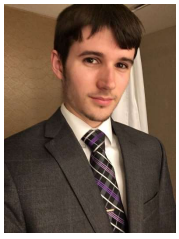
1

Speaker



Alan Johnson, FCAS

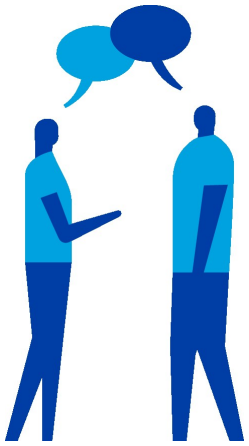
Actuarial Assistant Sr-P&C | Analytics

Sentry

2

Agenda

1. Terminology and Variables
2. What is SVD?
3. Why SVD?
4. Text Preprocessing and Building Dictionary
5. Set Up the Document Term Matrix
6. SVD Components
7. Apply SVD Results to New Documents
8. Use SVD Topic Variables in a Predictive Model



Terminology and Variables

Terminology used in this presentation

- **Token:** a single word
- **Document:** compilation of all tokens for an individual
- **Lemma:** *main form* of a word; several tokens can map to a single lemma
  - Ex. {follow, followed, following, follows} map to lemma {follow}
- **Dictionary:** collection of top lemmas we want to include in the analysis, and the token-to-lemma mapping for each
- **Document-Term Matrix (DTM):** a sparse matrix with one row for each document, showing their counts of included lemmas in the columns
- **Topic:** A set of weights to apply to each lemma count; part of SVD output

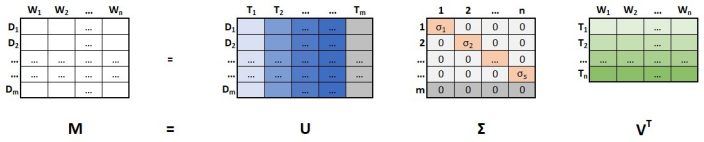
Variables used in this Presentation

- **m** = total # of documents
- **n** = total # of lemmas included in dictionary
- **s** = min(m, n)
- $\mathbf{D_i}$ = i^{th} document
- $\mathbf{W_i}$ = i^{th} lemma
- $\mathbf{T_i}$ = i^{th} topic
- $\boldsymbol{\sigma_i}$ = i^{th} singular value
- $\boldsymbol{ud_i}$ = number of unique documents lemma i appears in

What is SVD?

7

What is SVD?



Singular Value Decomposition

- Decomposes *m x n* matrix into three multiplicative components:
  - **U**: *m x m* matrix of **Left Singular Vectors**
  - **Σ**: *m x n* diagonal matrix of **Singular Values**
    - Larger singular values means that [the topic] explains a larger portion of the variance [between documents]
    - There will only be *s = min(m,n)* singular values, extra rows or columns will be all 0s
    - Can also be written as an *s x s* matrix (including only rows/columns with singular values), would make **U** *m x s* and $\mathbf{V}^T$ *s x n*
  - $\mathbf{V}^T$: *n x n* matrix of **Right Singular Vectors**

Note: U and V are matrices of orthogonal vectors, and Σ can be thought of as the "mapping" between the two to get to M.

8

Why SVD?

9

Why SVD?

Quickly parses text into topics

- Uncovers complex similarities and differences that could not be found manually
- The top topics explain a significant portion of variation in documents, which may uncover predictive attributes of the text

Flexibility

- Can control size and lemma content of dictionary
- User ultimately chooses how many "topics" are included

Easy to apply results to score new documents

10

Text Preprocessing and Building Dictionary

Illustrative Example

Direct Writer Prospect Model

- Direct sales producers have many new business opportunities in their territory to potentially pursue
- Model helps identify which opportunities have the best probability to sell (given they're quoted) at various points in sales process
- Producer activities (emails, phone calls, visits, etc.) are logged with freeform text descriptions
  - The data from these activities plays a large role in the model's predictions
  - We will look at an *illustrative* example of how we create the SVD topic variables for this model

Raw Activity Data

Opp_Index	Actv_Index	Type	Description
1	1	Visit	Long conversation with Eric. Is happy with CompetitorOne, but convinced him to let us quote. Fact find is Tuesday.
2	2	Visit	Asked George if he had given any thought to quoting with Sentry, he said that his wife Amy handles that, and she likes to keep it with a local. He said we can always check in case local guy makes her mad this year when up for renewal. Will call to try to reach Amy.
3	3	Phonecall	Talked to Brett. He was short, abrupt, and hung up on me.
4	4	Visit	Chris was unavailable. Left card, receptionist said he will call in a few days.
5	5	Phonecall	Did not want to transfer me to Tyler, emailed instead.
6	6	Visit	I stopped by to see Martin however he was busy in a manager's meeting. I left a brochure and business card with the receptionist.
7	7	Phonecall	Met with Judy, Kevin was also on the phone, seemed interested in looking at Sentry again. Follow up in a month or so to schedule FF.
...	...	...	...
23	26	Phonecall	Ross said he would speak with Jordan about quote this year.
23	27	Visit	Met owner Ross. We got a tour of the shop. Jordan was out on the road. And apparently the daughter also has a hand in the business. We got Ross's business card, contact him as soon as possible.
23	28	Visit	Checking with UW about 2018 denial of quote due to roof issue and ponding water evidence before I offer to quote. Owner was receptive to quote during visit.
24	29	Visit	Left contact information with HR. They wouldn't give me any of their specific contact information but she said she will pass it on. Will need to follow up, it's a pretty locked down place.
25	30	Phonecall	Called and spoke with Dennis. Has just him and his son. Said that he is happy with CompetitorSix and their pricing, but would be open to the quote. Send him an email and he will give me the policies.

