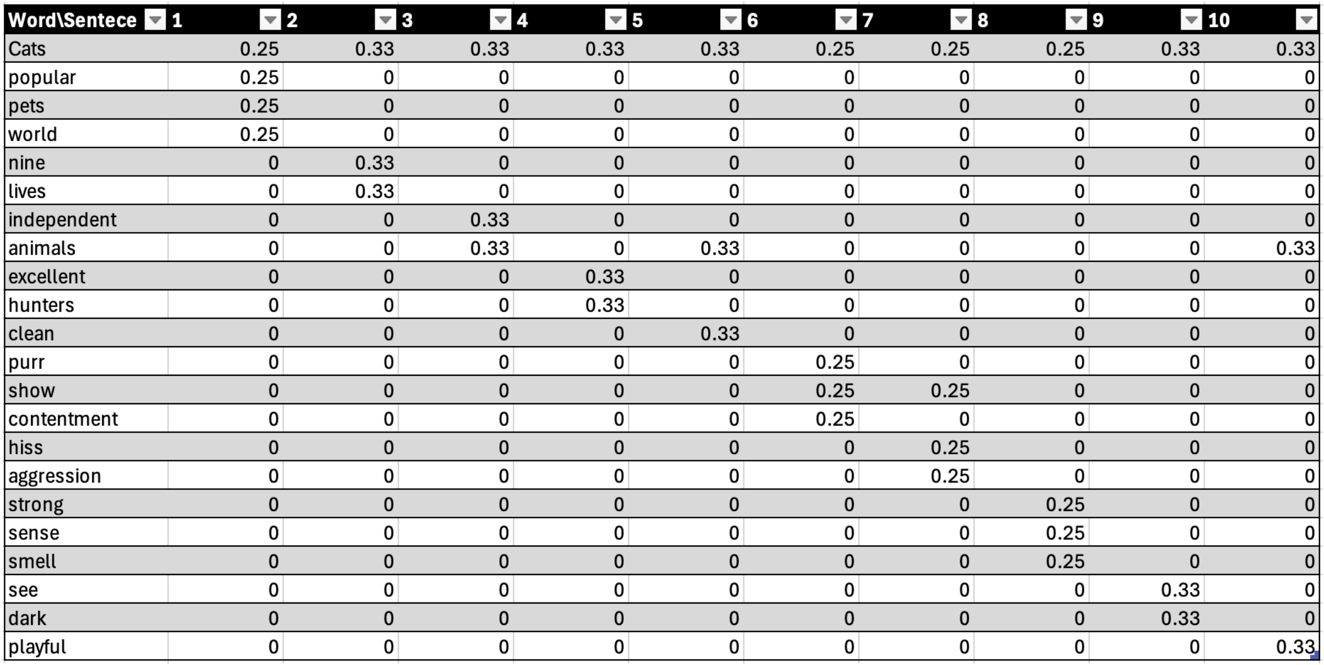
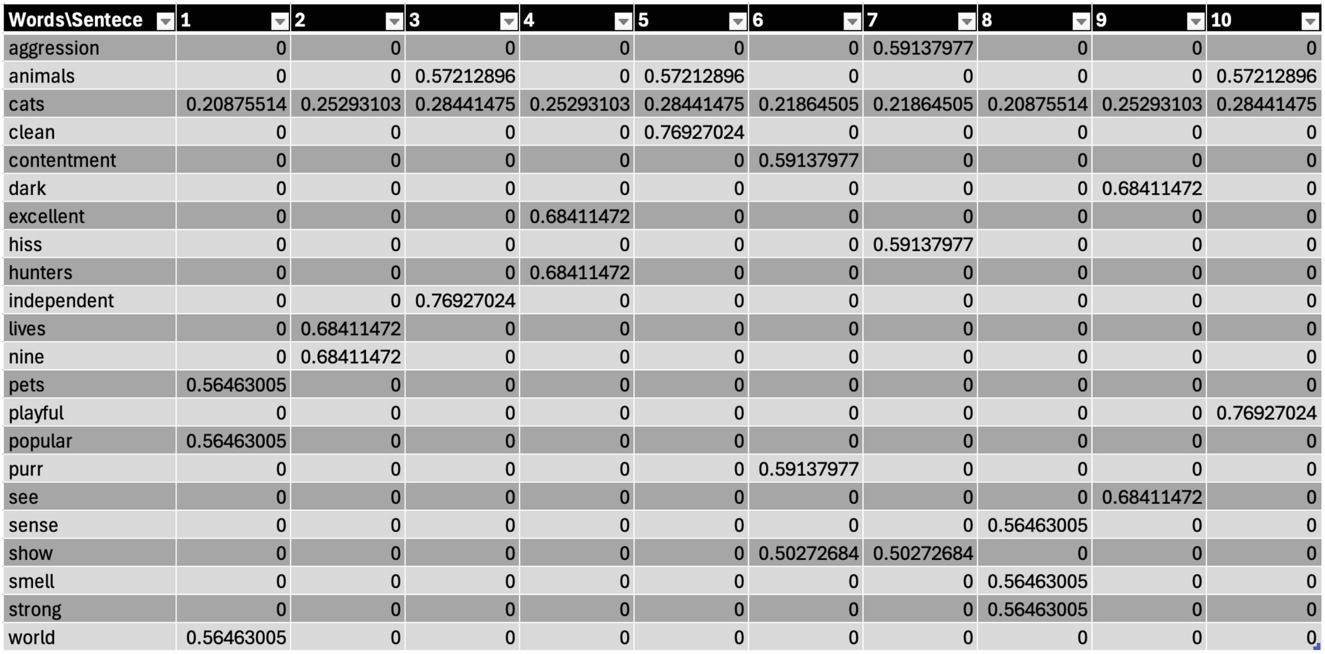
Lucas Sipos

