Lizzy Axworthy
11/1/2021
IST 664, HW1

I will be comparing the first and last books of the Dune original series; Dune (1965) by Frank Herbert and Sandworms of Dune(2007) by Brian Herbert and Kevin J Anderson. Frank Herbert died before the series was finished and the following two books were written by his son Brian and fellow science fiction author Kevin Anderson. The posthumous books were written based on Frank Herbert's notes. Many fans were dissatisfied by B. Herbert and Anderson's work, which poses the question of what major differences are there between the writings of Frank and his successors?

To get the actual text needed for the analysis I obtained a pdf with recognized text of each book. For Dune I was able to run a python script to scrape the text using PyPDF2. Sandworms of Dune was formatted in a way which made automated scraping with the earlier technique impossible and so after a time/benefit analysis I copied and pasted it manually rather than developing a new script. Sandworms, hereafter referred to as worms, required one more step of replacing the ’ used with the standard ' Both books were trimmed of everything except the central story, meaning for example that the appendixes are not included. Pre-chapter headings were saved however.

I wanted to keep abbreviations like “can’t” in their original forms for now, so I used regex to tokenize the texts rather than a built in nltk tokenizer. This also let me keep ‘...’ and ‘--’, two odd punctuations that Frank Herbert was very fond of using, and which I wanted to keep in the unfiltered tokens as they were a big part of his writing style. I also switched all words to lowercase, since there was no reason to keep capitalized versions as separate tokens. Lastly, I made Text objects for each corpus. They didn’t come into play here, but I think they will be useful in future analysis.

I made frequency distributions for the raw tokens, alphabetical only tokens, and tokens with stopwords removed. These can be seen in the Jupyter notebook. For stop words I used the basic nltk set with the addition of [would, could, like, us] as they were high frequency in both works, but seem to be in the same class as other stopwords. After normalizing both final frequencies, this was the distribution I obtained:

	dune	worm
0	(said, 0.01024607870334578)	(one, 0.002885036762915713)
1	(paul, 0.007089940833784319)	(face, 0.002854603885669767)
2	(jessica, 0.0038655867278490513)	(duncan, 0.0026233140186005745)
3	(one, 0.0028150566876924267)	(sheeana, 0.0021242148317670547)
4	(thought, 0.0028059611895092525)	(even, 0.0020085698982324584)
5	(baron, 0.002487618753098154)	(said, 0.001795539757510834)
6	(duke, 0.0022465880512440367)	(back, 0.0017590203048156985)
7	(man, 0.0020101050984815067)	(old, 0.001704241125772995)
8	(fremen, 0.0018873158730086544)	(murbella, 0.0016859813994254272)
9	(asked, 0.0016963104111619954)	(time, 0.0015520767395432633)
10	(must, 0.001682667163887234)	(ship, 0.0014546915323562352)
11	(stilgar, 0.0016690239166124727)	(paul, 0.0014181720796610994)
12	(back, 0.0016644761675208856)	(machines, 0.0014181720796610994)
13	(hawat, 0.0016553806693377113)	(new, 0.0013694794760675854)
14	(know, 0.00164173742206295)	(long, 0.0012903539952281250)
15	(mother, 0.0016235464256966015)	(thinking, 0.0012781808443297464)
16	(see, 0.0015917121820554917)	(mother, 0.0012538345425329891)

17	(eyes, 0.0015416869420480333)	(man, 0.0012173150898378536)
18	(saw, 0.0015234959456816848)	(ghola, 0.0011381896089983931)
19	(way, 0.0015144004474985106)	(bene, 0.001132103033549204)
20	(gurney, 0.001491661702040575)	(baron, 0.0011199298826508252)
21	(hand, 0.0014825662038574008)	(leto, 0.0011016701563032576)
22	(water, 0.0014370887129415296)	(see, 0.0011016701563032576)
23	(kynes, 0.0014370887129415296)	(erasmus, 0.0011016701563032576)
24	(voice, 0.0014325409638499425)	(many, 0.001089497005404879)
25	(time, 0.0013552292292929615)	(still, 0.001065150703608122)
26	(men, 0.0012870129929191546)	(teg, 0.0010590641281589327)
27	(sand, 0.0012642742474612192)	(yueh, 0.001022544675463797)
28	(looked, 0.0012324400038201092)	(enemy, 0.0010164581000146077)
29	(around, 0.0011915102619958252)	(though, 0.0010103715245654186)
30	(face, 0.0011778670147210638)	(know, 0.0010103715245654186)
31	(felt, 0.0011596760183547154)	(much, 0.0009921117982178507)
32	(arrakis, 0.001150580520171541)	(eyes, 0.0009860252227686613)
33	(old, 0.0011414850219883669)	(omnius, 0.0009677654964210937)
34	(turned, 0.0011323895238051927)	(dancers, 0.0009677654964210937)
35	(chani, 0.0010596255383397988)	(knew, 0.0009555923455227 15)