Activity Text Preprocessing

Opp_Index	Actv_Index	Type	Description
1	1	Visit	Long conversation with Eric. Is happy with CompetitorOne, but convinced him to let us quote. Fact find is Tuesday.
2	2	Visit	Asked George if he had given any thought to quoting with Sentry, he said that his wife Amy handles that, and she likes to keep it with a local. He said we can always check in case local guy makes her mad this year when up for renewal. Will call to try to reach Amy.
3	3	Phonecall	Talked to Brett. He was short, abrupt, and hung up on me.
4	4	Visit	Chris was unavailable. Left card, receptionist said he will call in a few days.
5	5	Phonecall	Did not want to transfer me to Tyler, emailed instead.
6	6	Visit	I stopped by to see Martin however he was busy in a manager's meeting. I left a brochure and business card with the receptionist.
7	7	Phonecall	Met with Judy, Kevin was also on the phone, seemed interested in looking at Sentry again. Follow up in a month or so to schedule FF.
...	...	...	...
23	26	Phonecall	Ross said he would speak with Jordan about quote this year.
23	27	Visit	Met owner Ross. We got a tour of the shop. Jordan was out on the road. And apparently the daughter also has a hand in the business. We got Ross's business card, contact him as soon as possible.
23	28	Visit	Checking with UW about 2018 denial of quote due to roof issue and ponding water evidence before I offer to quote. Owner was receptive to quote during visit.
24	29	Visit	Left contact information with HR. They wouldn't give me any of their specific contact information but she said she will pass it on. Will need to follow up, it's a pretty locked down place.
25	30	Phonecall	Called and spoke with Dennis. Has just him and his son. Said that he is happy with CompetitorSix and their pricing, but would be open to the quote. Send him an email and he will give me the policies.

Removing Permutations of a Phrase with Regex

```
def remove_negation_competitor(text):
    spl_phr = re.compile(
            r'\s?'
            r'(is (\w+\s)?|isn\'t (\w+\s)?|isnt (\w+\s)?|were (\w+\s)?|weren\'t (\w+\s)?|werent (\w+\s)?|was (\w+\s)?|wasn\'t (\w+\s)?'
            r'|wasnt (\w+\s)?|hasn\'t\sbeen (\w+\s)?|hasnt\sbeen (\w+\s)?|haven\'t\sbeen (\w+\s)?|havent\sbeen (\w+\s)?'
            r'|have\sbeen (\w+\s)?|aren\'t (\w+\s)?|arent (\w+\s)?|are (\w+\s)?|not (\w+\s)?)?'
            r'(happy|unhappy|comfortable)'
            r'\s(with)'
            r'\s('
            r'(competitorone|competitortwo|competitorthree|competitorfour|competitorfive|competitorsix)'
            r'((?<=[a-z])\'(?=[a-z])s)?)'
            r'|'
            r'((his\s|her\s|the\s|their\s|your\s)?'
            r'(current\s|current\sinsurance\s|insurance\s|incumbent\s)?'
            r'(advisor|agency|agent|broker|carrier|insurer|provider|them|company)'
            r'))')

    text = spl_phr.sub('',text)
    return text
```


Preprocessed Activity Data

Opp_Index	Actv_Index	Type	Mod_Description
1	1	Visit	long conversation with but convinced him to let us quote fact find is
2	2	Visit	asked if he had given any thought to quoting with sentry he said that his wife handles that and she likes to keep it with a local he said we can always check in case local makes her mad this year when up for renewal will call to try to reach
3	3	Phonecall	talked to he was short abrupt and hung up on me
4	4	Visit	was unavailable left card receptionist said he will call in a few days
5	5	Phonecall	did not want to transfer me to emailed instead
6	6	Visit	i stopped by to see however he was busy in a manager's meeting i left a brochure and business card with the receptionist
7	7	Phonecall	met with was also on the phone seemed interested in looking at sentry again follow up in a month or so to schedule fact find
…	…	…	…
23	26	Phonecall	said he would speak with about quote this year
23	27	Visit	met owner we got a tour of the shop was out on the road and apparently the daughter also has a hand in the business we got business card contact him as soon as possible
23	28	Visit	checking with underwriting about denial of quote due to roof issue and ponding water evidence before i offer to quote owner was receptive to quote during visit
24	29	Visit	left contact information with hr they wouldnt give me any of their specific contact information but she said she will pass it on will need to follow up its a pretty locked down place
25	30	Phonecall	called and spoke with has just him and his son said that he and their pricing but would be open to the quote send him an email and he will give me the policies

Build Dictionary

1. Install required R packages for SVD process
   - dplyr, tibble, tidyr, readr, stringr, textstem, lexicon, quanteda, quanteda.textmodels, purrr, ggplot2
2. Import preprocessed activity data
3. Find unique lemmas (using all activities) and remove non-negation stop words
   - Unique lemma list uses textstem::lemmatize_words function with lexicon::hash_lemmas (vocabulary to get token-lemma mapping)
4. Keep top k% of lemmas
   - Example uses top 12%; can customize further after applying cutoff
5. Grab token-lemma mapping for selected lemmas

