

Text Indexing Project

Natural Language Processing and Information Retrieval

Assumption

- Our hypothesis is: « Depending on features of movies, the word distribution is statistically different. »
- I choose two different genres for studying this assumption : horror movies and family movies.

Horror :

- The Haunting Of Hill House
- The Mummy
- Alien III
- Book Of The Dead
- Evil Dead II
- Friday The 13th
- The Grudge
- Halloween
- Hannibal
- Insidious
- It



Family:

- American Pie
- Bean
- The Brothers Bloom
- Crazy, Stupid, Love
- Happy Feet
- The Incredibles
- Kung Fu Panda
- The Mask
- The Proposal
- Shrek
- Ted



Programming
language

- I choose Python for this project: easier for manipulating files.



How I have proceeded

- Get movies scripts in one file for each genre
- Tokenize the scripts according to blank space and get off of punctuation
- Make a list of stop words and delete them from the scripts
- Count each words and sort them by frequency
- Get the 40 first terms and get a graph in excel for better results
- Analyze results

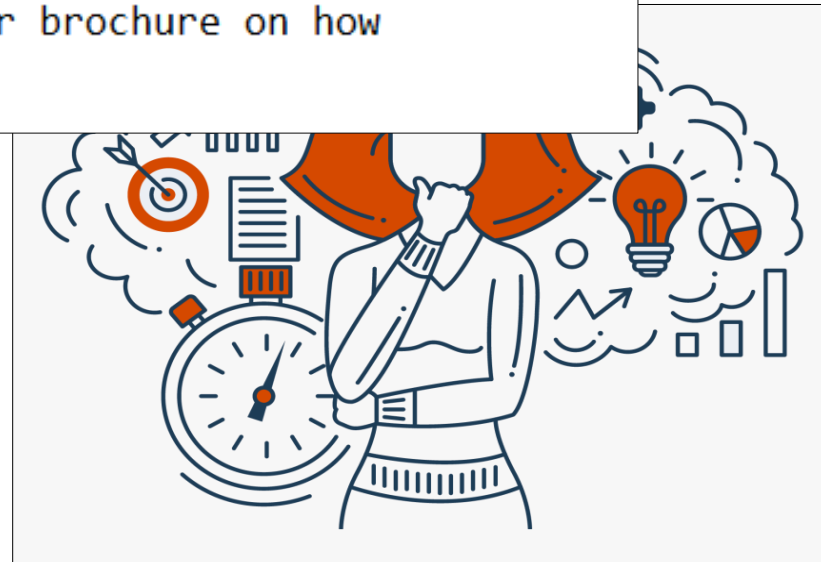


Difficu

VICKY
You think so?

She tears it open. Pulls out a course catalog, various forms, and a letter which she hands to Kevin.

KEVIN
"Dear Ms. Hughes. We're sorry, but after keeping you on the wait list for the past couple months, we've decided you are now rejected. Enclosed is a 100-page, full-color brochure on how rejected you are."



