Harry Potter Pages Classification Using a Naïve Bayes Algorithm

The Algorithm

**Background**

The Naïve Bayes Algorithm uses a list of probabilities and Bayes Rule to classify data into groups.

The idea of this project was to classify sets of pages from the Harry Potter books into the books in which they belong.

After classifying the Harry Potter books, we then try to use the Benchmark Test (using seven books from different genres) to asses how well our Naive Bayes’ algorithm performs.

**Data Processing**

We start off by splitting all the books into page sets (we will explain later how we optimized the `pageSetSize` which is the number of pages in a set). We then took around 20% of these from each book and set them aside for validation data which we used for training our hyperparameters, we set a further 20% aside for testing and the remaining 60% we used for training data.

We removed all punctuation from the data, except for question marks and exclamation marks which were separated into their own words as we felt this may give more insight as J.K. Rowling may have asked more questions or written more exclamations in some books than others.

**Procedure**

We then took our training data and we calculated the probability for each word appearing in any one page set from each book. We did this by counting the number of page sets a word appears in in each book and dividing it by the total number of page sets in that book. This gave us our probability tables, each of which contain probabilities for around 20 000 words. Each of the tables contains a different number of words because of how the data is split up and which data the trainer sees and which is used for testing (see folder probTables where we have stored each of our tables).

In order to narrow down our dictionary (to improve both speed and accuracy) we decided to remove common words that don’t differentiate well between books. We did this by comparing the probabilities for each word of it appearing in each book and if at least one of these seven probabilities where greater than `upperThreshold` and at least `requiredNum` probabilities are less than `lowerThreshold` we then kept the word. By doing this we only keep words that appear often in one or more books and barely appear in multiple other books.

Using the validation data we optimized the hyperparameters: `pageSetSize`, `upperThreshold`, `lowerThreshold` and `requiredNum`. We did this using the following algorithm:

for each pageSetSize ranging from 1 – 6

probailityTable = create probability table based off pageSetSize using trainingData

for each requiredNum in range 1 - 5

currTable = probailityTable with words removed based off upperThreshold, lowerThreshold and requiredNum

while accuracy on validation data is improving

while accuracy on validation data is improving

Decrease lowerThreshold by alpha

currTable = probailityTable with words removed based off new upperThreshold, lowerThreshold and requiredNum

Retest on validation data

while accuracy on validation data is improving

Increase upperThreshold by alpha

currTable = probailityTable with words removed based off new upperThreshold, lowerThreshold and requiredNum

retest on validation data

Change alpha by a factor of a half

If alpha is less than a threshold exit the loop

take the pageSetSize, upperThreshold, lowerThreshold and requiredNum that yield the best accuracy

test on testing data

**Results**

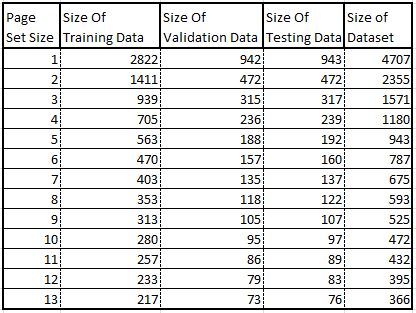
The following graph shows the accuracy given for the different pageSetSizes on our test data, it should be noted that since there are 7 options a random guess will be correct 14.29% of the time.



It is quite clear a pageSetSize of 5 gives the highest accuracy. Since the pageSetSize of 5 gives the highest accuracy, we use this pageSetSize in our analysis.

, we however feel that the pageSetSize of 3 pages with an accuracy of 50% is the best result as it balances as few pages as possible with the best accuracy possible. In addition despite a pageSetSize of 1 only having a 24% accuracy this is nearly double random and we feel it is still noteworthy and meaningful that based off a single page we can classify with that accuracy.