Build Dictionary

Remove Non-Negation Stop Words

ActivityID	OpportunityID	Token	Lemma	Non-Negation Stop Word?
1	1	long	long	N
1	1	conversation	conversation	N
~~1~~	~~1~~	~~with~~	~~with~~	Y
~~1~~	~~1~~	~~but~~	~~but~~	Y
1	1	convinced	convince	N
~~1~~	~~1~~	~~him~~	~~him~~	Y
~~1~~	~~1~~	~~to~~	~~to~~	Y
1	1	let	let	N
1	1	us	us	N
1	1	quote	quote	N
1	1	fact	fact	N
1	1	find	find	N
~~1~~	~~1~~	~~is~~	~~be~~	Y
…	…	…	…	…
17	17	called	call	N
~~17~~	~~17~~	~~and~~	~~and~~	Y
17	17	spoke	speak	N
~~17~~	~~17~~	~~with~~	~~with~~	Y
~~17~~	~~17~~	~~she~~	~~she~~	Y
17	17	still	still	N
17	17	hasn't	hasn't	N
17	17	confirmed	confirm	N
~~17~~	~~17~~	~~when~~	~~when~~	Y
~~17~~	~~17~~	~~their~~	~~their~~	Y
17	17	insurance	insurance	N
17	17	policies	policy	N
17	17	renew	renewal	N
…	…	…	…	…

Compile Unique Lemmas

Lemma	ud_i
abrupt	1
across	1
addition	1
advise	2
agency	1
agent	2
ago	1
agree	1
alot	1
already	1
also	4
always	1
anyway	1
apparently	1
approach	1
around	1
ask	3
auto	1
away	1
back	3
behind	1
…	…

Build Dictionary

Keep Top Unique Lemmas

Lemma	ud_l
also	4
ask	3
back	3
broker	3
business	3
busy	3
call	9
can	4
card	5
check	5
contact	3
doesn't	3
email	6
fact	4
find	4
follow	6
get	6
give	6
information	3
insurance	3
interest	3
…	…

Grab Mappings for Final Dictionary

Lemma_ID	Lemma	Token	Token_ID
1	also	also	1
2	ask	ask	2
2	ask	asked	3
2	ask	asking	4
2	ask	asks	5
3	back	back	6
3	back	backed	7
3	back	backing	8
3	back	backs	9
4	broker	broker	10
4	broker	brokered	11
4	broker	brokering	12
4	broker	brokers	13
5	business	business	14
5	business	businesses	15
6	busy	busied	16
6	busy	busier	17
6	busy	busies	18
6	busy	busiest	19
6	busy	busy	20
6	busy	busying	21
…	…	…	…

Set Up the Document-Term Matrix

Document Term Matrix (DTM) Setup

1. Compile *Term Frequency (TF)* of each lemma in the dictionary for each document (in matrix form)
   - Raw lemma counts are used for TF by document
2. Compute the *Inverse Document Frequency (IDF)* adjustment for each lemma
   - IDF weights "level the playing field"; prevents the most frequent terms from dominating / driving the SVD topics by default
3. Apply the IDF adjustments to the corresponding lemma columns in the matrix
   - Result will be a TF-IDF value in each cell of the matrix representing how important a lemma is to that document

Document Term Matrix (DTM) Setup

Initial DTM Matrix:
Term Frequency (TF)

	also W_1	*ask* W_2	*back* W_3	*broker* W_4	*business* W_5	*busy* W_6	*call* W_7	*can* W_8	*card* W_9	*check* W_{10}	*contact* W_{11}	*doesn't* W_{12}	*email* W_{13}	*fact* W_{14}	*find* W_{15}	*follow* W_{16}	*get* W_{17}	*give* W_{18}	*information* W_{19}	... ...	*year* W_{46}
D_1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	...	0
D_2	0	1	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	1	0	...	1
D_3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...	0
D_4	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	...	0
D_5	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	...	0
D_6	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	...	0
D_7	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	...	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
D_{19}	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	...	0
D_{20}	0	0	1	0	0	2	1	1	0	0	0	0	1	0	0	2	0	0	0	...	0
D_{21}	0	1	1	3	2	0	2	0	1	1	0	0	1	0	0	0	0	2	0	...	1
D_{22}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	...	0
D_{23}	1	0	0	0	2	0	0	0	1	1	1	0	0	0	0	0	2	0	0	...	1
D_{24}	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	1	0	1	2	...	0
D_{25}	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	...	0

Document Term Matrix (DTM) Setup

Compute IDF Weights for each Lemma

Lemma_ID	Lemma	m (total # documents)	ud_i (# unique documents with lemma)	IDF_Weight = LN(m / (0.5 + ud_i))
1	also	25	4	1.714798
2	ask	25	3	1.966113
3	back	25	3	1.966113
4	broker	25	3	1.966113
5	business	25	3	1.966113
6	busy	25	3	1.966113
7	call	25	9	0.967584
8	can	25	4	1.714798
9	card	25	5	1.514128
10	check	25	5	1.514128
11	contact	25	3	1.966113
12	doesn't	25	3	1.966113
13	email	25	6	1.347074
14	fact	25	4	1.714798
15	find	25	4	1.714798
16	follow	25	6	1.347074
17	get	25	6	1.347074
18	give	25	6	1.347074
19	information	25	3	1.966113
20	insurance	25	3	1.966113
21	interest	25	3	1.966113
...	...	...	...	...

