TTDS CW2 Report

December 2020

1 Overall Work

Coursework 2 is split into three main parts: IR evaluation, text analysis and text classification.

IR Evaluation
In this part, I build a module that evaluates six IR systems using different retrieval scores. The measures are P@10, R@50, r-precision, AP, nDCG@10 and nDCG@20. Based on the average scores achieved for each system, I get the best system according to each score. For each best system with a given score, I find this system is not statistically significantly better than the second system by using 2-tailed t-test, with p-value of 0.05.

Through programming, I have further mastered the calculation methods of various evaluation indicators.

Text Analysis
In this part, I use the Quran, New Testament, and Old Testament each to be a separate corpus, and their verses as individual documents. I use three methods to select features: Mutual Information, X^2 and LDA. The highest scoring words for each corpus obtained by MI and X^2 are relatively similar. But the information contained in the top ten tokens calculated by LDA is more accurate.

Through the calculations in this section, I have deepened my understanding of the three corpora and know their similarities and differences.

Text Classification
In this part, I do the text classification. I extract BOW features and train SVM classifiers to predict the labels. For each dataset (training, development and testing), I compute the precision, recall, and f1-score for each of the three classes, as well as the macro-averaged precision, recall, and f1-score across all three classes. Then, I managed to improve the performance of the baseline system. I tried two methods. The method of tuning parameters (C=10) increases the Macro-F1 scores on dev set and test set for 0.018 and 0.014 respectively.

Through this section, I learned to preprocess the data and then use the model SVM to make predictions. I also learned how to calculate a variety of metrics.

2 IR Evaluation

Table 1 shows the average scores achieved for each IR system.

	P@10	R@50	r-precision	AP	nDCG@10	nDCG@20
S1	0.390	0.834	0.401	0.400	0.363	0.485
S2	0.220	0.867	0.253	0.300	0.200	0.246
S3	0.410	0.767	0.448	0.451	0.420	0.511
S4	0.080	0.189	0.049	0.075	0.069	0.076
S5	0.410	0.767	0.358	0.364	0.332	0.424
S6	0.410	0.767	0.448	0.445	0.400	0.491

Table 1: The average scores achieved for each IR system

Table 2 shows the best performing IR system according to each metric.

	P@10	R@50	r-precision	AP	nDCG@10	nDCG@20
Best	S3,S5,S6	S2	S3,S6	S3	S3	S3
Value	0.410	0.867	0.448	0.451	0.420	0.511

Table 2: The best performing IR system for each metric

Use 2-tailed t-test with p-value of 0.05 to indicate if the best system is statistically significantly better than the second system with that score or not. Results are shown in table 3.

	P@10	R@50	r-precision	AP	nDCG@10	nDCG@20
p-value	0.888	0.703	0.759	0.967	0.882	0.869

Table 3: 2-tailed t-test results

We assume that the best system has a similar performance as the second system. We calculate the p-value to test this assumption. From table 3, it can be seen that the p-values are all greater than the significance level of 0.05. So we cannot reject the null hypothesis. It means that for each best system with a given score, this system is not statistically significantly better than the second system with that score.

3 Text Analysis

3.1 Token Analysis

Table 4 shows the top 10 highest scoring words for Mutual Information and X^2 for each corpus.

	MI			X^2		
	Quran	**OT**	**NT**	**Quran**	**OT**	**NT**
1st	allah 0.153	allah 0.087	jesus 0.065	punishment 965.474	jesus 1296.973	jesus 3268.989
2nd	god 0.027	jesus 0.041	christ 0.037	believers 913.689	king 818.822	christ 1790.610
3rd	man 0.020	israel 0.036	allah 0.019	unbelievers 760.787	christ 647.466	disciples 823.367
4th	punishment 0.019	lord 0.031	disciples 0.016	god 735.803	land 504.265	things 786.652
5th	believers 0.019	king 0.028	things 0.016	verses 641.677	house 474.266	faith 740.932
6th	israel 0.018	christ 0.020	lord 0.016	nation 634.772	allah 458.748	paul 588.945
7th	king 0.017	judah 0.017	faith 0.014	clear 633.465	judah 429.452	peter 560.751
8th	unbelievers 0.016	land 0.016	israel 0.013	messenger 615.957	sons 381.096	lord 533.517
9th	verses 0.014	house 0.014	paul 0.012	man 493.631	punishment 373.161	john 457.702
10th	house 0.013	sons 0.013	peter 0.011	disbelieve 471.661	believers 363.550	allah 423.394

Table 4: The top 10 highest scoring words for MI and X^2 for each corpus

I add some pronouns and verbs in the stop word list, including thou, thy, ye, thee, hath and shalt.

