Yapay Zeka Ödev 1

Text classifying using Genetic / Hill Climb Algorithm

Bakhtiyar Abbasov – 18011114 **Tural Azimov** – 19011903

Dataset link - https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

In this project, we chose to classify IMDB reviews into two categories: Positive and Negative.

You can find detailed information about these individuals in Tables 1 and 2, such as their elements and their initial state precision. Also, you can find a figure that shows precision changes during the execution of the program.

For Table 1, we created 5 different datasets from our original dataset and ran both the Hill Climb and Simulated Annealing algorithms with the individual that has an N (size) value of 6. In order to create our dataset, we divided our initial dataset into five sub datasets. For each sub dataset, we took 5000 reviews, so our datasets consist of reviews 1-5000, 5001-10000, 10001-15000, 15001-20000, and 20001-25000, respectively.

For the Hill Climb algorithm, the best precision was with the individual that was generated from the dictionary of dataset number 4, with a value of 0.5226. As for the highest number of individuals that were generated via the Hill Climb algorithm in these datasets, it was dataset number 4 with a total of 6 individuals.

When we performed Simulated Annealing in these datasets, the highest precision was for the individual from dataset number 4 with 0.5338. And this value was also the best precision value among other individuals in the datasets, which is the highest precision value that our program achieves while executing. Sometimes these values can be higher than the final precision that our program calculates because of the nature of Simulated Annealing algorithm, which allows for replacing elements of an individual with a specific probability that is calculated by the algorithm itself. The highest number of total individuals was in dataset number 3, with a value of 50.

Dataset	**Hill Climb**			**Simulated Annealing**			
	Precision	PrecisionΔ	Total Indv.	Precision	PrecisionΔ	$\text{Precision}_{\text{best}}$	Total Indv.
1	0.5152	0.0142	3	0.5166	0.0168	0.5224	38
2	0.5118	0.0072	3	0.5042	0.014	0.5112	44
3	0.517	0.0224	4	0.5094	0.0092	0.5184	50
4	0.5226	0.029	6	0.5338	0.356	0.5338	47
5	0.5154	0.0304	5	0.5198	0.24	0.5198	42

Table 1. Precisions for different datasets (N=6)

In Table 2, we showed individuals with five different sizes that were randomly generated from Dataset 1. This dataset was the same as the one in Table 1, with the same name. Used individual sizes are 4, 6, 10, 16, and 20.

The highest precision value for the Hill Climb algorithm was 0.5382, for an individual with a size of 16. And this individual also contains the highest number of unique individuals that were generated, with a total of 8.

When we used different individual sizes in Simulated Annealing algorithm, the highest precision and precision delta were from the individual that had 10 elements, with values of 0.52 and 0.0202, respectively. But the best precision that we got during the execution was with the individual that had 20 elements with a value of 0.5246. As expected, the highest number of total individuals during the execution was the individual with a size of 20.

N	**Hill Climb**			**Simulated Annealing**			
	Precision	PrecisionΔ	Total Indv.	Precision	PrecisionΔ	$\text{Precision}_{\text{best}}$	Total Indv.
4	0.5158	0.02	3	0.507	0.0054	0.5112	37
6	0.5152	0.0142	3	0.5166	0.0168	0.5224	38
10	0.5194	0.0266	4	0.52	0.0202	0.5214	80
16	0.5382	0.0552	8	0.5164	0.0018	0.5206	146
20	0.5212	0.0242	7	0.508	0.0066	0.5246	180

Table 2. Precisions by size of individual (Dataset 1)