Document Term Matrix (DTM) Setup

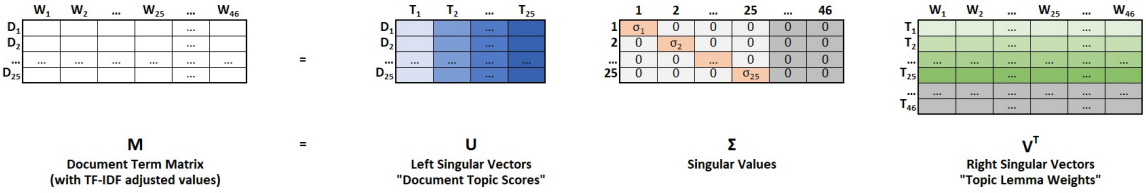
Final DTM Matrix:
Term Frequency Inverse Document Frequency
(TF-IDF)

	also W_1	*ask* W_2	*back* W_3	*broker* W_4	*business* W_5	*busy* W_6	*call* W_7	*can* W_8	*card* W_9	*check* W_{10}	*contact* W_{11}	*doesn't* W_{12}	*email* W_{13}	*fact* W_{14}	*find* W_{15}	*follow* W_{16}	*get* W_{17}	*give* W_{18}	*information* W_{19}	...	*year* W_{46}
D_1	0	0	0	0	0	0	0	0	0	0	0	0	0	1.7148	1.7148	0	0	0	0	...	0
D_2	0	1.9661	0	0	0	0	0.9676	1.7148	0	1.5141	0	0	0	0	0	0	0	1.3471	0	...	1.0788
D_3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...	0
D_4	0	0	0	0	0	0	0.9676	0	1.5141	0	0	0	0	0	0	0	0	0	0	...	0
D_5	0	0	0	0	0	0	0	0	0	0	0	0	1.3471	0	0	0	0	0	0	...	0
D_6	0	0	0	0	1.9661	1.9661	0	0	1.5141	0	0	0	0	0	0	0	0	0	0	...	0
D_7	1.7148	0	0	0	0	0	0	0	0	0	0	0	0	1.7148	1.7148	1.3471	0	0	0	...	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
D_{19}	0	0	0	0	0	0	0	0	0	0	0	0	0	1.7148	1.7148	0	0	0	1.9661	...	0
D_{20}	0	0	1.9661	0	0	3.9322	0.9676	1.7148	0	0	0	0	1.3471	0	0	2.6941	0	0	0	...	0
D_{21}	0	1.9661	1.9661	5.8983	3.9322	0	1.9352	0	1.5141	1.5141	0	0	1.3471	0	0	0	0	2.6941	0	...	1.0788
D_{22}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.3471	0	0	...	0
D_{23}	1.7148	0	0	0	3.9322	0	0	0	1.5141	1.5141	1.9661	0	0	0	0	0	2.6941	0	0	...	1.0788
D_{24}	0	0	0	0	0	0	0	0	0	0	3.9322	0	0	0	0	1.3471	0	1.3471	3.9322	...	0
D_{25}	0	0	0	0	0	0	0.9676	0	0	0	0	0	1.3471	0	0	0	0	1.3471	0	...	0

SVD Components

25

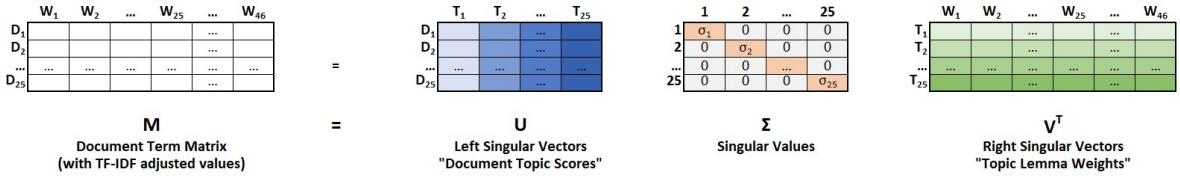
SVD Components



- Matrix "M" is the same DTM we just finished building in the prior slide
  - We have 25 document rows and 46 lemma columns in our DTM
- We can have at most min(25,46) = 25 singular values
  - And therefore we can also have at most 25 topics
- Grayed out sections of matrices will not affect the result so we could rewrite SVD without them and still get the same DTM

26

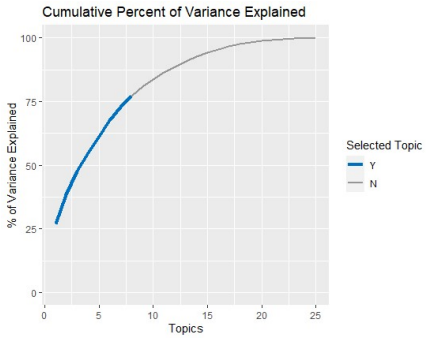
SVD Components



- The Left Singular Vectors show a "topic score" for each document
- The Right Singular Vectors show how the lemmas weigh into each topic
- The magnitude of the singular values shows how much of the variance between documents the topic (T_i) explains
  - The % Variance Explained for each topic is calculated as: $\sigma_i^2 / Sum_i(\sigma_i^2)$
  - Typically topics in the equation are ordered from largest singular value to smallest

