Paper 2647-2018

Identifying Semantically Equivalent Questions Using Singular Value Decomposition

Varsha Reddy Akkaloori, Graduate Student - Business Analytics, Oklahoma State University

ABSTRACT

In the past few decades, inquisitive people who are in constant pursuit of knowledge, are visiting Question & Answer sites, such as Quora, Stack Overflow, Yahoo! Answers etc. to find out new things authored by mavens in their respective fields. In order to maintain "content quality", most of the Q&A sites would want its visitors to search their website for an answer to their question before posting a new one. With over 100 million monthly visitors, it's not surprising that many people ask similarly worded questions causing site visitors to spend more time discovering the best response to their question. This also frustrates authors because they feel they need to answer multiple versions of the same question.

This paper aims at solving a challenge released by Quora to improve the experience of its authors and site visitors by grouping queries with similar intent using SAS. Two queries are assumed semantically equivalent, if they could be answered with the exact content. To ensure that different words are processed equivalently as the same representative parent term, Pydictionary module in Python was used for extracting synonyms for the most frequently occurring terms. With the help of SAS Enterprise Miner, singular value decomposition (SVD) was implemented to reduce the dimensions of the term-by-document frequency matrix. Euclidean distance was used to determine distance between sentences that have been projected into the SVD space. The accuracy of the classification is determined by comparing the similarity index to the Target variable present in the data. In addition to identifying a duplicate question pair, duplication could be avoided for the whole corpus by comparing the distance between a given question and corpus of questions.

Further research would be continued to make a utility which would predict if a question is duplicate based on the prior knowledge imbibed into it thereby acting as a recommender system for Quora.

INTRODUCTION

Where else but Quora can a traveler help a chef who was confused to make a list of must visit places and could get cooking tips in return? Quora is a platform to share and gain knowledge. Connecting people who have knowledge to the people who need it would empower everyone to share their understanding to better appreciate the rest of the world.

With numerous people visiting Quora every month, most likely many people ask similar questions with slightly different formations. Numerous questions with similar intent can cause explorers to spend more time discovering the best response to their question, and also could make authors feel they need to answer multiple versions of the same question.

DATA DICTIONARY

The public dataset released by Quora consists of over 400,000 records of potential question duplicate pairs. Each record contains IDs' for each question in the pair, the full text for each question, and a binary value that indicates whether the line truly contains a duplicate pair or not.

Variable	Description
ID	This field represents row number

QID1	Unique ID for Question 1
QID2	Unique ID for Question 2
Question1	Question asked in Quora
Question2	Question asked in Quora
Is_Duplicate	Binary Target Variable indicating if question pair is duplicate

Table 1. Data Dictionary

METHODOLOGY

Firstly, the public dataset on Quora Duplicate questions pairs is used as the data source. Exploratory Analysis on the dataset was performed using SAS. SAS Enterprise Miner is used to clean the data using techniques such as text parsing and filtering. The process flow is illustrated in Figure 2. Using PyDictionary module, consolidated synonym list for the most frequently occurring terms in the term-by-document matrix is created. Singular value Decomposition (SVD) dimensions are computed to transform the original weighted, term-by-document frequency matrix into a dense but low dimensional representation. The Distance Procedure is performed to compute Euclidean distances as a measure of distance/similarity between the documents using SVD dimensions. Finally, the similarity measure is compared with the target variable present in the original data to determine the accuracy of the classification.

Figure 1. Project Methodology

Figure 2. Process Flow

DATA EXPLORATION

Quora Duplicate Questions Dataset was read–in using SAS. Frequency distribution of the binary indicator variable illustrates that 36.92% of the data contains duplicate question pairs and the rest doesn't.

The FREQ Procedure

IS_DUPLICATE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	255045	63.08	255045	63.08
1	149306	36.92	404351	100.00

Figure 3. Frequency Distribution

DATA PREPARATION & CLEANING

For the purpose of classification, binary indicator variable present in the raw data was ignored. The raw dataset now consists of a single column of stacked question pairs along with their question ID's.