36	(feyd, 0.0010414345419734502)	(away, 0.000955592345522715)
37	(people, 0.0010232435456071017)	(machine, 0.000955592345522715)
38	(across, 0.0010096002983323404)	(another, 0.0009495057700735258)
39	(came, 0.0009732183055996434)	(waff, 0.0009434191946243365)
40	(away, 0.0009368363128669465)	(enough, 0.000931246043725958)
41	(place, 0.000918645316500598)	(never, 0.0009251594682767688)
42	(emperor, 0.000914075674090109)	(looked, 0.0009190728928275795)
43	(two, 0.000914075674090109)	(people, 0.0009129863173783902)
44	(thing, 0.000914075674090109)	(two, 0.0009129863173783902)
45	(desert, 0.0009050020692258367)	(around, 0.000906899741929201)
46	(spice, 0.0009050020692258367)	(tleilaxu, 0.0009008131664800117)
47	(think, 0.0009004543201342496)	(memories, 0.0008947265910308225)
48	(left, 0.0008731678255847269)	(first, 0.0008947265910308225)
49	(room, 0.0008686200764931397)	(way, 0.0008886400155816331)

For creating the bi-grams by mutual frequency I again filtered out punctuation and stopwords. Here I found one of the differences between the two works I had hoped to find. 27 of the top 50 bigrams from Dune occur in the form <character> <verb>, like ‘Paul said’. This form doesn’t show up in the worms distribution until the 29th entry (index 28) with ‘Sheeana said, and only has one other instance in ‘Duncan said’ at index 30. The majority of bigrams for worms describe factions and titles, such as “face dancers” and “mother commander” . Part of this could be caused by most of the entities of this category having two-part names that always occur

together, hence whenever “bene” occurs, “gesserit” must follow. There’s also a larger cast of characters and a shorter overall text length in worms, so the instances of any one character doing something would occur less as they have to share narrative space. There are seven shared bigrams: [('bene', 'gesserit'), ('kwisatz', 'haderach'), ('duncan', 'idaho'), ('reverend', 'mother'), ('thufir', 'hawat'), ('old', 'woman'), ('duke', 'leto')]

	dune	worm
0	((paul, said), 0.001337038232926613)	((face, dancers), 0.0009677654964210937)
1	((feyd, rautha), 0.0008185948364856815)	((thinking, machines), 0.0008825534401324438)
2	((bene, gesserit), 0.0006275893746390224)	((face, dancer), 0.0008642937137848761)
3	((reverend, mother), 0.000573016385539977)	((bene, gesserit), 0.0007912548083946049)
4	((baron, said), 0.0005684686364483899)	((kwisatz, haderach), 0.0007912548083946049)
5	((jessica, said), 0.0005684686364483899)	((mother, commander), 0.0007243024784535229)
6	((stilgar, said), 0.0005002524000745831)	((duncan, idaho), 0.0004078005550956809 5)
7	((kynes, said), 0.0004229406655176021)	((old, man), 0.0003651945269513561)
8	((duke, said), 0.0004047496691512536)	((leto, ii), 0.0003469348006037883)
9	((jessica, thought), 0.00040020192005966646)	((bene, gesserits), 0.000340848225154599)
10	((h, h), 0.00037291542551014375)	((honored, matres), 0.0003286750742562205)
11	((old, woman), 0.00030015144004474985)	((paul, atreides), 0.0002982421970102741 4)
12	((hawat, said), 0.0002865081927699885)	((thinking, machine), 0.0002921556215610849)
13	((paul, asked), 0.00026831719640364)	((reverend, mothers), 0.0002799824706627063)
14	((paul, thought), 0.00026831719640364)	((long, ago), 0.00023737644251838145)
15	((ah, h), 0.00025467394912887865)	((tleilaxu, master), 0.00023737644251838145)