It should also be noted that because we have a limited number of pages (there are 4707 pages in the series) when we increase the pageSetSize it decreases our dataset size. This means that the higher pageSetSizes have less training data and less testing data so the results aren’t as trustworthy. We believe this is the reason that the percentage accuracy doesn’t change as drastically over the last few pageSetSizes. The following table gives the breakup of pageSetSize and dataset size.



Thus the best classifier that we found was trained using the following parameters:

pageSetSize = 5

Training pageSets = 563

Testing pageSets = 192

Validation pageSets = 188

Words Used for Classifying = 1594

requiredNum = 1

upperThreshold = 0.051875

lowerThreshold = 0.03125

The classifier uses 1594 unique words to classify the pages, for interest’s sake theses words are:

Words:

send love charlie cover norbert mark agreed rid twice size pomfrey recognize dragon afternoon choice nasty fangs rushed hospital wing although threatening dog match calm midnight saturday bolt remembered charlie’s letter fang tail nothin’ handle leg jus’ baby banged good-bye hut peeves packed crate lots o’ rats brandy fer journey muffled torn sobbed underneath forget ticked nearer corridors panted tallest sudden movement drop invisible shrank shadows lamp detention points wandering potter’s rubbish lies steep throw breathe malfoy’s sing swooping safely hearts filch’s filch mcgonagall’s study wild accept classes add bags warn likely believed says yourselves failed longbottom neville’s granger gives fifty lose lead won hundred ruined cup bottom sobbing pillow gryffindors passing mistake yesterday famous hated hufflepuffs lower clapped cheering weeks miserably business sneaking wood offered team thundered win fun practice seeker drawing class exams studying potions learn charms spells goblin due test library somebody classroom quirrell’s quirrell turban cry peered ajar sorcerer’s brothers percy’s rat fat name’s scabbers prefect afford money month dudley’s clothes proper birthday presents cheer sounding impressed brave bet worst families train london speeding flick cart breakfast sandwiches candy dursleys pockets buy carry flavor chocolate pumpkin wands wanting paid eleven compartment tipped hungry lumpy package dry share eating pack card missing cards witches frog man’s wore beard picture albus considered greatest dragon’s nicolas flamel chamber music hang he’ll pile amazed weird interested careful warned ordinary ones grass touch neat knock spat locker nimbus thousand happier proud sweet damp lift distance jumping wooden setting sun shown figure swiftly faded gliding silently enter branches tree tight broomstick leaves strained severus private interrupted perfectly chat petrified squeaked crabbe goyle everyone’s party stole kitchens force fluffy guarding needs stone’s stands braver cracked check sweeping temper notes dumbledore’s useful steal food forgive saved people’s minds father’s protect sir mirror brilliant drinking elixir sometimes questions suggest sweets afraid liking choked absolutely rumors places thinks round contact ’ hurtled d’you sending exploded thoughtfully goes pretty taught feast tomorrow ravenclaw firmly k normal realize ruddy evil ter could’ve egg drink live somehow don’ crying calling wiped present anxiously weak dudley stick purpose kicked thrown roof wishing vernon aunt petunia darkly mail banging bedroom privet drive grabbed confused fighting cupboard received clock rang dressed mustn’t lights letters crept squashy clicked horror kitchen vernon’s lap addressed ink pieces stayed ways piece bathroom burning eggs furious post complain badly amazement tired ill reminded fireplace ducked argue wrenched car television sharp shake ’em mutter howling twin beds sheets awake ate fear follow instructions wherever foolish lord shivered easily gringotts trip diagon alley shaken leaky racing finding realizing noticing ankles master answered rounded desperately breathed smell reflection gotten impatiently devil’s snare glaring nostrils snake shadow unicorn stumbled fool save mercy hissed sprang wrist scar struggling wildly knocking clean landing blinding cannot bewildered burned play chess visit norris sank insides ice desperate relieved slipping hundreds discovered shows gently whatever desire neither men prepared dreams socks christmas struck convinced holidays trunk wished flash term view row makes dying defense copying werewolf wiping field wonderful lessons possibly rooms wood’s meanwhile grim practicing using sleeve early snitch finish game hufflepuff blimey mistaking teams marched ouch poked grinned broadly bludger squinting choose weasleys stammered stretched breaking warning dive cheers streaked happening hesitated clambered leaping sped rolling scarlet inches erupted dancing parvati patil nearby busy yer you-know-who liked forehead powerful lived age orders load clenched expected sofa coat sword dursley danger vol- you- human ours outta powers ’cause somethin’ stumped warmth lock kick positively sorts lily james born teach whirled poking trousers cast pig bushy keen meself expelled gamekeeper hedwig borrowed quill lucky lesson start-of-term banquet dungeons floating roll paused names potion-making gift effort curled sneer dursleys’ herbs ignoring dungeon seamus information plant cheek continued pairs simple swept whom perfect slugs clouds smoke twisted