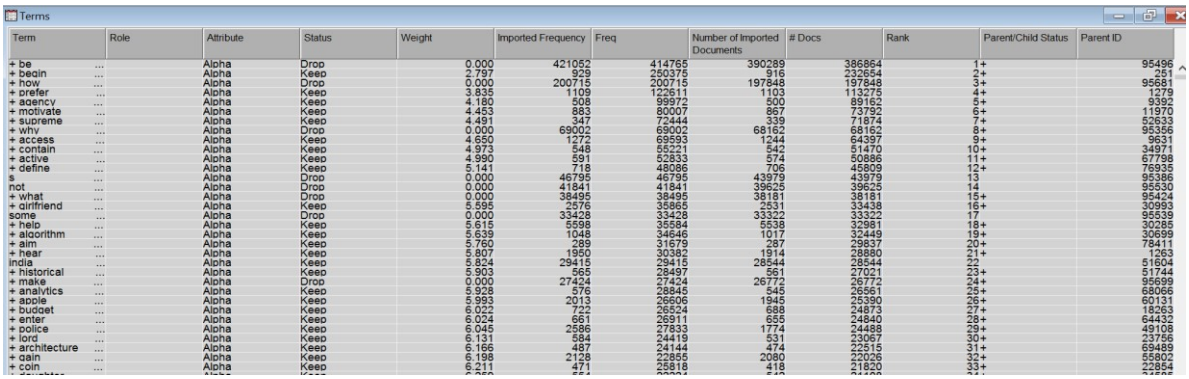
Raw data is imported and cleaned using Text Parsing and Text Filter nodes in SAS Enterprise Miner. In order to reduce the dimensionality of the term by document matrix, parts of speech are turned off but spell check and stemming are performed.

The Log frequency weighting option is used in the Text Filter node to dampen the effect of terms that occur many times in a document. Inverse Document Frequency was used as the term weighting method to give greater weight to terms that occur infrequently in the document collection.

SYNONYM LIST

Usually Information Retrieval is performed by literally matching terms in documents with those of a query, but based on the concept of synonymy, the literal terms sometimes might not match with the query. A synonym list enables us to specify different words that should be processed equivalently, as the same representative parent term.

In order to address this issue in the paper, most frequently occurring terms in the term-by-document matrix has been exported and synonyms have been identified for the top 2,000 terms sorted by descending frequency.


Terms

Term	Role	Attribute	Status	Weight	Imported Frequency	Freq	Number of Imported Documents	# Docs	Rank	Parent/Child Status	Parent ID
+ be	...	Alpha	Drop	0.000	421052	414765	390289	386864	1+		95496
+ begin	...	Alpha	Keep	2.797	929	250375	916	232654	2+		251
+ how	...	Alpha	Drop	0.000	200715	200715	197848	197848	3+		95681
+ prefer	...	Alpha	Keep	3.835	1109	122611	1103	113275	4+		1079
+ agency	...	Alpha	Keep	4.180	508	99972	500	89162	5+		9392
+ motivate	...	Alpha	Keep	4.453	883	80007	867	73792	6+		11970
+ supreme	...	Alpha	Keep	4.491	347	72444	339	71874	7+		52633
+ why	...	Alpha	Drop	0.000	69002	69002	68162	68162	8+		95356
+ access	...	Alpha	Keep	4.650	1272	69593	1244	64397	9+		9631
+ contain	...	Alpha	Keep	4.973	548	56221	542	51470	10+		34971
+ active	...	Alpha	Keep	4.990	591	52833	574	50886	11+		67798
+ define	...	Alpha	Keep	5.141	718	48086	706	45809	12+		76935
s	...	Alpha	Drop	0.000	46795	46795	43979	43979	13		95386
not	...	Alpha	Drop	0.000	41841	41841	39625	39625	14		95530
+ what	...	Alpha	Drop	0.000	38495	38495	38181	38181	15+		95424
+ girlfriend	...	Alpha	Keep	5.595	2576	35865	2531	33438	16+		30993
some	...	Alpha	Drop	0.000	33428	33428	33322	33322	17		95539
+ help	...	Alpha	Keep	5.615	5598	35584	5538	32981	18+		30285
+ algorithm	...	Alpha	Keep	5.639	1048	34646	1017	32449	19+		30699
+ aim	...	Alpha	Keep	5.760	289	31679	287	29837	20+		78411
+ hear	...	Alpha	Keep	5.807	1950	30382	1914	28880	21+		1263
india	...	Alpha	Keep	5.824	29415	29415	28544	28544	22		51604
+ historical	...	Alpha	Keep	5.903	565	28497	561	27021	23+		51744
+ make	...	Alpha	Drop	0.000	27424	27424	26772	26772	24+		95699
+ analytics	...	Alpha	Keep	5.928	576	28845	545	26561	25+		68066
+ apple	...	Alpha	Keep	5.993	2013	26606	1945	25390	26+		60131
+ budget	...	Alpha	Keep	6.022	722	26524	688	24873	27+		18263
+ enter	...	Alpha	Keep	6.024	661	26911	655	24840	28+		64432
+ police	...	Alpha	Keep	6.045	2586	27833	1774	24488	29+		49108
+ lord	...	Alpha	Keep	6.131	584	24419	531	23067	30+		23756
+ architecture	...	Alpha	Keep	6.166	487	24144	474	22515	31+		69489
+ gain	...	Alpha	Keep	6.198	2128	22855	2080	22026	32+		55802
+ coin	...	Alpha	Keep	6.211	471	25818	418	21820	33+		22854