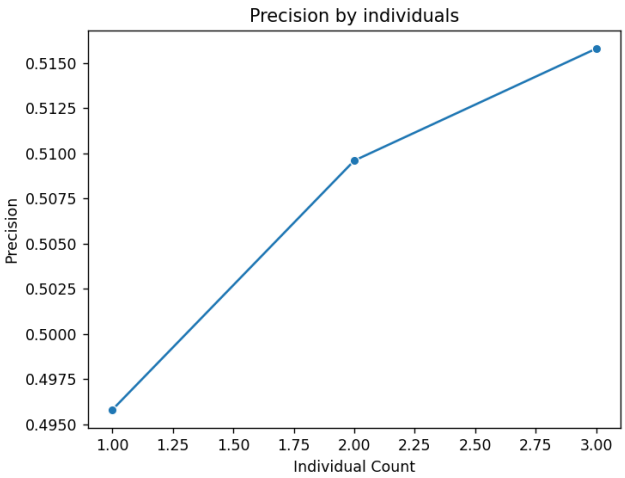
IMDB_1 (1-5000)

Hill climb N = 4

Total Individual Count = 3

First result : ['uncomfortable', 'Visconti', 'food', 'act.'] **Precision = 0.4958**

Last result : ['skills', 'Visconti', 'desire', 'act.'] **Precision = 0.5158**



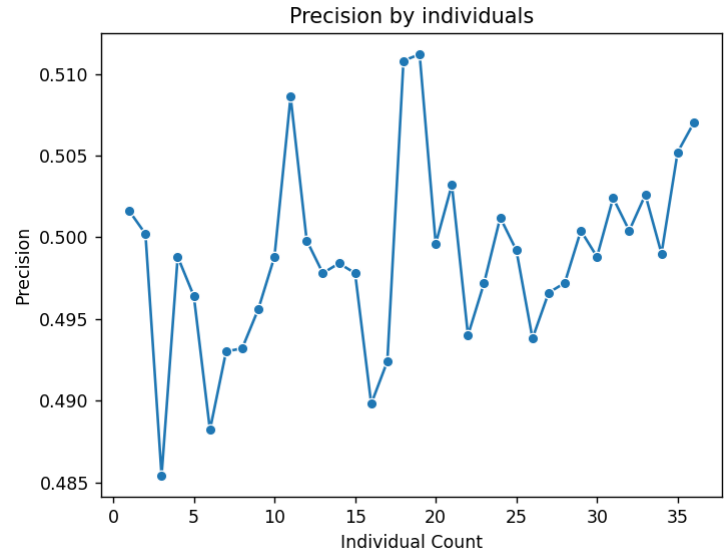
Simulated Annealing N = 4

Total Individual Count = 37

First result : ['screenplay', 'awful,', 'available', '7/10'] **Precision = 0.5016**

Last result : ['thousand', 'laughs,', 'Personally,', '/>3.'] **Precision = 0.507**

Best result : ['thousand', 'laughs,', 'refused', '7/10'] **19th Individual Precision = 0.5112**

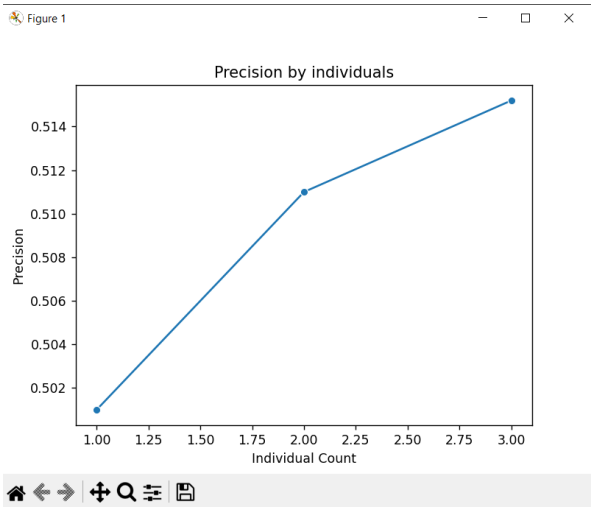


Hill Climb N=6

Total Individual Count = 3

First result : ['/>Now', 'wake', 'starred', 'flimsy', 'excitement', 'blows'] **Precision = 0.501**

Last result : ['/>Now', 'dogs', 'starred', 'peoples', 'excitement', 'blows'] **Precision = 0.5152**