16	((gurney, halleck), 0.00022283970548776883)	((reverend, mother), 0.0002130301407216244)
17	((princess, irulan), 0.00022283970548776883)	((axlotl, tanks), 0.00020085698982324584)
18	((gurney, said), 0.00020919645821300748)	((thufir, hawat), 0.00020085698982324584)
19	((chani, said), 0.00020010096002983323)	((old, woman), 0.0001886838389248673)
20	((duke, leto), 0.00018645771275507188)	((liet, kynes), 0.00018259726347567804)
21	((jessica, asked), 0.00016371896729713628)	((young, man), 0.00018259726347567804)
22	((paul, looked), 0.00015917121820554917)	((honored, matre), 0.00017651068802648878)
23	((baron, thought), 0.00015007572002237492)	((god, emperor), 0.0001704241125772995)
24	((jessica, saw), 0.00014552797093078781)	((ghola, children), 0.00015825096167892098)
25	((lady, jessica), 0.00014552797093078781)	((miles, teg), 0.00015825096167892098)
26	((paul, saw), 0.00014552797093078781)	((new, sisterhood), 0.00015825096167892098)
27	((piter, said), 0.00014098022183920068)	((serena, butler), 0.0001521643862297317)
28	((thufir, hawat), 0.00014098022183920068)	((sheeana, said), 0.00014607781078054244)
29	((left, hand), 0.00013643247274761357)	((independent, robot), 0.00013999123533135315)
30	((kwisatz, haderach), 0.00013188472365602646)	((duncan, said), 0.0001339046598821639)
31	((shield, wall), 0.00013188472365602646)	((suk, doctor), 0.0001339046598821639)
32	((baron, asked), 0.00012733697456443932)	((year, old), 0.0001339046598821639)
33	((deep, breath), 0.00012733697456443932)	((house, atreides), 0.00012781808443297464)
34	((halleck, said), 0.00012733697456443932)	((human, race), 0.00012781808443297464)
35	((al, gaib), 0.00012278922547285221)	((machine, fleet), 0.00012781808443297464)

36	((lisan, al), 0.00012278922547285221)	((navigation, bridge), 0.00012781808443297464)
37	((turned, away), 0.00012278922547285221)	((old, empire), 0.00012781808443297464)
38	((fremen, said), 0.00011824147638126509)	((years, ago), 0.00012781808443297464)
39	((harah, said), 0.00011369372728967797)	((de, vries), 0.00012173150898378537)
40	((paul, muad'dib), 0.00011369372728967797)	((enemy, ships), 0.00012173150898378537)
41	((right, hand), 0.00011369372728967797)	((duke, leto), 0.00010955835808540682)
42	((paul, glanced), 0.00010914597819809086)	((shape, shifters), 0.00010955835808540682)
43	((emperor, said), 0.00010459822910650374)	((wellington, yueh), 0.00010955835808540682)
44	((gom, jabbar), 0.00010459822910650374)	((axlotl, tank), 0.00010347178263621756)
45	((man, said), 0.00010459822910650374)	((commander, murbella), 0.00010347178263621756)
46	((duncan, idaho), 0.00010005048001491662)	((final, kwisatz), 0.00010347178263621756)
47	((leto, said), 0.00010005048001491662)	((memories, back), 0.00010347178263621756)
48	((one, side), 0.00010005048001491662)	((butlerian, jihad), 9.738520718702829e-05)
49	((alia, said), 9.550273092332949e-05)	((holtzman, engines), 9.738520718702829e-05)

In creating the mutual information bi-grams, I did the same process as above for filtering. The only bigram that occurred in both lists were (shai, halud) and (de, vries) . I don't see a pattern for the differences between the two lists. Most entries are for names of persons, planets, titles, or objects.

	dune	**worm**
0	((bela, tegeuse), 15.424487814381784)	((van, gogh), 15.004009735976137)
1	((ajax, niner), 15.161453408547988)	((gold, hilted), 14.325937830863499)