Figure 4. Term by Document Matrix

PyDictionary module in python based on Thesaurus.com is used for creating synonym list. The scraped synonyms are imported back into the Text Parsing Node and the process flow ran until Text Filter Node.

```
from PyDictionary import PyDictionary
dictionary=PyDictionary()



dictionary = PyDictionary(" best", " good"," people", " learn",  " difference", " life",  " know",  " time",  " money", " work",  " year",  " mean",  " thing",
print (dictionary.getSynonyms())
```

Figure 5. Python Code Snippet illustrating the approach used for scraping synonyms from the list of words given as inputs.

COMPUTATION OF SVD

Parsing a document collection generates a term-by-document frequency matrix that is often large. Several thousand documents would require too much of computational time and space to analyze the matrix effectively. To address the challenge of dealing with high dimensional data, singular value decomposition

(SVD) is implemented to reduce the dimensions of the term-by-document frequency matrix by transforming the matrix into a lower dimensional, more compact, and informative form.

Text Cluster Node determines the number of SVD dimensions based on the SVD Resolution and Max SVD Dimensions properties in SAS Enterprise Miner. SVD Resolution was set to low and Max SVD Dimensions are specified as 100. Text Cluster Node resulted in 51 SVD dimensions.

DISTANCE PROCEDURE

Euclidean Distance was used to determine distance between sentences projected into the SVD space because the vectors have been normalized to unit length in SAS Text Miner. In addition, since Inverse Document Frequency weight has been chosen, frequently occurring terms will have already been down-weighted so that the rarer but concentrated terms have the greatest influence on similarity.
Text_Cluster_docs dataset in the workspace of Enterprise Miner folder is considered for the computation of distances using distance procedure. The distance is computed between every question in the corpus to every other question.

SAS Code used to compute the distance:

```
DATA cosine.txtcluster(keep = index textcluster_svd1-textcluster_svd51);
      Set cosine.textcluster_docs;
Run;

DATA cosine.cluster_svd;
      Set  cosine.txtcluster;
      Doc = PUT(index,$8.);
Run;

PROC DISTANCE DATA=cosine.cluster_svd OUT=cosine.cosine_svd_euclid
      METHOD=euclid nostd;
      Var interval(textcluster_svd1--textcluster_svd51);
      Id doc;
Run;
```