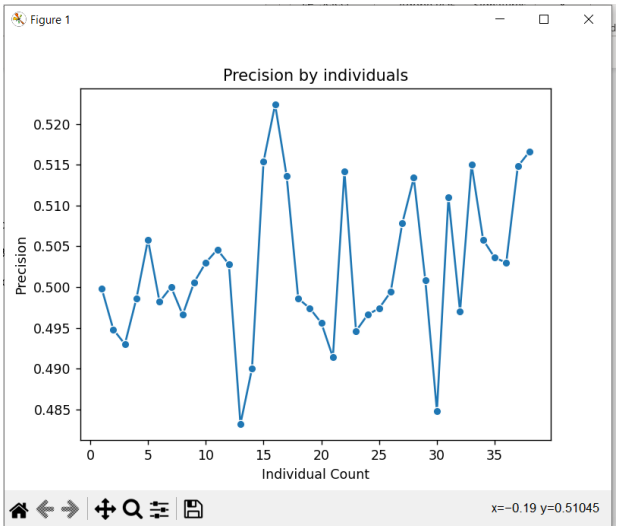
Simulated Annealing N = 6

Total Individual Count = 38

First result : ['tank', 'Craven', 'Yes', 'bucket', 'MY', 'authentic'] **Precision = 0.4998**

Last result : ['mature', 'milk', 'made.<br', 'pay', 'zombies', 'Britain']] **Precision = 0.5166**

Best result : ['mature', 'milk', 'Yes', 'bucket', 'MY', 'authentic'] **16th Individual Precision = 0.5224**

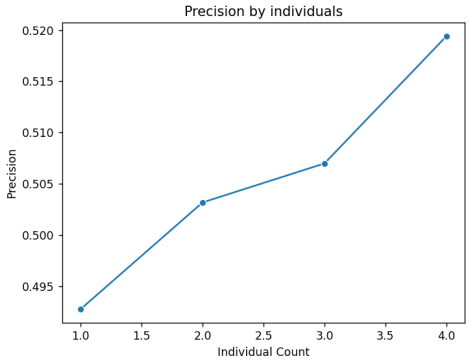


Hill climb N = 10

Total Individual Count = 4

First result : ['food', 'trash.', 'location', '/>Also', 'delivered', 'text', 'promising', 'anything,', 'mystery,', 'Chinese'] **Precision = 0.4928**

Last result : ['food', 'trash.', 'location', '/>Also', 'delivered', 'areas', 'promising', 'anything,', 'mystery,', 'Chinese'] **Precision = 0.5194**



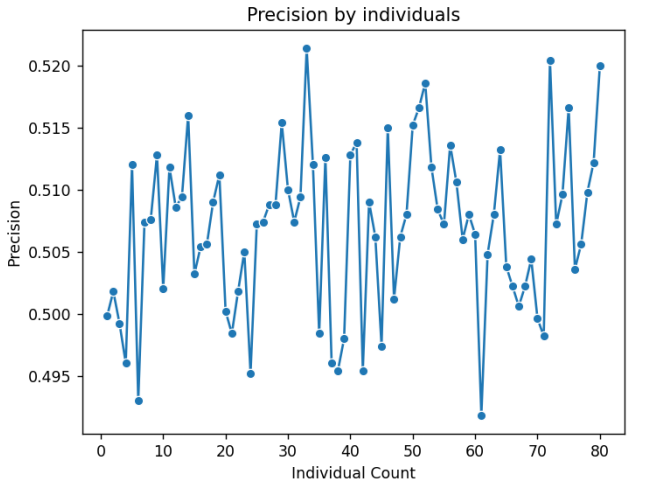
Simulated Annealing

Total Individual Count = 80

First result : ['sharp', 'top,', 'formulaic', 'grew', 'danger', 'middle-class', 'Again,', 'Tiny', 'Four', 'floor']
Precision = 0.4998

Last result : ['first,', 'personal', 'gigantic', 'release,', 'say.', 'throwing', 'wont', 'True', 'flick', '/>Now']
Precision = 0.52

Best result : ['first,', 'personal', 'gigantic', 'release,', 'danger', 'middle-class', 'Again,', 'Tiny', 'Four', 'floor']
33rd individual Precision = 0.5214