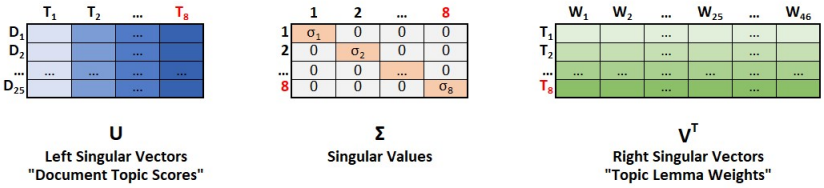
Percent of Variance Explained

topic_id	Singular Value	% Variance Explained	Cumulative % Variance Explained
1	14.430	26.9%	26.9%
2	9.460	11.6%	38.5%
3	8.373	9.1%	47.6%
4	7.516	7.3%	54.9%
5	7.076	6.5%	61.4%
6	6.910	6.2%	67.5%
7	6.469	5.4%	72.9%
8	5.649	4.1%	77.1%
9	5.336	3.7%	80.8%
10	4.799	3.0%	83.7%
...	...	...	...
25	0.861	0.1%	100.0%



- Judgmentally select how many topics to keep in analysis
  - Later topics have very little variance explained, and may amount to noise
- For this illustrative example, we will select 8 topics, which explains >75% of the variance in a full SVD

SVD Components



U
Left Singular Vectors
"Document Topic Scores"

Σ
Singular Values

V^T
Right Singular Vectors
"Topic Lemma Weights"

- These are the final SVD components after selecting 8 as our number of topics
  - The [absolute] values in the matrices will be the same as the full 25-topics decomposition, just with the less important topics removed from each matrix
    - Note these will no longer multiply out to the DTM because of the dropped topics
- The Singular Values and Topic Lemma Weights can be reused on new documents to score these same "top topics"

SVD Components: Results from Example

	T1	T2	T3	T4	T5	T6	T7	T8
D1	-0.0224	0.0886	0.0757	0.0900	-0.0965	0.0776	-0.0944	0.0343
D2	-0.1447	0.1841	-0.0605	-0.2989	0.0788	-0.0446	-0.0690	-0.0717
D3	-0.0062	0.0211	0.0089	-0.0078	-0.0403	-0.0013	-0.0159	-0.0025
D4	-0.0551	-0.0111	-0.0314	-0.0114	0.0066	0.0501	-0.0231	-0.0609
D5	-0.0428	-0.0058	0.0106	0.0407	-0.1168	-0.0327	-0.0266	-0.1074
...	...	...	...	...	...	...	...	...
D21	-0.7823	-0.5063	-0.1889	0.0120	-0.0824	0.1351	-0.0645	0.0716
D22	-0.0530	0.0612	0.0578	-0.0782	-0.0034	-0.1480	-0.0523	-0.1689
D23	-0.2748	0.2964	0.5235	0.1356	0.5735	0.3425	-0.1139	-0.0715
D24	-0.1064	0.1852	-0.0291	0.0962	-0.1063	0.2346	0.8033	-0.2238
D25	-0.0990	0.0535	0.0865	-0.0014	-0.0158	-0.1086	-0.0389	-0.1701

U
Left Singular Vectors
"Document Topic Scores"

	1	2	3	4	5	6	7	8
1	14.43	0	0	0	0	0	0	0
2	0	9.46	0	0	0	0	0	0
3	0	0	8.373	0	0	0	0	0
4	0	0	0	7.516	0	0	0	0
5	0	0	0	0	7.076	0	0	0
6	0	0	0	0	0	6.91	0	0
7	0	0	0	0	0	0	6.469	0
8	0	0	0	0	0	0	0	5.649

Σ
Singular Values

	also W1	ask W2	back W3	broker W4	business W5	...	open W32	phone W33	policy W34	quote W35	renewal W36	...	year W46
T1	-0.0557	-0.1431	-0.1387	-0.4212	-0.3067	...	-0.1301	-0.0482	-0.0807	-0.1658	-0.0685	...	-0.1556
T2	0.1793	-0.0333	-0.0008	-0.2596	-0.0975	...	-0.0264	0.1573	0.0914	0.3145	0.1944	...	0.1531
T3	0.0740	-0.1137	-0.2371	0.0426	0.1413	...	0.0406	0.0092	0.1284	0.4170	-0.1235	...	0.0458
T4	0.1400	-0.0968	0.1039	0.0079	0.1126	...	0.0248	-0.0465	-0.0680	0.1143	-0.4311	...	-0.1451
T5	0.0570	0.0593	0.0873	-0.0523	0.3235	...	-0.1778	-0.2127	-0.0855	0.1348	-0.0401	...	0.0498
T6	0.1325	-0.0122	-0.0550	-0.3172	0.3216	...	0.0142	0.0498	-0.3545	0.2096	0.1073	...	-0.0210
T7	-0.1190	-0.0541	-0.0389	0.1655	-0.1330	...	-0.0752	-0.0893	-0.1137	-0.2355	0.0133	...	-0.0523
T8	0.2758	-0.1301	-0.1292	0.3105	-0.0087	...	-0.1639	0.0494	-0.1682	-0.1703	0.0959	...	0.2062