For Quran, the rankings calculated using the two methods are quite different. There are only six words in the top ten scores in both rankings. For OT and NT, the rankings calculated by the two methods are relatively close. There are eight and nine words in the top ten scores in both rankings.

Some words in the top ten rankings did not appear in that corpus. Based on these rankings, I learned that the Quran mainly records the story of Allah. The Old Testament tells about the development of human beings such as Israelis and Jews. The New Testament records the life experience of Jesus.

3.2 Topic Analysis

Topics are labeled from 0-19. Table 5 shows the top 10 tokens and their probability scores for each of the 3 topics that have been identified as being most associated with

each corpus.

	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
Quran	dwell	wrath	power	jesus	beasts	poor	rich	taught	people	sheep
Topic 1	0.048	0.042	0.033	0.031	0.030	0.027	0.023	0.023	0.022	0.021
OT	king	law	mine	house	david	servant	cast	feet	bread	brother
Topic 2	0.089	0.057	0.031	0.029	0.027	0.026	0.025	0.022	0.021	0.020
NT	son	jesus	israel	spirit	children	lord	jews	gospel	midst	filled
Topic 4	0.139	0.129	0.074	0.066	0.052	0.051	0.031	0.022	0.021	0.015

Table 5: The top 10 tokens and their probability scores for each of the 3 topics that have been identified as being most associated with each corpus

Below are the labels I give to each of the three topics for each corpus.

- Topic 1: Power and Wealth
- Topic 2: Privileges and Obligations
- Topic 4: Jesus and Gospel

From the results of the LDA model, I know that the Quran explains the relationship between God and his people. It also includes systems to alleviate the gap between the rich and poor. The Old Testament involves some moral standards and codes and religious rules. The New Testament tells about Jesus preaching the gospel and paradise is not heaven.

I compared the top five topics in the three corpora.

Topic 9 and Topic 10 appear to be common in Quran and NT but not the OT.
Topic 9: [faith god earth lord kingdom man grace heaven made great]
Topic 10: [father lord peace works angel stood delivered good egypt mother]

Topic 4 and Topic 7 appear to be common in OT and NT but not the Quran.
Topic 4: [son jesus israel spirit children lord jews gospel midst filled]
Topic 7: god lord hand heart receive laid sin stone offering part]

The topics obtained by LDA analysis is more representative. Make the characteristics of the corpus more distinct.

4 Classification

Below shows three instances from the development set that the baseline system labels incorrectly.

- For the mystery of iniquity doth already work: only he who now letteth will let, until he be taken out of the way.
  Predict: OT True: NT
- For they have refreshed my spirit and your's: therefore acknowledge ye them that are such.
  Predict: OT True: NT
- And all men shall fear, and shall declare the work of God; for they shall wisely consider of his doing.
  Predict: NT True: OT

These instances contain few tokens that have high scores in MI or X^2 or LDA. So the features are not obvious. The classifications would be incorrect.

I tried two methods to improve performance.

Use the top N features with the highest MI scores
Using tokens with the highest MI scores can better reflect the characteristics of each corpus. I used the top 50%, 40%, 30%, 20% and 10% of the data as features. The gain in the Macro-F1 score that achieved with the method when evaluated on the dev set and test set are shown in table 6.

	50%	**40%**	**30%**	**20%**	**10%**
dev	-0.032	-0.020	-0.016	-0.018	-0.033
test	-0.037	-0.024	-0.022	-0.024	-0.044

Table 6: Use the top N features with the highest MI scores

From table 6, we can see that these five ratios all make macro-F1 score decrease. So this improvement method did not work as expected. It may be because most of the tokens with high MI scores have a high proportion in all three corpora.

Change the SVM parameter C
In the LinearSVC, C is the penalty coefficient used to control the loss function. The greater the C, the greater the punishment for the wrong sample. Therefore, the higher the accuracy in the training sample, and the generalization ability will decrease. That is, the classification accuracy of test data is reduced. On the contrary, if C is reduced, some misclassification error samples are allowed in the training samples, and the generalization ability is strong.

I adjust the parameter C based on the remaining settings of the baseline unchanged. Traverse C equal to 100, 300, 500 in turn. The best result is when C is 100. Then traverse C equal to 10, 30, 50. The gain in the Macro-F1 score that achieved with the method when evaluated on the dev set and test set are shown in table 7.

C	100	300	500	10	30	50
dev	0.004	0.001	0.001	0.018	0.005	0.006
test	0.003	0.001	0.000	0.014	0.013	0.007

Table 7: Change the SVM parameter C

From table 7, we can see that the Macro-F1 scores on dev set and test set increase the most when C is equal to 10. They are 0.018 and 0.014 respectively. Reducing the value of C effectively improves the generalization ability of the model. Therefore, when C is equal to 10, the model is improved.

The final scores for all metrics of the baseline model and the improved model are shown in the $classification.csv$.