Reviewing Large Quantities of Documents

White Paper

Prepared by RINA Systems

October, 2017

Introduction

This paper describes a problem and a possible solution for reviewing large quantities of documents.

When companies are faced with a need to review a large set of documents, e.g. 100,000 or more, they typically address it by using a team of reviewers. The reviewers are given a set of documents and a list of issue tags. Each reviewer who reads a document must decide which issue tags are relevant to this document. As the time to review a document set is typically limited, the more documents need to be reviewed the larger the team of reviewers is typically used. When a reviewer reads a document, he makes a judgment regarding which issue tags should be assigned to the document. It is common in the legal industry today to review a certain number of documents out of the total set and then, using some kind of modeling technology, build a model for each issue tag based on the classified documents and then apply the model on the rest of the population to determine which documents belong to each issue tag without performing a human review.

Problem

Each reviewer has a certain chance of making an error by misclassifying a document with an incorrect issue tag. Errors result in building lower quality models which may lead to misclassifying and/or missing important documents.

This paper shows how the quality of reviewing the total set of documents is affected by the number of reviewers and their ability to correctly review documents. The paper also suggests an approach to reduce errors in reviewing documents, which results in building better models and therefore significantly reducing the number of misclassified and/or missed important documents.

Study

Definitions

There is a population of n documents to be reviewed. There are m reviewers who review documents. During the review process a reviewer reviews a document and tags it with a certain issue tag(s). A human reviewer can do it correctly or make a mistake.

Let us assume that a reviewer r_i reviews q_i documents and tags $f_i \le q_i$ documents correctly.

Let's assume that together all m reviewers review n documents, or $n = \sum_{i=1}^{m} q_i$.

Assumptions

1. Each reviewer's reviews are independent of other reviewers.
2. A reviewer's review of a document is independent of the same reviewer's review of another document.
3. The probability of each reviewer reviewing documents is consistent over time.

Let us also assume that each reviewer's probability of reviewing a document correctly is $p_i, i = \overline{1,m}$.

Let N be the total number of documents tagged correctly by all reviewers. $N = \sum_{i=1}^{m} f_i$.

Then the probability P of all reviewers correctly tagging at least N documents is

$$P = \prod_{i=1}^{m} \sum_{k=f_i}^{q_i} p_i^k (1-p_i)^{q_i-k} C_{q_i}^{\ k}. \qquad (1)$$

It follows from (1) that the larger the number of reviewers who all have the same probability smaller than 1 of correctly tagging a document, the smaller the probability that at least N documents will be correctly tagged. The opposite is true, namely the smaller the number of reviewers whose probabilities of correctly tagging a document are less than 1, the larger the probability that at least N documents will be tagged correctly.

Therefore, we can formulate

A reviewers' law

The larger the number of people who independently review not completely trivial or obvious different and independent blocks of information on the same topics presented over time, the more difficult it is to make a conclusion based on their opinions about the information.

Corollary

Adding reviewers to independently review not completely trivial or obvious different and independent blocks of information on the same topics makes it more difficult to make a conclusion based on their opinions about the information.

Example 1

Let the number of reviewers be 10 and each reviewer's probability of correctly reviewing a document is 0.9. Then the probability of these 10 reviewers correctly reviewing at least 80 documents out of 100 so that each of them would review 10 documents is 0.48299. If, however, the number of the same reviewers is 5, and each reviewer has to review 20 documents, then the probability of these 5 reviewers correctly reviewing at least 80 documents out of 100 is 0.80198. If we reduce the number of reviewers to two, with each reviewing 50 documents, then the probability of these 2 reviewers correctly reviewing at least 80 documents out of 100 is 0.98138. This probability is more than two times higher than using 10 reviewers.

	# of reviewers	Probability of each reviewer	# of docs to review	# of docs to be correctly reviewed by each	Probability of the correct review
1	10	0.9	10	8	0.48299
2	5	0.9	20	16	0.80198
3	2	0.9	50	40	0.98138

Table 1

This example shows that increasing the number of reviewers with the same or close to the same probabilities reduces the probability of correctly reviewing the total document set.

Example 2

Let's reduce each reviewer's probability in the Example 1 from 0.9 to 0.8. Then for 10 reviewers, the probability of them correctly reviewing at least 80 documents out of 100 by each of them reviewing 10 documents, is only 0.02047.

If the number of reviewers is 5, with each reviewing 20 documents, then the probability of these 5 reviewers correctly reviewing at least 80 documents out of 100 is 0.09897.

If the number of reviewers is two, with each reviewing 50 documents, then the probability of these 2 reviewers correctly reviewing at least 80 documents out of 100 is 0.34054, which is almost 17 times more than with 10 reviewers.