V^T
Right Singular Vectors
"Topic Lemma Weights"

SVD Components: Results from Example

	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
D_1	-0.0224	0.0886	0.0757	0.0900	-0.0965	0.0776	-0.0944	0.0343
D_2	-0.1447	0.1841	-0.0605	-0.2989	0.0788	-0.0446	-0.0690	-0.0717
D_3	-0.0062	0.0211	0.0089	-0.0078	-0.0403	-0.0013	-0.0159	-0.0025
D_4	-0.0551	-0.0111	-0.0314	-0.0114	0.0066	0.0501	-0.0231	-0.0609
D_5	-0.0428	-0.0058	0.0106	0.0407	-0.1168	-0.0327	-0.0266	-0.1074
...	...	...	...	...	...	...	...	...
D_{21}	-0.7823	-0.5063	-0.1889	0.0120	-0.0824	0.1351	-0.0645	0.0716
D_{22}	-0.0530	0.0612	0.0578	-0.0782	-0.0034	-0.1480	-0.0523	-0.1689
D_{23}	-0.2748	0.2964	0.5235	0.1356	0.5735	0.3425	-0.1139	-0.0715
D_{24}	-0.1064	0.1852	-0.0291	0.0962	-0.1063	0.2346	0.8033	-0.2238
D_{25}	-0.0990	0.0535	0.0865	-0.0014	-0.0158	-0.1086	-0.0389	-0.1701

U
Left Singular Vectors
"Document Topic Scores"

	1	2	3	4	5	6	7	8
1	14.43	0	0	0	0	0	0	0
2	0	9.46	0	0	0	0	0	0
3	0	0	8.373	0	0	0	0	0
4	0	0	0	7.516	0	0	0	0
5	0	0	0	0	7.076	0	0	0
6	0	0	0	0	0	6.91	0	0
7	0	0	0	0	0	0	6.469	0
8	0	0	0	0	0	0	0	5.649

Σ
Singular Values

	also W_1	ask W_2	back W_3	broker W_4	business W_5	... ...	open W_{32}	phone W_{33}	policy W_{34}	quote W_{35}	renewal W_{36}	... ...	year W_{46}
T_1	-0.0557	-0.1431	-0.1387	-0.4212	-0.3067	...	-0.1301	-0.0482	-0.0807	-0.1658	-0.0685	...	-0.1556
T_2	0.1793	-0.0333	-0.0008	-0.2596	-0.0975	...	-0.0264	0.1573	0.0914	0.3145	0.1944	...	0.1531
T_3	0.0740	-0.1137	-0.2371	0.0426	0.1413	...	0.0406	0.0092	0.1284	0.4170	-0.1235	...	0.0458
T_4	0.1400	-0.0968	0.1039	0.0079	0.1126	...	0.0248	-0.0465	-0.0680	0.1143	-0.4311	...	-0.1451
T_5	0.0570	0.0593	0.0873	-0.0523	0.3235	...	-0.1778	-0.2127	-0.0855	0.1348	-0.0401	...	0.0498
T_6	0.1325	-0.0122	-0.0550	-0.3172	0.3216	...	0.0142	0.0498	-0.3545	0.2096	0.1073	...	-0.0210
T_7	-0.1190	-0.0541	-0.0389	0.1655	-0.1330	...	-0.0752	-0.0893	-0.1137	-0.2355	0.0133	...	-0.0523
T_8	0.2758	-0.1301	-0.1292	0.3105	-0.0087	...	-0.1639	0.0494	-0.1682	-0.1703	0.0959	...	0.2062

V^T
Right Singular Vectors
"Topic Lemma Weights"

SVD Components: Results from Example

	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
D_1	-0.0224	0.0886	0.0757	0.0900	-0.0965	0.0776	-0.0944	0.0343
D_2	-0.1447	0.1841	-0.0605	-0.2989	0.0788	-0.0446	-0.0690	-0.0717
D_3	-0.0062	0.0211	0.0089	-0.0078	-0.0403	-0.0013	-0.0159	-0.0025
D_4	-0.0551	-0.0111	-0.0314	-0.0114	0.0066	0.0501	-0.0231	-0.0609
D_5	-0.0428	-0.0058	0.0106	0.0407	-0.1168	-0.0327	-0.0266	-0.1074
...	...	...	...	...	...	...	...	...
D_{21}	-0.7823	-0.5063	-0.1889	0.0120	-0.0824	0.1351	-0.0645	0.0716
D_{22}	-0.0530	0.0612	0.0578	-0.0782	-0.0034	-0.1480	-0.0523	-0.1689
D_{23}	-0.2748	0.2964	0.5235	0.1356	0.5735	0.3425	-0.1139	-0.0715
D_{24}	-0.1064	0.1852	-0.0291	0.0962	-0.1063	0.2346	0.8033	-0.2238
D_{25}	-0.0990	0.0535	0.0865	-0.0014	-0.0158	-0.1086	-0.0389	-0.1701

U
Left Singular Vectors
"Document Topic Scores"

	1	2	3	4	5	6	7	8
1	14.43	0	0	0	0	0	0	0
2	0	9.46	0	0	0	0	0	0
3	0	0	8.373	0	0	0	0	0
4	0	0	0	7.516	0	0	0	0
5	0	0	0	0	7.076	0	0	0
6	0	0	0	0	0	6.91	0	0
7	0	0	0	0	0	0	6.469	0
8	0	0	0	0	0	0	0	5.649

Σ
Singular Values

Ross said he would speak with Jordan about quote this year.