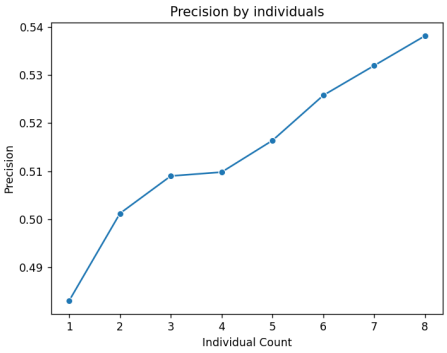


Hill climb N = 16

Total Individual Count = 8

First result : ['Van', 'allowed', 'murdered', 'president', 'humor.', 'core', 'winning', "didn't", 'minutes.', 'Another', 'Fans', 'note,', 'work', 'daily', 'Dee', 'fans,'] **Precision = 0.483**

Last result : ['Van', 'delivers', 'murdered', 'president', 'humor.', 'core', 'winning', 'film,', 'minutes.', 'Another', 'contest', 'ONLY', 'least', 'Elizabeth', 'Dee', 'attempting'] **Precision = 0.5382**



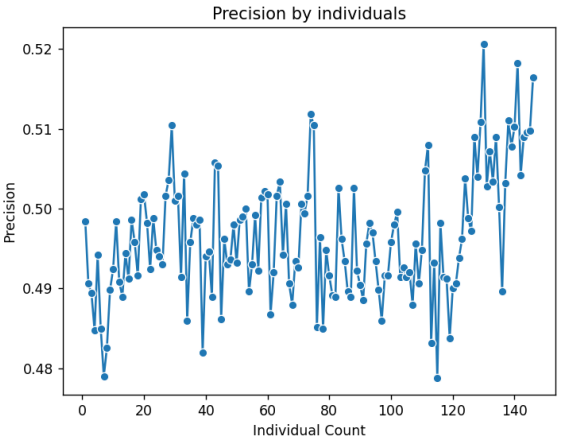
Simulated Annealing N = 16

Total Individual Count = 146

First result : ['purely', 'Howard', 'Michele', 'imaginative', 'designed', 'Ryan,', 'roll.', 'Art', 'top', 'Chase', 'ability', 'actress', 'avoiding', 'performance', 'student', 'survive'] **Precision = 0.4984**

Last result ['during', 'hour,', 'past,', 'Furthermore,', 'entertaining.', 'essential', 'Philip', '`The', 'Hamlet', 'surprising', 'Homer', 'right?', '/>Great', 'mental', 'seems', 'asked'] **Precision = 0.5164**

Best result : ['during', 'hour,', 'past,', 'Furthermore,', 'entertaining.', 'essential', 'Philip', '`The', 'Hamlet', 'surprising', 'Homer', 'right?', '/>Great', 'mental', 'student', 'survive'] **130th Individual Precision = 0.5206**

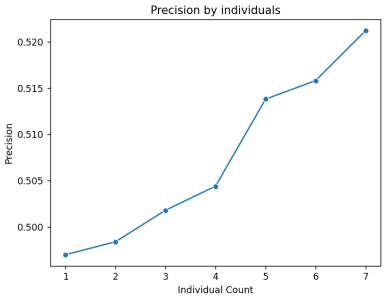


Hill climb N = 20

Total Individual Count = 7

First result : ['(no', 'discussing', 'checking', 'rock', 'transfer', 'dead', 'spiritual', 'pal', 'Last', '....', 'Never', 'Polish', 'Rambo', 'femme', 'destroyed', 'disabled', 'hero', 'anyway', 'Franco', 'Hell']
Precision = 0.497

Last result : ['masterpiece.', 'painful.', 'crush', 'rock', 'transfer', 'relationship', 'sitcom', 'pal', 'Last', 'happiness', 'Never', 'Polish', 'Rambo', 'femme', 'destroyed', 'disabled', 'hero', 'anyway', 'Franco', 'Hell']
Precision = 0.5212



