to pick your category. | Did you see the game last night? |
| 0 | Mm-hm. | I'm sure I'll pick it up. |
| 0 | Are you asking me to throw the game? | Don't worry about it. |
| 0 | Break a leg. | I'm still working my way through the last 30 years of technology. |
| 0 | You have to pay close attention, it's subtle, but when he gets excited, his ears ever so slightly flop up and give him away. | What's going on? |
| 0 | No? | Correct. |
| 0 | Mr. Peanutbutter | Tonight we answer the age-old question: "Hollywood stars and celebrities, what do they know? " |
| 0 | Oh, no thanks, Mia. | Listen to that crowd. |
| 0 | No, it doesn't. | What the? |
| 0 | Uh You want to talk about that, BoJack? | Sorry, no time for idle chit-chat, because we've entered |
| 0 | Are you wearing roller skates? | Someone unplugged the main power line and plugged in a George Foreman Grill. |
| 0 | Daniel Radcliffe! | Uh, what? |
| 0 | Bow-wow-wow-wow. | Well, there are roadside cameras all along the PCH, I have a few friends in the highway patrol, and she told me, of course! We're married! Wait, so you've known all this time? |
| 0 | Did you see the game last night? | Let's really make him sweat. |
| 0 | "To what extent was feudalism a cause of the French Revolution?" | And BoJack barely got beyond his thesis statement. |
| 0 | And that color was green. | I can walk into any Chili's with a child under 12 and get crayons for free. |
| 0 | Well, have fun watching me win that pen. | Which is an exciting surprise. |
| 0 | Also, it has four different colors. | Mia, get ready to drop the bomb. |
| 0 | Or maybe she went to help herself get away from her awful marriage. | Hi, I'm Daniel Radcliffe. |
| 0 | No! | You earned my respect and this pen. |
| 0.09623 | It is great to be here. | And as my own dog, I decided to forgive BoJack live on the air! Bring it in, buddy! |
| 0.10911 | Do you know anything about Secretariat? | Hey! Hey, hey! Well, that stern but supple kiss puts you over the top. |
| 0.20412 | Watch the ears. | What else could the universe possibly owe you? |

The sentences were so unique that only occasionally would words match between randomly selected sentences. The way that sentence calculation works, is that sentences are more similar the more words that they share. This is perhaps a naïve way to measure this, as it takes no contextual meaning or more than just words. The word play and playing are not treated the same, even though a human may look to that as saying the sentences are similar. The only non-zero scores I received were from common words such as ‘it’, ‘to’, ‘the’, ‘you’, etc. These are majority non-meaningful words. To get more from the cosine similarity scores, we could look to use word vectors we already have and do some complex computation there. Or do some stemming of words so that conjugation won’t influence similarity scores.