Problem 1:



| Word\Sentece | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Cats | 0.25 | 0.33 | 0.33 | 0.33 | 0.33 | 0.25 | 0.25 | 0.25 | 0.33 | 0.33 |
| popular | 0.25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| pets | 0.25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| world | 0.25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| nine | 0 | 0.33 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| lives | 0 | 0.33 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| independent | 0 | 0 | 0.33 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| animals | 0 | 0 | 0.33 | 0 | 0.33 | 0 | 0 | 0 | 0 | 0.33 |
| excellent | 0 | 0 | 0 | 0.33 | 0 | 0 | 0 | 0 | 0 | 0 |
| hunters | 0 | 0 | 0 | 0.33 | 0 | 0 | 0 | 0 | 0 | 0 |
| clean | 0 | 0 | 0 | 0 | 0.33 | 0 | 0 | 0 | 0 | 0 |
| purr | 0 | 0 | 0 | 0 | 0 | 0.25 | 0 | 0 | 0 | 0 |
| show | 0 | 0 | 0 | 0 | 0 | 0.25 | 0.25 | 0 | 0 | 0 |
| contentment | 0 | 0 | 0 | 0 | 0 | 0.25 | 0 | 0 | 0 | 0 |
| hiss | 0 | 0 | 0 | 0 | 0 | 0 | 0.25 | 0 | 0 | 0 |
| aggression | 0 | 0 | 0 | 0 | 0 | 0 | 0.25 | 0 | 0 | 0 |
| strong | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.25 | 0 | 0 |
| sense | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.25 | 0 | 0 |
| smell | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.25 | 0 | 0 |
| see | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.33 | 0 |
| dark | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.33 | 0 |
| playful | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.33 |