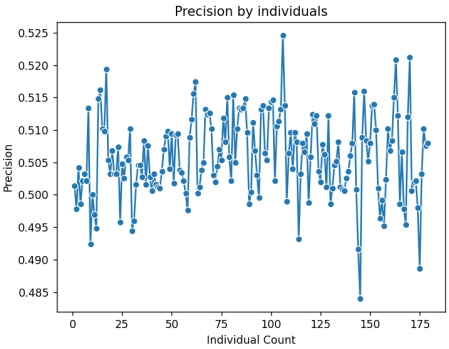
Simulated Annealing N = 16

Total Individual Count = 180

First result : ['cute,', '(that', 'due', 'get.', 'response', 'treatment', 'cinematography,', 'flash', 'era,', 'dialog', 'uninspiring', '(just', 'Benny', 'cinematic', 'presumably', 'play.', 'divorce', 'awaiting', 'reviews.', 'And']
Precision = 0.5014

Last result : ['adolescent', 'past.', '/>By', 'overall,', 'Native', 'Roberts', 'Eva', 'yourself,', 'accident', 'fighting', 'again!', 'boasts', 'names', 'reaching', 'formed', 'documentary', 'dream.', 'sound', 'disjointed', 'crude']
Precision = 0.508

Best result : ['adolescent', 'past.', '/>By', 'overall,', 'Native', 'Roberts', 'Eva', 'yourself,', 'accident', 'fighting', 'again!', 'boasts', 'films.<br', 'cinematic', 'presumably', 'play.', 'divorce', 'awaiting', 'reviews.', 'And']
106th Individual Precision = 0.5246



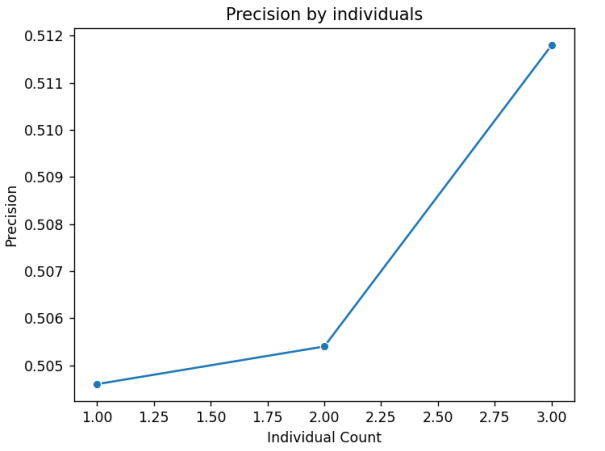
IMDB_2 (5001-10000)

Hill climb N = 6

Total Individual Count = 3

First result : ['happening,', 'ones.', 'case,', 'assist', 'up.', 'biting'] **Precision = 0.5046**

Last result : ['happening,', 'ones.', 'reminds', 'assist', 'battle', 'biting'] **Precision = 0.5118**



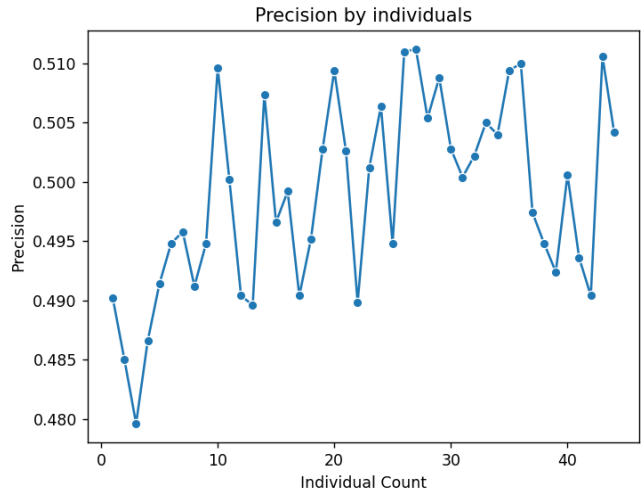
Simulated Annealing N = 6

Total Individual Count = 44

First result : ['interests', 'it,', 'hanging', '&', 'Fonda', 'That'] **Precision = 0.4902**