collapsed moaned wave spirits hate hairy crack kettle speed ball wind screams pull fist faster dived flashed furiously excuse curiously writing barked bin swooped boys oliver expression clue captain decent training player joking youngest century secret lee passageway coolly wizard’s yours deadly building voldemort’s glinting blinked piled shop toilet doubt prevent blankly wife order truly merely losing battle beautiful therefore 11 killing deeply older mother’s protection forever soul platform nine three-quarters toad keeps hell yellow interesting brown cleared waved extra history rise asking sounds mom imagine ends houses romania dragons daily prophet security news must’ve happens you-know- mentioned entering balls players interest slight freckles children suggested locked rain beat bought sea perched miserable shack certain clapping kindly wicked grin bobbing whipped sliding mood blew fierce filthy blankets rumbling dangling hoped keeper keys jerked where’s crash doorway completely squeezed easy frozen budge lump stranger hide yeh’ve knot box cake written rubbing stronger flickering sunk weight yeh’ll stare abou’ anger marks parents’ rage eagerly gasp twins prefects pinned peas plastic bang blast merrily joke flaming wine including sleepy stolen badge whoever bother curtains four-poster silk anywhere grunted dormitory lady horribly faint whispering difficulty piercing screaming stuffing shrieking tall suit armor floors directly replied narrow solid backed ages important lines homework moaning months summer what’re st who’d ireland garden putting refused twitched beamed sprout glance nervously bundle lightning map kiss handkerchief grip laid twinkling shone roar street orange cat murmured beside special scream bottles meeting country owls mantelpiece pictures blond fair woke dream demanded bacon groaned spiders troll club oy hitting yell ugly snout terror shouting escape jump blow thud bringing fury teacher rules pretending handing disappointed plates motionless student bloody baron croaked rising upon blame noses sniffed madly slumped trapdoor approached ring lowered struggled fought leaning fan angelina glared hooch gathered fifth fluttering flashing brooms whistle alicia katie bell goal posts space raising attacked rush wings echoed spun foul angrily thomas soccer village doin’ pleasure herbology lives flushed excited chairs jet sparks color gap feathers lip brother dragged animal she’ll mine sorting black-haired cross fit magnificent crowded shortly honor whichever she’d fixed glancing ought ghost oddly double candles tables among avoid stars bewitched placed dirty judge hurry likes flames pay honestly unpleasant attack sprinted overhead squinted examined darted impossible stuffed crashed flight reveal knight persuade frightened anxious discuss pressed godric’s hollow bowed spectacles bring future child seriously growing trust grew steadily louder road vast sirius shapes pounding trick shape three-headed bill cats snap vision younger bus bald everybody camera visited sticks ickle handsome ribs sink board smashed blank determined massive poison fourth hey sister mumbled moments grinning wrote flourish loss level list mixture roots galleons cage center brightly scroll plate eaten headless nick finnigan irritably dad’s married gran guard eight accidentally transfiguration goblet harder stadium bat smaller zoomed forcing guards member understood problem fallen halloween objects flicked clever startled uncomfortable overheard lavender deserted search heading gigantic passage emerged patch dull trunks edged ollivander shout company removed annoyed ’re law armchair creatures frantically pick protected smoothly surrounding proof pity moon enjoying cool werewolves striding fiercely gaze track lifted silvery unable path group pushing headed hello steam express surrounded raise sweaty forgot beyond dragging tongue dress password scrambled girls smooth lawn falling slip higher sticking oak soared cheeks sneered finger picking familiar c’mon lake opening tunnel checking ripped shooting appearance monster basilisk dreadful diary riddle clatter floo powder riddle’s awkwardly fawkes echoing lockhart puzzled glow gilderoy phoenix extraordinary myrtle’s myrtle grimly slime tense tom closing attacks daughter scene secrets cornelius fudge anybody circular award slug amount bow examining examine smirking necklace cabinet entirely sunlight alongside dobby galloping swinging vegetable accompanied whomping willow shining travels lockhart’s gleaming page plenty mandrakes mandrake narrowly slytherin’s serpent spitting winking elbow sinking spreading mudblood dobby’s flung settled weasley’s weakly remained hearing avoided pages daresay gazed hoarsely infirmary attacker final fellow thoroughly mudbloods loose justin tip ernie macmillan heir stall spare mud gazing one’s pitch applause marching further penelope lowering argus popping highly shivering hopefully contents vanishing alarmed write dimly parseltongue remaining salazar azkaban lucius elf attempting apparently arthur definitely permission appear act minister traveling cabin flickered panting message ended mass murdered surface inch abruptly why’s signed pixies shrilly suspiciously guilty bleeding attempted allow aware expecting symbol colin creevey conversation muggle-born hastily trace advanced evidence searching innocent admit comes fresh effect often dim tone dashed signs it’d position fake amusin