2	((diamond, tattoo), 15.161453408547988)	((optic, threads), 13.78161731463969)
3	((delta, ajax), 14.93906098721154)	((shai, hulud), 13.625498112722408)
4	((shaddam, iv), 14.93906098721154)	((nullentropy, capsule), 13.419047235254979)
5	((giedi, prime), 14.746415909269146)	((crystal, sheets), 13.325937830863499)
6	((bi, lal), 14.57649090782683)	((shakkad, station), 13.2555485029721)
7	((lal, kaifa), 14.57649090782683)	((god's, messenger), 13.226402157312585)
8	((e, e), 14.4536341600413)	((chief, fabricator), 13.131559785696648)
9	((gaius, helen), 14.424487814381784)	((crystalline, teeth), 12.892978423587392)
10	((helen, mohiam), 14.424487814381784)	((gurney, halleck), 12.878478853892275)
11	((de, vries), 14.42448781438178)	((shayama, sen), 12.802375874806486)
12	((colonel, bashar), 14.117059289189537)	((de, vries), 12.802375874806485)
13	((dew, collectors), 13.898419002714196)	((butlerian, jihad), 12.740975330142343)
14	((collected, sayings), 13.839525313660626)	((nullentropy, tube), 12.74097533014234)
15	((missionaria, protectiva), 13.839525313660626)	((choam, representative), 12.571050328700032)
16	((shai, hulud), 13.746415909269142)	((surveillance, imagers), 12.518582908805897)
17	((per, cent), 13.653306504877662)	((piter, de), 12.480447779919123)
18	((litter, bearer), 13.498488395825559)	((mathematical, compilers), 12.39047808305821)
19	((salusa, secundus), 13.424487814381783)	((mortal, wound), 12.303570017835046)
20	((inkvine, scar), 13.383845829884436)	((shape, shifter), 12.11648446523455)
21	((hunter, seeker), 13.331378409990302)	((shape, shifters), 12.11648446523455)

22	((gom, jabbar), 13.222853953212132)	((xavier, harkonnen), 12.004009735976137)
23	((puzzled, frown), 13.171507073211911)	((serena, butler), 11.983952083634884)
24	((ducal, signet), 13.102559719494419)	((junction, shipyards), 11.885365239477519)
25	((soo, soo), 12.930498973708115)	((golden, path), 11.847890534058855)
26	((kwisatz, haderach), 12.888434914141571)	((sentinel, robots), 11.834084734533825)
27	((choam, company), 12.613016783851947)	((vladimir, harkonnen), 11.811364658033742)
28	((al, gaib), 12.498488395825559)	((breeding, program), 11.79790239822718)
29	((lisan, al), 12.498488395825559)	((famine, times), 11.653512488892005)
30	((race, consciousness), 12.461013690406896)	((proctor, superior), 11.653512488892003)
31	((suk, school), 12.254562812939469)	((tachyon, net), 11.489436563146379)
32	((instrument, panel), 12.216594962740452)	((navigation, bridge), 11.464407768654855)
33	((palm, lock), 12.051535716469953)	((holtzman, engines), 11.374653115896527)
34	((ring, segment), 11.87605118968574)	((suk, doctors), 11.27065539536231)
35	((nose, plugs), 11.868671659320144)	((medical, center), 11.255548502972102)
36	((landing, field), 11.761522801659353)	((observation, window), 11.230013410864963)
37	((princess, irulan), 11.729948697680374)	((we've, got), 11.16606649408511)
38	((houses, minor), 11.683406111743343)	((suk, doctor), 11.126265486027133)
39	((golden, box), 11.676026581377748)	((edrik's, heighliner), 11.116484465234551)
40	((message, cylinder), 11.630938691849208)	((desperately, needed), 11.085623501529788)
41	((shadout, mapes), 11.624019277909419)	((fifteen, thousand), 11.016082568276714)

42	((communications, equipment), 11.617132892324179)	((star, system), 11.004009735976139)
43	((projectile, weapons), 11.617132892324179)	((navigator, faction), 10.992782480552883)
44	((command, post), 11.598928296946589)	((twelve, year), 10.973421416142715)
45	((transparent, end), 11.59328615036981)	((reverend, mothers), 10.93735893062479)
46	((stillsuit, manufacturer), 11.556591350389128)	((stolen, mines), 10.916546894725796)
47	((stillsuit, manufacturer's), 11.556591350389128)	((honored, matres), 10.899673076161402)
48	((poison, snooper), 11.386666348946816)	((honored, matre), 10.8996730761614)
49	((death, commandos), 11.37137647792222)	((get, rid), 10.840511003693257)