	doc	_1	_2	_3	_4	_5	_6	_7	_8	_9	_10	_11	_12
1	1	0	.	.	.	.	.	.	.	.	.	.	.
2	2	0.6382327512	0	.	.	.	.	.	.	.	.	.	.
3	3	1.3994019325	1.3961006507	0	.	.	.	.	.	.	.	.	.
4	4	1.4056995537	1.4013880503	1.4127832256	0	.	.	.	.	.	.	.	.
5	5	1.4154241434	1.4095127402	1.4264126393	1.2692044478	0	.	.	.	.	.	.	.
6	6	1.4048630237	1.4000743606	1.3670791774	1.4116440045	0.7351852111	0	.	.	.	.	.	.
7	7	1.4537222488	1.455002068	1.4177543628	1.4451433676	1.398728743	1.3823705535	0	.	.	.	.	.
8	8	1.3016992177	1.2592265847	1.4651992599	1.4420588258	1.4139317879	1.3810038364	1.5024239083	0	.	.	.	.
9	9	1.3477046892	1.3615563903	1.4375234099	1.3825931963	1.3722551701	1.3645693066	1.2246788444	1.380150486	0	.	.	.
10	10	1.4141215053	1.4160551662	1.4145667491	1.4232270329	1.4217714176	1.421665843	1.3962615754	1.3887002526	1.1913774714	0	.	.
11	11	1.3633792962	1.368113139	1.4061830165	1.338374921	1.24074066	1.2138399285	1.3847768806	1.1951426916	1.281088416	1.3863147748	0	.
12	12	1.4135764786	1.3941619272	1.4374328292	1.3475071422	1.2949725342	1.2375099781	1.3214896402	1.2712422578	1.3436338643	1.4136362375	0.7755123649	0
13	13	1.4130251503	1.412338812	1.417966891	1.0425164821	1.0749608117	1.4201585548	1.3804892494	1.49321323	1.3822935445	1.4168642816	1.408482444	*******
14	14	1.4130781979	1.4076023668	1.4343115902	1.2676750507	1.2332590619	1.2878359909	1.3872074545	1.4158015378	1.3522339093	1.386110963	1.3881046777	*******
15	15	1.4065232505	1.4040274325	1.4121084874	1.4274443871	1.418244542	1.4197858284	1.387002999	1.3658006992	1.3657704812	1.4066086627	1.4264975185	*******
16	16	1.3986088241	1.393932987	1.4224055523	1.4301553535	1.4119487431	1.4049445806	1.3350653594	1.2897514311	1.3150953782	1.3984000392	1.1527467895	*******
17	17	1.4026732838	1.3990986046	1.4276602128	1.1306680549	1.1383074988	1.4008385072	1.3186547011	1.3878227075	1.3198192821	1.4045979622	1.0568119487	*******
18	18	1.4026199384	1.3991007379	1.4296265579	1.1274743873	1.1368406674	1.4019448311	1.3198459183	1.3871556837	1.3206990044	1.4054329213	1.0681363165	*******
19	19	1.3751632111	1.3869143426	1.3199765947	1.4076984008	1.3882918829	1.3689784428	1.3772693533	1.3673366431	1.4179401753	1.3925984584	1.3696853284	*******
20	20	1.3269450002	1.3119636722	1.3954779922	1.4607290272	1.3688687378	1.3021225459	1.3677368024	0.9992450135	1.3298306802	1.4087016314	1.1399342624	*******

Figure 6. Output of Distance Procedure

RESULTS & FINDINGS:

After a trial and error method, distance of 0.75 is chosen to identify semantically equivalent queries. If the similarity metric for a given question pair is <= 0.75, then the pair is considered to be a duplicate. The accuracy of the classification was **62.4%**.