	# of reviewers	Probability of each reviewer	# of docs to review	# of docs to be correctly reviewed by each	Probability of the correct review
1	10	0.8	10	8	0.02047
2	5	0.8	20	16	0.09897
3	2	0.8	50	40	0.34054

Table 2

This example together with the Example 1 show that even a small decrease in the probability of correctly reviewing a document leads to a significant decrease in the probability of correctly reviewing the total document set. So, that for 10 reviewers, the reduction in the probability from 0.9 to 0.8 leads to a decrease in probability for the total set more than 24 times, for 5 reviewers the decrease is 8 times and for 2 reviewers, the decrease is almost 3 times (see the tables 1 and 2).

Example 3

Now let us consider two reviewers with probabilities 0.9 and 0.8 respectively. If each person reviews 50 documents, then the probability of correctly reviewing at least 80 documents out of 100 is 0.57810. If the 1st reviewer with the probability of 0.9 would review 80 documents and the 2nd would review 20, the probability of correctly reviewing at least 80 documents out of 100 would be 0.62831.

For two reviewers with probabilities 0.9 and 0.7 the numbers are 0.07811 and 0.23700 respectively.

This example shows that the better the reviewer, more documents should be allocated for him to review (seems obvious), and that there is an optimum allocation of the number of documents to be reviewed between available reviewers.

Example 4

Let us consider changing the number of documents that must be tagged correctly by each reviewer from 9 documents to 8, 7, 6 and 5 to achieve the correct review of the total set and calculate the probability of the correct review for the total set in each case using (1).

	# of reviewers	Probability of each reviewer	# of docs to review	# of docs to be correctly reviewed by each	Probability of a correct review
1	10	0.9	10	9	0.04670
2	10	0.9	10	8	0.48299
3	10	0.9	10	7	0.87917
4	10	0.9	10	6	0.98377
5	10	0.9	10	5	0.99853

Table 3

This Table 3 shows that the lower the percent of documents that must be reviewed correctly out of the total number reviewed by each reviewer, the higher the probability of correctly reviewing the total document set.

Example 5

Let us consider 5 reviewers, with each reviewer's probability of correctly reviewing a document of 0.9. Let us assume that it is necessary that at least 80 documents are reviewed correctly, with the probability of 0.99. Let us ask a question: how many documents does each reviewer have to review to accomplish this assuming each of the 5 reviewers would review the same number of documents? This means we are trying to solve (1) for n.

The answer is 23 and the total number of documents to be reviewed by 5 reviewers is 115.

If each reviewer's probability of correctly reviewing a document is 0.8, then each has to review 28 documents for the total number of documents in the review set of 140.

For the probability of 0.7, each reviewer must review 34 documents, for the total number of 170 documents in the review set. For the probability of 0.6, the numbers are 42 and 210 respectively and for the probability of 0.45 the numbers are 59 and 295 respectively (see the Table 4).

	# of reviewers	Probability of each reviewer	Probability of the correct review	# of docs required to review by each reviewer	Total # of documents need to be reviewed
1	5	0.9	0.99	23	115
2	5	0.8	0.99	28	140
3	5	0.7	0.99	34	170
4	5	0.6	0.99	42	210
5	5	0.45	0.99	59	295

Table 4

This example shows that a decrease in the quality of reviewers led to an increase in the number of documents that need to be reviewed to achieve the required quality of the review. A 33%

decrease in the quality of reviewers (from 0.9 to 0.6) leads to an 82% increase in the number of documents that need to be reviewed (from 115 to 210) to achieve the required review quality. This will translate into the higher cost and longer time to perform the review.

As our objective is to review documents to build a model, we could ask what would be the model's quality that would generate the results shown in the Table 4.

It is well known that

$$F1 = 2\frac{precision * recall}{(precision + recall)}, \quad (2)$$

where

$$precision = \frac{tp}{tp + fp}, recall = \frac{tp}{tp + fn},$$

tp, fp and fn are true positive, false positive and false negative respectively. We can then rewrite (2) as

$$F1 = \frac{2tp}{2tp + fp + fn} \quad (3)$$

Let us assume that it is possible to build the best models, which means the highest F1 score, given the data in Table 4. The last column in Table 4 shows the total number of documents to be reviewed correctly to achieve 99% quality.

This number is the sum of number of documents identified as true positive and true negative. It is clear from (3) above, that F1 is the highest when the number of documents identified as true positive is the highest. So, let's put the numbers in the last column of Table 4 as all true positive.

