Compare the statistics of the two different corpora – What are the differences in the most common unigrams between the two? What interesting things do you see about the bigrams?

The two different corpora I used were different movie reviews from different reviewers from the same movie (The Godfather). The first one was from Roger Ebert (https://www.rogerebert.com/reviews/the-godfather-1972) and the second from Peter Bradshaw (https://www.theguardian.com/film/2022/feb/23/the-godfather-review-a-brutal-sweep-of-magnificent-storytelling). Unsurprisingly, both corpora have some of the same unigrams consistently; "the" is number one in both. And (different orders for both corpora) it is followed by "of", "a", "and", "to", and "is". The only word in the top 10 unigrams that would even give a hint to what the two corpora are only appears in Ebert's, and that's the word movie. The bigrams are actually interesting though; ignoring the usual suspects ("to the", etc.) we can see how each reviewer's word choice differs. There are similarities; "of the" and "in the" are present in both, and of course "the godfather" appears, although it is only second in both. "The family", a pretty typical way to refer to the Corleone crime syndicate/family, is only found in Ebert. There's another interesting detail; Ebert appears to deeply prefer the word movie to film. Not only does it appear as a unigram, the bigram "the movie" appears compared to "the film" which appears in Bradshaw's review. Ebert also refers to the titular godfather by his full title "don corleone" whereas Bradshaw refers to him as "the don". Weirdly, in Bradshaw's, the bigram "his daughter" appears. This in reference to the opening wedding scene and to the fact that the don is visited in that scene by a man who wants revenge for his daughter's assault, but it is strange that the phrase is used three times. Even though it is a famous scene, both parts are relatively small parts in the larger story, and for the most part are forgotten by the time the true main plot of the movie begins. That being said, mention of the larger plot in both is lacking, with only one use of "drugs" and that is only found in Bradshaw's review. At least Ebert has an excuse, since he mostly talks about casting choice and cinematography. Bradshaw goes into a sort of synopsis, but he stops before the end of the beginning of the movie without saying why, though he does mention the ending. I feel confident in saying that Ebert seems more familiar with the work, even referring to Mario Puzo (the author of the original novel and writer of the film's screenplay) upwards of 6 times, barely missing the top 10 unigrams (typically referred to by only his last name).

Choose a sentence (fragment) from your first corpus and a sentence (fragment) of the same length from your second corpus.
Using Add-1 Smoothing, compute the probabilities of the 2 different sentences on each corpus. What do you find? Explain the results.

The two sentence fragments I chose were "The Godfather" and "his daughter". I picked "The Godfather" because I knew it occurred in both (just the first one slightly more) and I picked "his daughter" because it appears in the second corpora 3 times but it seemed out of place, and it didn't appear in the first one at all. My results, naturally, reflected this; when performed on the first corpora (Ebert) "The Godfather" showed a 0.00476190476190476% chance of appearing, which seems low, but it is taken from a review with 878 words. 2 out of 878 comes to 0.00227790433, so "The Godfather" appearing at 0.00476190476190476% implies it appears more than once ( and it did, as a top 10 bigram). When performed on the second (Bradshaw) a

similar, if lower, number came up (0.003116883116883117% from a 1000 word piece). The more interesting result is with the phrase “his daughter”. Using the same Add-1 smoothing as the previous phrase “his daughter” only appears at a remarkably low 0.0005952380952380953% in the first corpora. This means it should not appear at all and that it wouldn’t make sense if it did (it of course doesn’t appear at all and this result didn’t surprise me). With the second corpora, it came back at 0.002077922077922078%, a percentage nearly four times higher than the first corpora’s result. This is due to the fact that the phrase is repeated three times in the second corpora.

Raw results below:
Corpora 1 (Ebert)
878
[(('the',), 69), (('of',), 29), (('a',), 28), (('and',), 24), (('to',), 18), (('is',), 13), (('in',), 12), (('as',), 12), (('his',), 12), (('movie',), 9)]
Frequency of being in article
0.0785876993166287
0.03302961275626424
0.03189066059225513
0.02733485193621868
0.02050113895216401
0.014806378132118452
0.01366742596810934
0.01366742596810934
0.01366742596810934
0.010250569476082005
[(('of', 'the'), 8), (('the', 'godfather'), 7), (('the', 'movie'), 7), (('don', 'corleone'), 5), (('in', 'the'), 5), (('as', 'a'), 5), (('at', 'the'), 4), (('the', 'family'), 4), (('and', 'a'), 3), (('up', 'to'), 3)]
Frequency of being in article
0.009111617312072893
0.007972665148063782
0.007972665148063782
0.0056947608200455585
0.0056947608200455585
0.0056947608200455585
0.004555808656036446
0.004555808656036446
0.003416856492027335
0.003416856492027335
Appearances of sentence fragments
0.004761904761904762
0.0005952380952380953

Corpora 2 (Bradshaw)
1000
[(('the',), 68), (('a',), 29), (('is',), 29), (('of',), 25), (('to',), 24), (('and',), 23), (('in',), 23), (('his',), 17), (('as',), 12), (('with',), 11)]
Frequency of being in article
0.068
0.029
0.029
0.025
0.024
0.023
0.023
0.017
0.012
0.011
[(('in', 'the'), 10), (('the', 'godfather'), 5), (('of', 'the'), 5), (('the', 'don'), 5), (('to', 'the'), 4), (('is', 'the'), 4), (('the', 'film'), 3), (('his', 'daughter'), 3), (('it', 'is'), 3), (('is', 'to'), 3)]
Frequency of being in article
0.01
0.005
0.005
0.005
0.004
0.004
0.003
0.003
0.003
0.003
Appearances of sentence fragments
0.003116883116883117
0.002077922077922078