Last result : ['mission', 'watch,', 'faster', 'alive.', 'wisdom', 'decade'] **Precision = 0.5042**

Best result : ['mission', 'watch,', 'faster', 'alive.', 'Fonda', 'That'] **27th Individual Precision = 0.5112**



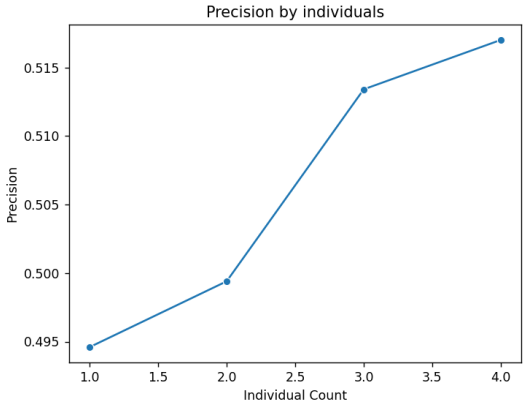
IMDB_3 (10001-15000)

Hill climb N = 6

Total Individual Count = 4

First result : ['(in', 'destroys', 'mesmerizing', 'screenplay', 'lucky', 'tapes'] **Precision = 0.4946**

Last result : ['creepy', 'popping', 'mesmerizing', 'screenplay', 'remotely', 'tapes'] **Precision = 0.517**



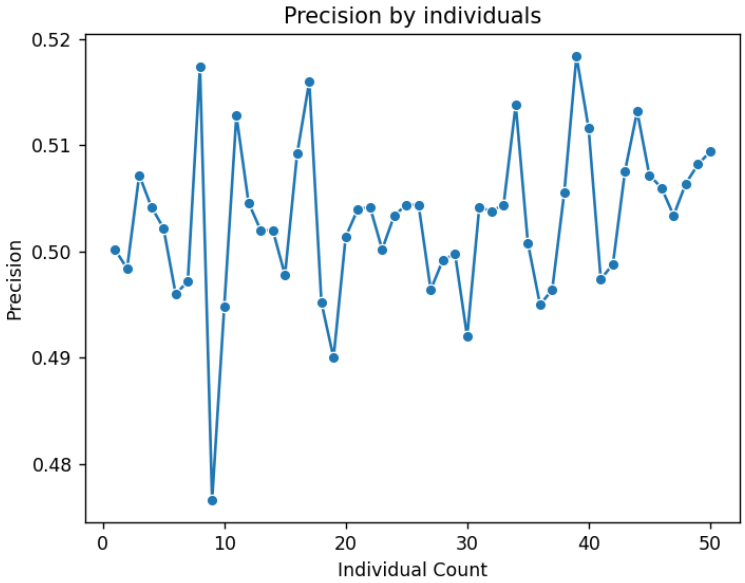
Simulated Annealing N = 6

Total Individual Count = 50

First result : ['awe', 'truth,', 'nominated', 'lesbian', 'praise', 'magazine'] **Precision = 0.5002**

Last result : ['Mormon', 'pictures', 'Stewart', 'cliche', 'Watching', 'black,'] **Precision = 0.5094**

Best result : ['Mormon', 'pictures', 'Stewart', 'cliche', 'school,', 'magazine']
39th Individual Precision = 0.5184



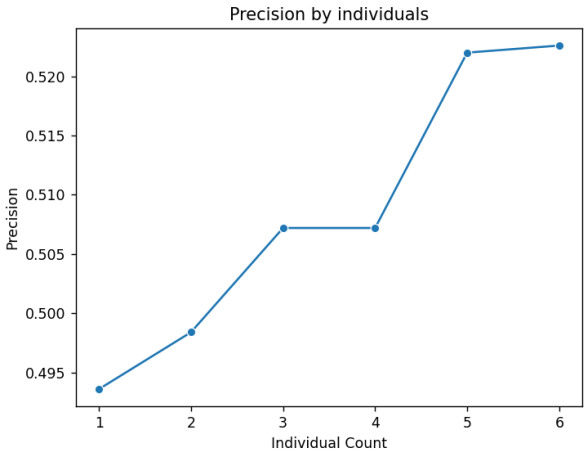
IMDB_4 (15001-20000)