Then, based on (3) we get the following best possible F1 scores for each reviewer:

F1 for reviewer1 = 2*80/(2*80+35)=0.820513

F1 for reviewer2 = 2*80/(2*80+60)=0.727273

F1 for reviewer3 = 2*80/(2*80+90)=0.64

F1 for reviewer4 = 2*80/(2*80+130)=0.551724

F1 for reviewer5 = 2*80/(2*80+215)=0.426667

	# of reviewers	Probability of each reviewer	Probability of the correct review	Total # of documents that need to be reviewed	Highest possible F1 score
1	5	0.9	0.99	115	0.820513
2	5	0.8	0.99	140	0.727273
3	5	0.7	0.99	170	0.64
4	5	0.6	0.99	210	0.551724
5	5	0.45	0.99	295	0.426667

Table 5

It follows, that in addition to spending more time and higher cost, lower probability reviewers result in producing data that could be modeled by models with a lower predicting ability even if these reviewers review a predefined number of documents correctly.

If we use the eDiscovery conventionally acceptable F1 score of 0.7, we can see from Table 5 that only reviewers with probabilities of 0.9 and 0.8 may be able to do the document review with acceptable quality even though they reviewed a predefined number (80) of documents correctly.

Let us now consider the process of reviewing documents differently. Let us assume that each document is reviewed by two independent reviewers where the second reviewer does not know the document has already been reviewed. This means there are twice as many reviewers as in our previous examples. As before we also assume that each reviewer's probability of reviewing a document correctly is $p_i, i = \overline{1,2m}$. For simplicity and without limiting general conclusions let's assume that the reviewers from 1 to m review documents and the reviewers from $m+1$ to 2m review the documents already reviewed. Then the probability of each document being correctly reviewed is

$$p = 1-(1-p_i)*(1-p_{i+m}) \qquad (4)$$

Then the probability P of all reviewers correctly tagging at least N documents using this process could be obtained by replacing p_i in (1) by p in (4)

$$P = \prod_{i=1}^{m} \sum_{k=f_i}^{q_i} (p_i + p_{i+m} - p_i p_{i+m})^k (1-p_i)^{q_i-k} (1-p_{i+m})^{q_i-k} C_{q_i}^{\ k}. \qquad (5)$$

As in the Table 5 let us consider 5 pairs of reviewers and assume that it is necessary to correctly review at least 80 documents with the probability of 0.99. Let us recalculate using (5) the number of documents that need to be reviewed by each of the two reviewers and compare it to the numbers shown in Table 4 above.

	# of reviewers	Probability of each reviewer	Probability of the correct review	# of docs required to review by each reviewer	Total # of documents need to be reviewed
1	5	0.9	0.99	18	90
2	5	0.8	0.99	20	100
3	5	0.7	0.99	23	115
4	5	0.6	0.99	26	130
5	5	0.45	0.99	34	170

Table 6

It follows that using the process of two reviewers per document requires fewer documents to be reviewed by each reviewer, which reduces time to reach the final result. In the example above, comparing one reviewer with the probability of 0.45 vs two reviewers with the probability of 0.45 each, reduced the number of documents to be reviewed by each reviewer from 295 to 170, or about 42%. Even though the total number of documents to be reviewed by two reviewers in

this case doubles to 340, and is more than the 295 documents to be reviewed in the first case, the reviews of the same documents could be done simultaneously. This reduces the time to reach the final result as in the case of 0.45 probability by 42%. The tradeoff then is in reducing elapsed time by 42% vs increasing cost by 15%, or expressed as a ratio 2.8. Let's call it Time/Cost Ratio or "TCR". Obviously, the higher TCR, the more effective the process and the bigger the time reduction per $100 of cost.

It's easy to see that the lower the probability of a reviewer, the higher the TCR.

Interestingly, if there are 3 reviewers to review the same document then for each reviewer's probability of 0.45 only 26 documents need to be reviewed by each, which leads to 56% elapsed time reduction vs 33% cost increase or 1.7 TCR.

Conclusions

The fewer number of reviewers, the higher the quality of the review of the document set, the better models could be built to apply to the total document set and the fewer misclassified and/or missed documents.

Using reviewers whose probabilities are lower than 0.9 risks a very low review quality of the document set and the chance of building a low-quality model which, if used as is, may lead to more misclassified and/or missed documents than acceptable.

A process of reviewing the same document by more than one reviewer (without the reviewers' knowledge) could be used, if the elapsed time to review a set of documents is very limited. This leads to a reduction in elapsed time but also to an increase in cost.

To make conclusions based on this study we need to be able to estimate each reviewer's probability to correctly review a document. This could be achieved by providing each reviewer with the same predefined set of documents and making a probability calculation based on the review of these documents. (This process could be repeated during the review process at different times as the reviewers' behavior may change both over time and during a day). Once the probabilities are estimated, it is possible to determine if a reviewer should be involved in the review process and, if yes, to estimate the number of documents that should be given to each reviewer to maximize the review quality.

It is recommended to measure the reviewers' probabilities to correctly review a document periodically during the review process to reduce the impact of assumption 3, that the probability of each reviewer to review documents is consistent over time.