Met owner Ross. We got a tour of the shop. Jordan was out on the road. And apparently the daughter also has a hand in the business. We got Ross's business card, contact him as soon as possible.

Checking with UW about 2018 denial of quote due to roof issue and ponding water evidence before I offer to quote. Owner was receptive to quote during visit.

	also W_1	ask W_2	back W_3	broker W_4	business W_5	... ...	open W_{32}	phone W_{33}	policy W_{34}	quote W_{35}	renewal W_{36}	... ...	year W_{46}
T_1	-0.0557	-0.1431	-0.1387	-0.4212	-0.3067	...	-0.1301	-0.0482	-0.0807	-0.1658	-0.0685	...	-0.1556
T_2	0.1793	-0.0333	-0.0008	-0.2596	-0.0975	...	-0.0264	0.1573	0.0914	0.3145	0.1944	...	0.1531
T_3	0.0740	-0.1137	-0.2371	0.0426	0.1413	...	0.0406	0.0092	0.1284	0.4170	-0.1235	...	0.0458
T_4	0.1400	-0.0968	0.1039	0.0079	0.1126	...	0.0248	-0.0465	-0.0680	0.1143	-0.4311	...	-0.1451
T_5	0.0570	0.0593	0.0873	-0.0523	0.3235	...	-0.1778	-0.2127	-0.0855	0.1348	-0.0401	...	0.0498
T_6	0.1325	-0.0122	-0.0550	-0.3172	0.3216	...	0.0142	0.0498	-0.3545	0.2096	0.1073	...	-0.0210
T_7	-0.1190	-0.0541	-0.0389	0.1655	-0.1330	...	-0.0752	-0.0893	-0.1137	-0.2355	0.0133	...	-0.0523
T_8	0.2758	-0.1301	-0.1292	0.3105	-0.0087	...	-0.1639	0.0494	-0.1682	-0.1703	0.0959	...	0.2062

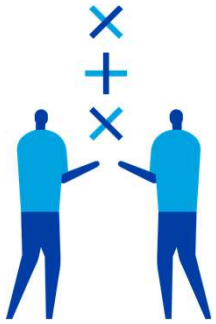
V^T
Right Singular Vectors
"Topic Lemma Weights"

Apply SVD Results to New Documents

33

Apply SVD Results to New Documents

1. Apply the text preprocessing to the new documents
2. Create the DTM for the new documents (using dictionary that was built)
3. Apply the IDF adjustments (from SVD analysis) to the appropriate lemma columns in the DTM
4. Using the *Singular Values* matrix and the *Topic Lemma Weights* matrix (from SVD analysis), compute the Document Topic Scores (U) as follows:



$$\mathbf{U = M\ V\ \Sigma^{-1}}$$

(Note: this calculation is derived from a simple rearrangement of the SVD equation)

34

Apply SVD Results to New Documents

Step 1: Activity Preprocessing

Opp_Index	Actv_Index	Type	Description	Mod_Description
26	31	Phonecall	Drew said that he does want to reschedule FF but says he wants to get owner Bob in meeting too... said call back in afternoon as he will have seen Dan by then.	said that he does want to reschedule fact find but says he wants to get owner in meeting too said call back in afternoon as he will have seen by then
26	32	Phonecall	VM Drew cell (sent email too) - follow up, still wanting to re-set FF appointment?	voicemail sent email too follow up still wanting to reset fact find appointment
27	33	Visit	I spoke with the receptionist and this looks like an excellent account for commercial insurance for us. She informed me that their x-date is September 15, need to update and check back with Harry next year.	i spoke with the receptionist and this looks like an excellent account for insurance for us she informed me that their xdate is need to update and check back with next year
28	34	Visit	Carl was busy and couldn't meet. Dropped off business card and got his. Will update information and send over email / call.	was busy and couldn't meet dropped off business card and got his will update information and send over email call

Step 2 : DTM

Initial DTM Matrix:
Term Frequency (TF)

	also W_1	ask W_2	back W_3	broker W_4	business W_5	... ...	stop W_{42}	talk W_{43}	want W_{44}	wc W_{45}	year W_{46}
D_{26}	0	0	1	0	0	...	0	0	3	0	0
D_{27}	0	0	1	0	0	...	0	0	0	0	1
D_{28}	0	0	0	0	1	...	0	0	0	0	0

Step 3 : IDF Adjustment

Final DTM Matrix:
Term Frequency Inverse Document Frequency (TF-IDF)

	also W_1	ask W_2	back W_3	broker W_4	business W_5	... ...	stop W_{42}	talk W_{43}	want W_{44}	wc W_{45}	year W_{46}
D_{26}	0	0	1.9661	0	0	...	0	0	5.1444	0	0
D_{27}	0	0	1.9661	0	0	...	0	0	0	0	1.0788
D_{28}	0	0	0	0	1.9661	...	0	0	0	0	0