SAS Code used to compute the accuracy of Classification:

```
PROC TRANSPOSE DATA =cosine.cosine_svd_euclid OUT = cosine.test
      PREFIX = similarity;
      BY doc;
run;

DATA cosine.test1(KEEP = pair similarity1);
      SET cosine.test;
      If similarity1 ne . Or qid1 ne qid2;
      Qid1 = SUBSTR(_name_,2);
      Qid2 = substr(compress(doc),1);
      Pair = CATX (",", qid1, qid2);
run;

DATA cosine.rawdata_target(KEEP = pair is_duplicate);
      SET cosine.raw_data;
      Pair = CATX (",", qid1, qid2);
run;

PROC SQL;
      Create table cosine.result as
      Select a. pair, is_duplicate, similarity1
      From cosine.test1 a, cosine.rawdata_target b
      Where a.pair = b.pair
      Order by a.pair;
quit;
```

	pair	IS_DUPLICATE	similarity1
1	1,2	0	0.6382327512
2	101,102	1	1.0988133793
3	103,104	1	0.0053137938
4	105,106	0	1.048253893
5	107,108	1	0.7854749055
6	109,110	0	1.4097787535
7	11,12	1	0.7755123649
8	111,112	0	1.4091269321
9	113,114	0	0.8135216805
10	115,116	0	1.3281126398
11	117,118	1	0.7142178655
12	119,120	0	0.8924533506
13	121,122	0	0.8718087925
14	123,124	0	1.303151708
15	125,126	1	0.6130749454
16	127,128	0	0.060145663
17	129,130	0	0.3258036818
18	13,14	0	1.089968061
19	131,132	1	0.5449814005
20	133,134	1	1.0126010804
21	135,136	1	0.473800657
22	137,138	0	0.5192818418
23	139,140	0	1.4149108837
24	141,142	0	0.7665176319
25	143,144	1	0

Figure 7. Assessment of the Classification

CONCLUSION

This paper illustrates the application of SAS Enterprise Miner to solve a challenge released by Quora. The research is intended to identify semantically equivalent queries in Quora Duplicate questions dataset in order to improve the experience of both the groups of active seekers and writers. Computation of Euclidean distance using Distance Procedure on SVD dimensions of the data, resulted an accuracy of 62.4%.

Since using PyDictonary module in python would only fetch five synonyms for every word, the accuracy could be improved by refining the synonym list considered in the current analysis. Also, considering parts of speech in the text analysis could enhance the capability of classification.

FUTURE SCOPE

The further research would be continued to make a utility which would predict if a question is duplicate based on the prior knowledge imbibed into it thereby acting as a recommender system for Quora.

REFERENCES

- Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS® ® by Goutam Chakraborty, Murali Pagolu, Satish Garla.
- Getting More from the Singular Value Decomposition (SVD): Enhance Your Models with Document, Sentence, and Term Representations, Russ Albright, James Cox, and Ning Jin, SAS Institute Inc., Cary, NC.
- Text Mining Reveals the Secret of Success: Identification of Sales Determinants Hidden in Customers' Opinions, Rafał Wojdan, Warsaw School of Economics
- Using SAS® at SAS: The Mining of SAS Technical Support, Annette Sanders, SAS Institute Inc., Cary, NC, Craig DeVault, SAS Institute Inc., Cary, NC
- SAS® Institute Inc. 2014. Getting Started with SAS® ® Text Miner 13.2. Cary, NC: SAS® Institute Inc.
- SAS/STAT® 13.1 User's Guide The DISTANCE Procedure, SAS® Institute Inc.
- Identifying Semantically Equivalent Questions Using Singular Value Decomposition, Varsha Reddy Akkaloori, Oklahoma State University.

ACKNOWLEDGMENTS

I would like to thank my Program Director Dr. Goutam Chakraborty, SAS Professor of Marketing Analytics and Dr. Miriam McGaugh, Clinical Professor, Spears School of Business, Oklahoma State University for their constant guidance and support throughout my research. Also, I would to thank Mr. Satish Garla, Analytical consultant at SAS for his constant inputs throughout my research.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Varsha Reddy Akkaloori
Phone: (513)282-9496
Email: varshareddya94@gmail.com
LinkedIn: https://in.linkedin.com/in/varsha-reddy-akkaloori

Varsha Reddy Akkaloori is a Graduate student in Business Analytics at Oklahoma State University. She is currently working as a Data Analyst Intern at Epsilon. She is SAS® Certified Advance programmer and Base programmer for SAS 9.