| Words\Sentece | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| aggression | 0 | 0 | 0 | 0 | 0 | 0 | 0.59137977 | 0 | 0 | 0 |
| animals | 0 | 0 | 0.57212896 | 0 | 0.57212896 | 0 | 0 | 0 | 0 | 0.57212896 |
| cats | 0.20875514 | 0.25293103 | 0.28441475 | 0.25293103 | 0.28441475 | 0.21864505 | 0.21864505 | 0.20875514 | 0.25293103 | 0.28441475 |
| clean | 0 | 0 | 0 | 0 | 0.76927024 | 0 | 0 | 0 | 0 | 0 |
| contentment | 0 | 0 | 0 | 0 | 0 | 0.59137977 | 0 | 0 | 0 | 0 |
| dark | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.68411472 | 0 |
| excellent | 0 | 0 | 0 | 0.68411472 | 0 | 0 | 0 | 0 | 0 | 0 |
| hiss | 0 | 0 | 0 | 0 | 0 | 0 | 0.59137977 | 0 | 0 | 0 |
| hunters | 0 | 0 | 0 | 0.68411472 | 0 | 0 | 0 | 0 | 0 | 0 |
| independent | 0 | 0 | 0.76927024 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| lives | 0 | 0.68411472 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| nine | 0 | 0.68411472 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| pets | 0.56463005 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| playful | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.76927024 |
| popular | 0.56463005 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| purr | 0 | 0 | 0 | 0 | 0 | 0.59137977 | 0 | 0 | 0 | 0 |
| see | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.68411472 | 0 |
| sense | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.56463005 | 0 | 0 |
| show | 0 | 0 | 0 | 0 | 0 | 0.50272684 | 0.50272684 | 0 | 0 | 0 |
| smell | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.56463005 | 0 | 0 |
| strong | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.56463005 | 0 | 0 |
| world | 0.56463005 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

d. “Cats” has a high TF score because it calculates the total count of that word over the total words in all the documents, but a much lower TF-IDF score, because the IDF penalizes words that appear across many documents, and they become irrelevant.
Words like “independent”, “playful” and “hunters” get the IDF boost because they appear once in every document, meaning it will carry more meaning for distinguishing between documents.

Problem 2:

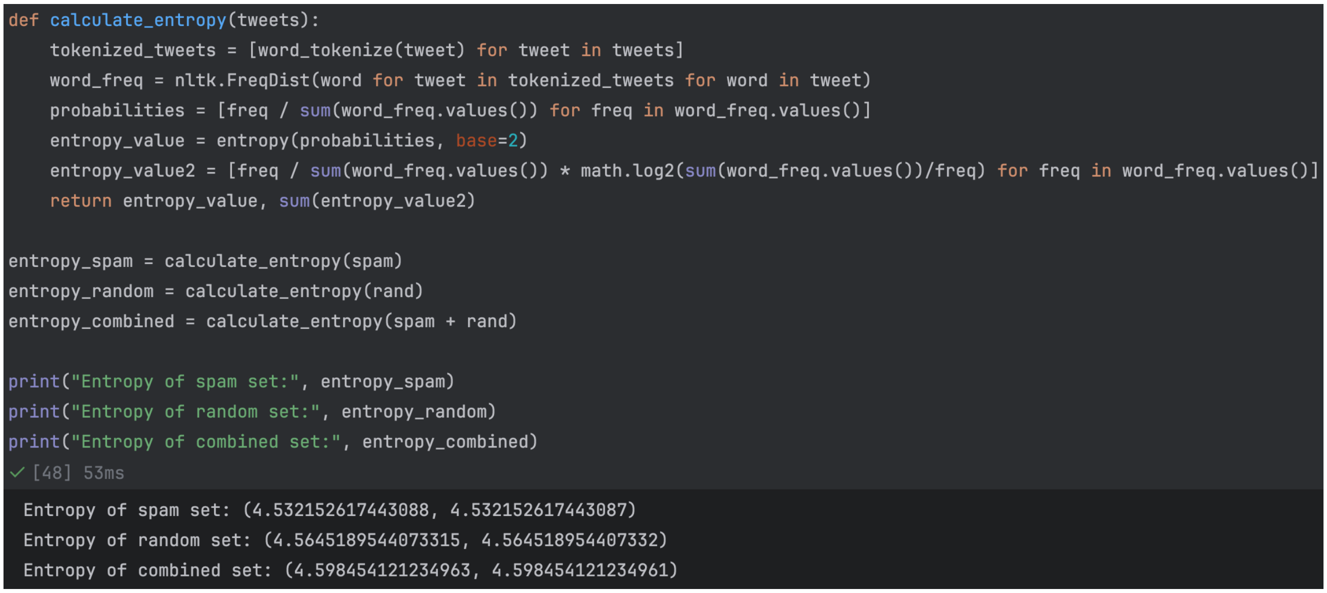
Top 5 PMI Scores:

[(('popular', 'pets'), 3.321928094887362), (('nine', 'lives'), 3.321928094887362), (('excellent', 'hunters'), 3.321928094887362), (('strong', 'sense'), 3.321928094887362), (('show', 'contentment'), 2.321928094887362)]

It makes sense because these values are pretty unique.

Problem 3:

I have built this program and had the following results:



```
def calculate_entropy(tweets):
    tokenized_tweets = [word_tokenize(tweet) for tweet in tweets]
    word_freq = nltk.FreqDist(word for tweet in tokenized_tweets for word in tweet)
    probabilities = [freq / sum(word_freq.values()) for freq in word_freq.values()]
    entropy_value = entropy(probabilities, base=2)
    entropy_value2 = [freq / sum(word_freq.values()) * math.log2(sum(word_freq.values())/freq) for freq in word_freq.values()]
    return entropy_value, sum(entropy_value2)

entropy_spam = calculate_entropy(spam)
entropy_random = calculate_entropy(rand)
entropy_combined = calculate_entropy(spam + rand)

print("Entropy of spam set:", entropy_spam)
print("Entropy of random set:", entropy_random)
print("Entropy of combined set:", entropy_combined)
✓ [48] 53ms
```

```
 Entropy of spam set: (4.532152617443088, 4.532152617443087)
 Entropy of random set: (4.5645189544073315, 4.564518954407332)
 Entropy of combined set: (4.598454121234963, 4.598454121234961)
```

(context: I’m doing ML and I thought I could use this import “from scipy.stats import entropy” to get the entropy and it actually worked) So the first one is done by myself and the second one I did it using Shannon’s Entropy Theory, but basically I’m using the same thing only one uses a function from “scipy” and the second one is done by hand.

Tweets Spam:

Buy our amazing new product now! Limited time offer. #spam #ad
Don't miss out on our incredible deal! Buy now and save big. #spam #sale

Our product is the best on the market. Order yours today! #spam #discount
Act fast! Our limited-time offer ends soon. Buy now and get a free gift. #spam #promo
Our product is guaranteed to satisfy. Buy with confidence! #spam #satisfaction
Amazing deal! Buy our product and get a bonus. #spam #bonus
Don't wait! Our product is selling out fast. Buy now before it's gone. #spam #urgent
Our product is the perfect gift for everyone. Buy now and show you care. #spam #gift
Limited-time offer! Buy our product and get a free shipping. #spam #free
Our product is the best investment you'll ever make. Buy now and start saving. #spam #investment

Tweets Random:

Just had the most amazing pizza! #food #delicious
Feeling so grateful for my friends and family. #love #blessed
Excited to start my new job tomorrow! #work #newbeginnings
Watching the sunset over the ocean. #beautiful #nature
Finally finished reading that book. Highly recommend! #reading #books
Can't wait to travel to Europe next summer. #travel #vacation
Had a great workout today. Feeling strong and energized. #fitness #health
Listening to my favorite playlist. #music #vibes
Just adopted a new puppy! So excited to start training. #dog #puppy
Went for a hike in the mountains. Breathtaking views! #hiking #nature