Step 4 : Compute Topic Scores

$$U = M\ V\ \Sigma^{-1}$$

Document Topic Scores

	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
D_{26}	-0.1511	0.2362	0.0365	0.2266	-0.4084	-0.0640	-0.1555	-0.2119
D_{27}	-0.1197	0.1041	-0.0979	-0.1026	0.0742	-0.0896	0.0860	0.0841
D_{28}	-0.1318	0.0644	0.0140	0.1455	0.1972	0.0348	0.1270	-0.0866

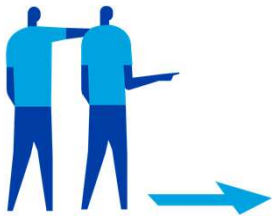
Use SVD Topic Variables in a Predictive Model

Use SVD Topic Variables in a Predictive Model

There are two main routes you can take:

1. Use **Topic Score variables** directly

- Allows the capture of complex multivariate relationships between the individual SVD topics and the non-SVD variables
- Ideal for **machine learning models**

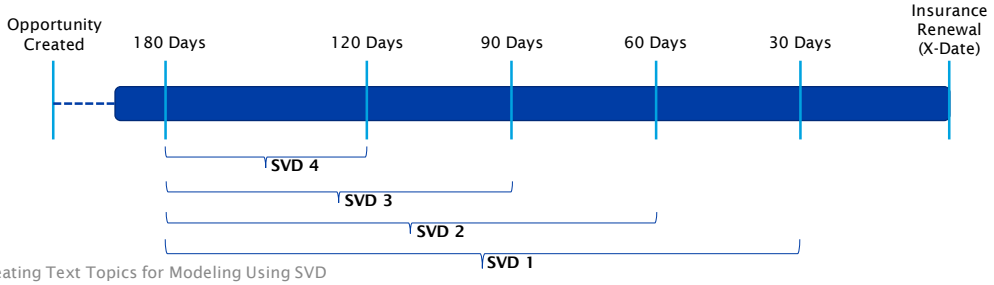


2. Create a **"SVD Score" sub-model**

- Consolidates, into a single variable, the freeform text's positive/negative impact on the prediction target
- Ideal for models that prioritize intuition and simplicity in an effort to generalize well, like manually-constructed **GLMs**
  - *e.g. Much easier to say "the sales discussions to date on this prospect indicate it's more likely to sell" than explain that "large positive/negative scores on topic t driven by words x, y, z indicate the prospect is more likely to sell", as the latter explanation is complex and will be different for everyone.*

SVD for the Direct Writer Prospect Model

- Dictionary was ~3500 Lemmas
  - Top 9% of lemmatized tokens (excluding non-negation stopwords)
    - The top 9% was performed on 3 different activity type groups (email, visit/phonecall/appt, other), then consolidated into a unique list
- Kept 25 topics for each SVD
- Ran a separate SVD for each point in the sales process that we score prospects:



SVD for the Direct Writer Prospect Model

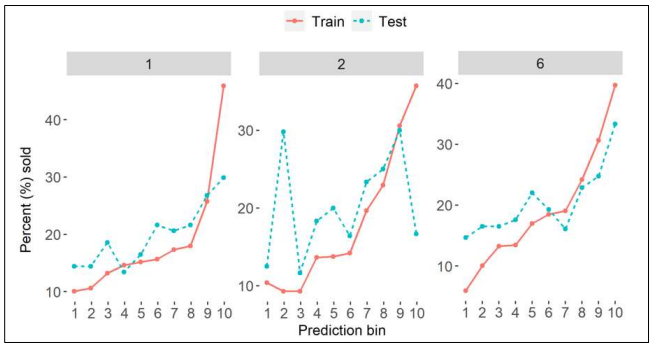
- Final scoring (for each time period's model) was an ensemble of several Machine Learning models and a GLM model
  - Used the individual **topic score variables** in the **ML models**
  - Created **SVD Score sub-model** to use in the **GLM**
- End-user scores seen in our Customer Relationship Management system are given on a simple 1-5 scale (1=Least likely to sell, 5=Most likely to sell)

Creating an "SVD Score" Sub-Model

1. Add the Document Topic Scores (25 variables) to a modeling dataset of quoted opportunities
   - Target variable was 1=Sold, 0=Not Sold
   - 25% of quoted opportunities were held out in a Test set for model validation
2. Fit a GBM (gradient-boosting machine) model to the Train modeling dataset
   - Hyperparameters were selected via grid search (n.trees, interaction.depth, n.minobsinnode, shrinkage, bag.fraction)
   - 5-fold cross validation (CV)
   - Selection of best hyperparameters:
     - Primarily based on highest average AUC (area under receiver operating characteristic curve) on the CV Test folds
     - Additional rule to exclude hyperparameter sets with too much overfitting [set a max threshold for (CV Train AUC - 0.5)/(CV Test AUC – 0.5) ratio]

Creating an "SVD Score" Sub-Model

3. Validate GBM model results



4. Output GBM model predictions ("*SVD Score*") for each modeling record

Thank you.

Resources

43

Resources

Contact Info:

alan.johnson@sentry.com

Illustrative Example data and R code are available along with this Powerpoint on the RPM website.

Requires R version 4.0.0+ to install proper packages and run.

44