ating

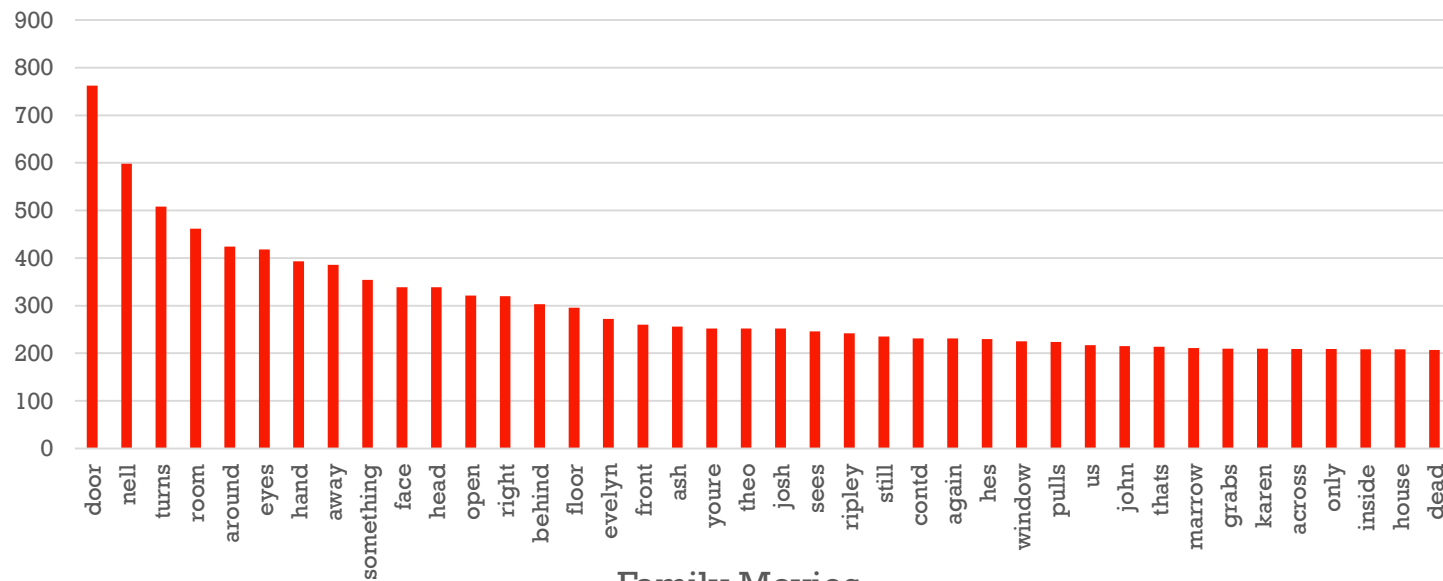
ds list

Results - files

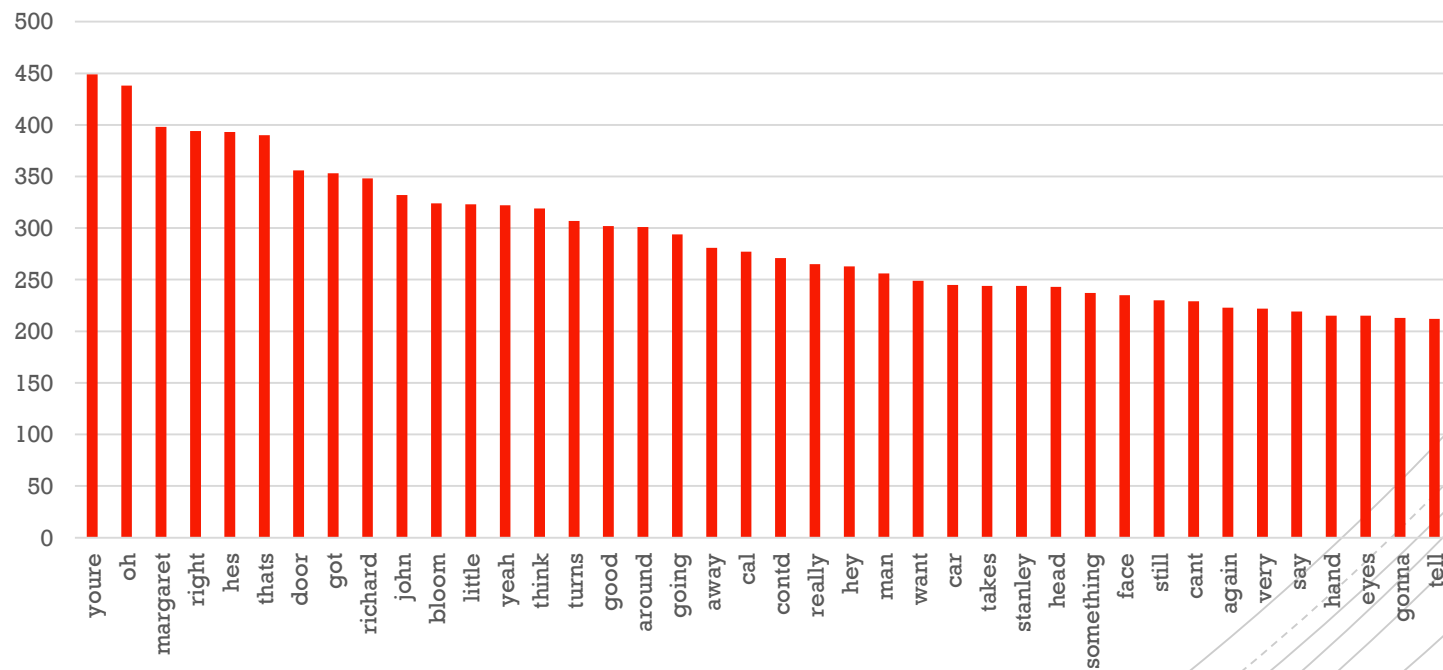
result.txt	result.txt
1 762:door	1 449:youre
2 598:nell	2 438:oh
3 508:turns	3 398:margaret
4 462:room	4 394:right
5 424:around	5 393:hes
6 418:eyes	6 390:thats
7 393:hand	7 356:door
8 386:away	8 353:got
9 354:something	9 348:richard
10 339:face	10 332:john
11 339:head	11 324:bloom
12 321:open	12 323:little
13 320:right	13 322:yeah
14 303:behind	14 319:think
15 296:floor	15 307:turns
16 272:evelyn	16 302:good
17 260:front	17 301:around
18 256:ash	18 294:going
19 252:youre	19 281:away
20 252:theo	20 277:cal
21 252:josh	21 271:contd
22 246:sees	22 265:really
23 242:ripley	23 263:hey
24 235:still	24 256:man
25 231:contd	25 249:want
26 231:again	26 245:car
27 230:hes	27 244:takes
28 225>window	28 244:stanley
29 224:pulls	29 243:head
30 217:us	30 237:something
31 215:john	31 235:face
32 214:thats	32 230:still
33 211:marrow	33 229:cant
34 210:grabs	34 223:again

Result - graph

Horror Movies



Family Movies



Results - Interpretation

Horror Movies

762	door
598	nell
508	turns
462	room
424	around
418	eyes
393	hand
386	away
354	something
339	face
339	head
321	open
320	right
303	behind
296	floor
272	evelyn
260	front
256	ash
252	youre
252	theo
252	josh
246	sees
242	ripley
235	still
231	contd
231	again
230	hes
225	window
224	pulls
217	us
215	john
214	thats
211	marrow
210	grabs
210	karen
209	across
209	only
208	inside
208	house
207	dead

Family Movies

449	youre
438	oh
398	margaret
394	right
393	hes
390	thats
356	door
353	got
348	richard
332	john
324	bloom
323	little
322	yeah
319	think
307	turns
302	good
301	around
294	going
281	away
277	cal
271	contd
265	really
263	hey
256	man
249	want
245	car
244	takes
244	stanley
243	head
237	something
235	face
230	still
229	cant
223	again
222	very
219	say
215	hand
215	eyes
213	gonna
212	tell

Sources

- <https://www.guru99.com/reading-and-writing-files-in-python.html>
- https://www.tutorialspoint.com/python/string_replace.html
- <https://www.programiz.com/python-programming/methods/list/index>
- https://snakify.org/fr/lessons/dictionaries_dicts/
- <https://www.codespeedy.com/sorting-associative-array-in-python/>
- <https://www.programiz.com/python-programming/regex>
- <https://www.rypeapp.com/most-common-english-words/>