CA#0 Report

Mahdy Mokhtari - 810101515

Main purpose:

The main goal of this project was to write a program that can get a random writing and identify what genre it is.

Basic description:

At first, we had a database that basically was our “books_train” file and we had to collect some information about the repetition, uniqueness and variety of words in an intended genre (specifically; Novels, Short stories, Children’s and teenager stories, Sociology, Business management, Islamic books) and convert it into a useful set of information, our main data Frame, Bag-of-Words.

Then using Bayes theorem, we calculated the probability of the writing for each genre and the one with the most significant probability is chosen as the book’s genre (category).

At last, we checked the accuracy of our program with different methods of implementation and use of techniques such as Additive-smoothing, garbage-words, etc.

Flow:

0. file **faze1** starts

1. imported needed **libraries**

2. defined some useful **global variables** (sets, dictionaries, tuples)

3. got **permissions** for all the different fazes and for a complete report

4. **initialized** the const values (non-useful characters of Persian language)

5. started the **pre-process**

    5.1. got the name of the train file and opened it with panda’s **read_csv()**

    5.2. founded **all categories** from our data frame “df_train”

    5.3. **iterated** through **all the words** in “df_train”

        5.3.1**. filtered** each word using different methods (lemmatizer, stemmer, normalizer, garbage words, punctuation marks, numbers, ect.)

        5.3.2. added the words to the set “**all_words**”

5.3.3. added the words to the dictionary, ”**all_words_in_each_category**” that is our main storage of our data that we will use to make the BoW in near future

5.4 convert our data into a pandas data frame(**BoW**) using “all_words” and “all_words_in_each_category” (**the columns are our words and the rows our categories**)

5.5. **cleaned up** all the data structures to pre-process the test file

5.6 **test file** normalized and all the words of each writing including the description and the title were poured into a dictionary that the keys are the index of the row and the values are all the words in each row

6. some summations and functions were done in **re-initializing** to minimize the run time of the program

7. file **faze2** starts

8. iterate in all the **rows** of the test file

8.1. using the BoW we calculated the probability of **log(p(C)) + sum(log(p(x[1] | C) ... log(p(x[n]|c))**

8.2 chose the **max probability** for that row

9. returned all the **founded categories** for each row(book)

10. file **judge** starts

11. checked the **accuracy** of our program in details

12. checked the **run time** of the program

supplementary information:

stemmer and lemmatizer: to get the root and the singular form of each word; this will increase the programs accuracy because the concept of each word matters not the exact writing of it and it will increase the precision of the finding of the genre

garbage words: punctuation marks, numbers and the prepositions and some words that are used a lot in sentences like some verbs aren’t useful for us to estimate the genre and from the experimental evidence collected from the project it will decrease the accuracy

additive smoothing: for different alphas it will give different results but if it is near to the alpha 1 we have the best accuracy and in general it doesn’t differ a lot and if we don’t use it we will get terrible results specially when a word isn’t in the BoW and that will cause a log(0) that isn’t defined so it will cause some kind of error or even if we ignored zero, our accuracy wasn’t even slightly compatible to the version when we have used additive smoothing

Files:

Zip -> judgeCA0.py, CA0faze1, CA0faze2

Run the judgeCA0.py as the main file

Report

Exact accuracy report from my program for each section in bonus faze:

ACCURRACY:

+without additive smoothing we will get a error because of the log(0) but if alpha is a very low number it is equivalent to not using additive smoothing

With additive smoothing: 81%

Without additive smoothing: 66%

Reports of other wanted states:

---

perm_lemm: True, perm_stem: True, perm_clear_garbage_words: True, alpha: 1

Accuracy of each category:

74% داستان کوتاه:

80% کلیات اسلام:

92% مدیریت و کسب و کار:

75% داستان کودک و نوجوانان:

90% جامعه‌شناسی:

82% رمان:

runTime: 3.0min 19.7sec

accuracy: %81

---

perm_lemm: True, perm_stem: True, perm_clear_garbage_words: True, alpha: 0.7

Accuracy of each category:

92% مدیریت و کسب و کار:

75% :داستان کودک و نوجوانان

82% :کلیات اسلام

72% :داستان کوتاه

78% :رمان

88% :جامعه‌شناسی

runTime: 3.0min 49.0sec

accuracy: %80

---

perm_lemm: True, perm_stem: True, perm_clear_garbage_words: True, alpha: 1.5

Accuracy of each category:

71% :داستان کودک و نوجوانان

80% :کلیات اسلام

92% :مدیریت و کسب و کار

90% :جامعه‌شناسی

87% :رمان

71% :داستان کوتاه

runTime: 3.0min 43.7sec

accuracy: %81

---

perm_lemm: False, perm_stem: False, perm_clear_garbage_words: True, alpha: 1

Accuracy of each category:

75% :داستان کودک و نوجوانان

78% :رمان

83% :کلیات اسلام

74% :داستان کوتاه

91% :مدیریت و کسب و کار

90% :جامعه‌شناسی

runTime: 3.0min 51.5sec

accuracy: %81

---

perm_lemm: True, perm_stem: True, perm_clear_garbage_words: False, alpha: 1

Accuracy of each category:

74% :داستان کودک و نوجوانان

87% :رمان

88% :جامعه‌شناسی

78% :کلیات اسلام

92% :مدیریت و کسب و کار

62% :داستان کوتاه

runTime: 7.0min 7.1sec

accuracy: %79

---

perm_lemm: True, perm_stem: True, perm_clear_garbage_words: True, alpha: 1e-07

Accuracy of each category:

50% :داستان کوتاه

80% :مدیریت و کسب و کار

62% :داستان کودک و نوجوانان

76% :کلیات اسلام

88% :جامعه‌شناسی

84% :رمان

runTime: 3.0min 45.218256999971345sec

accuracy: %73

---

perm_lemm: True, perm_stem: True, perm_clear_garbage_words: True, alpha: 1e-13

Accuracy of each category:

76% :کلیات اسلام

83% :رمان

60% :داستان کودک و نوجوانان

74% :مدیریت و کسب و کار

جامعه‌شناسی: 86%

داستان کوتاه: 44%

runTime: 3.0min 21.8sec

accuracy: %70

---

perm_lemm: True, perm_stem: True, perm_clear_garbage_words: True, alpha: 1e-90

Accuracy of each category:

کلیات اسلام: 75%

داستان کودک و نوجوانان: 55%

مدیریت و کسب و کار: 66%

داستان کوتاه: 43%

جامعه‌شناسی: 82%

رمان: 83%

runTime: 3.0min 26.4sec

accuracy: %66

---

perm_lemm: True, perm_stem: True, perm_clear_garbage_words: True, alpha: 1e-100

Accuracy of each category:

مدیریت و کسب و کار: 66%

داستان کوتاه: 43%

جامعه‌شناسی: 82%

داستان کودک و نوجوانان: 55%

رمان: 83%

کلیات اسلام: 75%

runTime: 3.0min 24.2sec

accuracy: %66