Hill climb N = 6

Total Individual Count = 6

First result : ['forget', 'pale', 'Lou', 'came', 'happy.', 'downbeat'] **Precision = 0.4936**

Last result : ['unique', 'sells', 'Lou', 'So', 'happy.', 'downbeat'] **Precision = 0.5226**



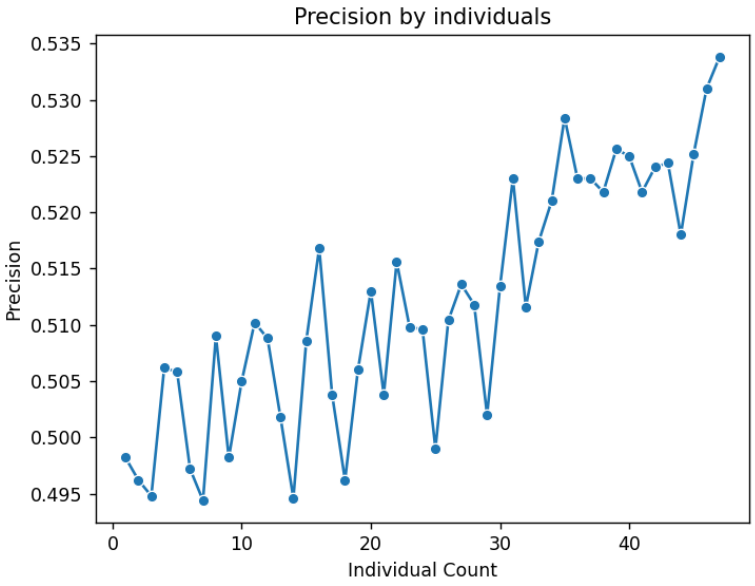
Simulated Annealing N = 6

Total Individual Count = 47

First result : ['joke.', 'among', '3)', 'costs', 'cast.<br', 'angles,'] **Precision = 0.4982**

Last result : ['choose', 'impressed', 'two', 'whole', 'skip', 'FOR'] **Precision = 0.5338**

For this initial individual, the last result was also the best result .



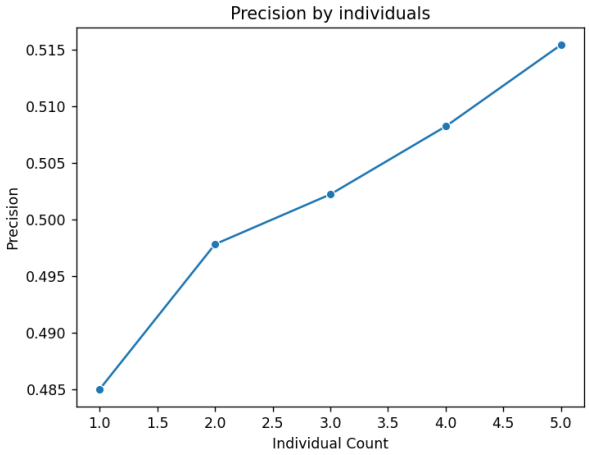
IMDB_5 (20001-24999)

Hill climb N = 6

Total Individual Count = 5

First result : ['reveal', 'least', 'minimal', 'much.<br', 'Slater', 'rides'] **Precision = 0.485**

Last result : ['shallow', 'new', 'favorites', 'much.<br', 'Ross', 'rides'] **Precision = 0.5154**



Simulated Annealing N = 6

Total Individual Count = 42

First result : ['originality', 'composition', 'UK', 'film-making', 'Vincent', 'are.'] **Precision = 0.4958**

Last result : ['undoubtedly', 'war', 'kid.', 'Hell,', 'laughing', 'affair,'] **Precision = 0.5198**

For this initial individual, the last result was also the best result .