g tear urgently froze boring escaped murder replaced gripping widening avoiding banshee jeered house-elves instantly house-elf ancient department burrow rooster trolls staff impression troubled members fully subject divination chose bedside aloud emerging bearing causing memories squealed elder planning results knife photograph proved measures task half-blood gulped copy disbelief remain conjured swear evidently dust crouch vaguely mist feebly teaching champion amongst unconscious previous figures intently thoughts dont black’s dementors lupin crookshanks peter pettigrew paw transformed betrayed tied lupin’s buckbeak patronus prongs diggory cedric drifting descended guardian wedding pettigrew’s thirteen caused bar attempt honeydukes wormtail padfoot firebolt boggart crystal essay weather staggered marauder’s hogsmeade indignantly cho presence dementor firebolt’s ginger sirius’s chain carriage remus squeaking enchanted occurred buried lunch trelawney’s trelawney features undoubtedly applauding normally hi opportunity shrugged chosen mysteries decide apparate behavior despite pointless rough offer connection sought continue tournament viktor krum triwizard mad-eye moody gloves skrewts fleur delacour durmstrang ship moody’s destroy summoning campsite beauxbatons madame maxime karkaroff winky imperius barty tent eaters crouch’s bertha jorkins auror karkaroff’s pensieve cedric’s bagman plans bagman’s skeeter champions hex bulgarian fleur’s rita correct flesh injured judges ludo pigwidgeon eater maze krum’s nagini tasks protective requirement blackened astonished jinx mungo’s fix aberforth husband pouch tonks mundungus kingsley umbridge luna lovegood dolores kreacher umbridge’s occlumency prophecy bellatrix grimmauld bellatrix’s prince narcissa mclaggen slughorn felix felicis rufus scrimgeour slughorn’s prince’s one’ locket morfin demelza greyback horcrux horace horcruxes yaxley griphook xenophilius beaded

The following is the confusion matrix we got after testing this classifier on 192 test cases

| 1 | 2 | 3 | 4 | 5 | 6 | 7 |

----|-------|-------|-------|-------|-------|-------|-------|

1 | 12 | 2 | 1 | 0 | 2 | 0 | 0

2 | 1 | 13 | 1 | 0 | 0 | 0 | 0

3 | 1 | 0 | 15 | 2 | 2 | 0 | 0

4 | 0 | 0 | 0 | 27 | 1 | 0 | 1

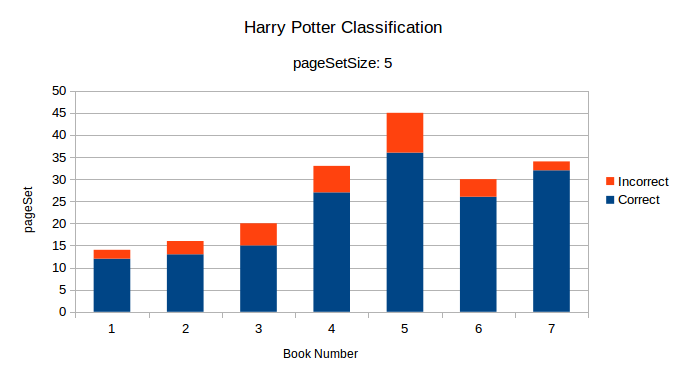
5 | 0 | 1 | 2 | 3 | 36 | 1 | 0

6 | 0 | 0 | 1 | 1 | 2 | 26 | 1

7 | 0 | 0 | 0 | 0 | 2 | 3 | 32

This confusion matrix has an accuracy of 83.85%. There a quite a few interesting observations that can be made about the harry potter books from this matrix.

The first thing that jumps out as obvious in this matrix (and is reflected in all the others) is an obvious bias to book 5 and book 7.



This could indicate multiple things. The bias towards book 5 could be because book 5 is the largest book out of all seven Harry Potter books. There are could be many references and similarities to the earlier books. Similarly, the bias towards book 7 can be explained in this way. Book 7 is the last book of the entire series which could try summarize the entire plot.

We expect a misclassified page set is more likely to be classified as either the previous book or the following book. According to the confusion matrix, this is mostly the case. However, we note that book 5 has the highest number of misclassifications and they are spread out. We attribute this to the fact that book 5 is the largest book in the series and contains the most words.

Interestingly, book 1 and book 7 have the least number of misclassifications. It makes sense to think of book 1 as the introductory book, hence it is a bit unique. This shows J. K. Rowling’s writing style did change from the first book. Book 7 is obviously not referenced throughout the series (as it is the last book) but can reference previous books. We see that book 7 is only misclassified once as book 4 and 6, which interestingly tells us book 7 is very much different to the other books.

We can conclude that the later books tend to be more similar to the earlier books than the earlier books are to the later ones. These results are expected and it is most likely that characters, places, jargon and ideas are introduced in the earlier books and are carried throughout the later books.

Another interesting aspect of this confusion matrix is that other than the bias to book 5 it is nearly an upper triangular matrix, with majority of the incorrect guesses being in the upper triangle. This means that when it chooses incorrectly the classifier is more likely to pick an earlier book than it is to pick a later one. This result is to be expected as it means the later books are more similar to the earlier books than the earlier books are to the later ones. This is most likely because names of people and places, jargon and ideas are introduced in one book and carried through to the later books but all the books before the one in which this was introduced will have no reference to it. This is not only evident in this matrix but in all the other matrices as well, see the following graph which shows the split up of incorrect guesses between those that where later than the correct book and those that where earlier, as you can see majority of the time we classify as an earlier book rather than a later book.

A further surprising idea seen from the confusion matrices is that you would expect that when a page set is classified incorrectly it is most likely to be classified as either the previous book or the following book. The confusion matrices show that on average 29% of the incorrect guesses are classified as either the previous or following book. This however is very close to random as the probability of a page set being any one book is 14%, so the probability of it being any two is 28%, thus not much information can be derived from this.

Interestingly however, it seems from the confusion matrices that book 6 in particular is easily confused with and thus very similar to books 5 and 7.

**Benchmarking**

In this test we try to classify the following books