Content Analysis Dictionaries 2

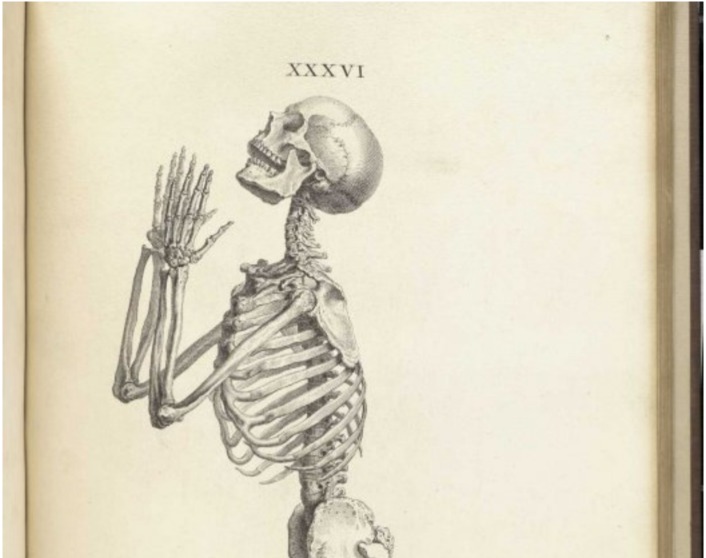
William Lowe

Hertie School

28th September 2020

Solutions: Some theological approaches

Solutions: Some theological approaches



Solutions: Do not sin in the first place

Solutions: Do not sin in the first place

An often non-obvious fact about content dictionaries:

→ *Recall*: the proportion of words used that way that are in your dictionary

→ *Precision*: the proportion of words used the way your dictionary assumes they are used

Sins

Every field reinvents this distinction:

- → precision and recall
- → specificity and sensitivity
- → users and producer's accuracy
- → type 1 and type 2 error
- → sins of omission and sins of commission

Precision

Keyword in context analyses (KWIC) allow you to scan all contexts of a word

→ How many of them *are* the sense or usage you want?

Let's take a look at `benefit*` as a 'pro government intervention in the economy' word

	pre	keyword	post
1	also keep all the other	benefits	that pensioners currently receive ,
2	regulation will have to have	benefits	exceeding costs , and regulations
3	and Controlled Immigration Britain has	benefited	from immigration . We all
4	positive contribution But if those	benefits	are to continue to flow
5	Northern Ireland brings	benefits	to all parts of our
6	their home , will also	benefit	first-time buyers . Empowering individuals
7	you help yourself ; you	benefit	and the country benefits .
8	you benefit and the country	benefits	. So now , I
9	result of our tax and	benefit	measures compared to 1997 .
10	result of personal tax and	benefit	measures introduced since 1997 ,
11	, the savings on unemployment	benefits	will go towards investing more
12	trebled the number on incapacity	benefits	. We will help 17
13	Work programme and reform Incapacity	Benefit	, with the main elements
14	main elements of the new	benefit	regime in place from 2008
15	stronger penalties . To the	benefit	of business and household consumers
16	effective directive to provide real	benefits	to consumers and new opportunities
17	better.We are examining the potential	benefits	of a parallel Expressway on
18	ways to lock in the	benefit	of new capacity . We
19	are determined to spread the	benefits	of enterprise to every community
20	to get ahead , to	benefit	from improving public services

Precision

Of the 20 instances, these are (arguably)

- → 6 used the way we expect from the topic
- → 3 used in the opposite sense: anti-government intervention in the economy
- → 11 used in ways that are neither

So... 0.3 correct, 0.15 mistaken, and 0.55 unrelated 'noise'

- → Perhaps not an amazing choice

There are two kinds of precision failures here with different consequences

- → Mistaking an *topic-unrelated* word for this topic (11 of these)
- → Mistaking a word used in the sense of a *different* topic for this one (3 of these)

The first mistake does not really harm precision, but the second does

Recall

Bad recall is a mixture of two problems

- Assigning words to a topic that are mostly used for a different one: $P(W \mid Z = k) < P(W \mid Z = j)$ but we assigned it to k anyway
- Failing to assign a topic-informative word to any topic: Dictionary says $P(W \mid Z = k) = 0$, but it's not. This is about *coverage*

Let's consider coverage first

Coverage

One possible checking procedure:

- → Take a random matched sample of words not in the dictionary but present in the corpus e.g. match each dictionary word to another of the same frequency
- → Examine their KWICs to see if they should have been assigned to a topic

If we were feeling even more energetic

- → Assign them their most likely topic manually
- → Compare this $\tilde{\theta}$ to the dictionary's own estimate θ
- → These should not be wildly different

Recall

The other kind of mistake is difficult because the natural procedure is

- → Assign *every word* (or at least every instance of a word that the dictionary knows about in a document) to a topic
- → For each topic, see what proportion of times the dictionary agrees it is in that topic

This also promises to be very tiring.

Using precision to estimate recall

However, two facts may help us:

- We have a tireless computer available
- Recall and precision relate $P(W \mid Z)$ and $P(Z \mid W)$ respectively
- …and we know how

For clarity, let's call the true topic of a word Z as before, and the dictionary's idea of the topic of a word $\hat{Z}$ because it's kind of an estimate of that. So,

$$\text{Recall:} \quad \sum_{k}^{K} P(\hat{Z} = k \mid Z = k)$$

$$\text{Precision:} \quad \sum_{k}^{K} P(Z = k \mid \hat{Z} = k)$$

Using precision to estimate recall

If we have a sense of dictionary topic precision, could we use it to get a sense of dictionary topic recall? Why yes, by borrowing methods from King and Lowe (2003)

Using precision to estimate recall

If we have a sense of dictionary topic precision, could we use it to get a sense of dictionary topic recall? Why yes, by borrowing methods from King and Lowe (2003)

According to the Rev. Bayes

$$
\begin{aligned}
P(\hat{Z} = k \mid Z = k) &= \frac{P(Z = k \mid \hat{Z} = k)P(\hat{Z} = k)}{\sum_j^K P(Z = j \mid \hat{Z} = j)P(\hat{Z} = j)} \\
&= P(Z = k \mid \hat{Z} = k)\frac{P(\hat{Z} = k)}{P(Z = k)} \qquad \text{(recall is reweighted precision)} \\
&\propto P(Z = k \mid \hat{Z} = k)P(\hat{Z} = k)
\end{aligned}
$$

Conveniently

→ we don't need the denominator because it only ensure the recall measures add to one

→ We can get $P(\hat{Z} = k)$ by running the dictionary over the entire corpus

Using precision to estimate recall

Confession and forgiveness

Under measurement error

- → A observed category proportions are generated by a *mixture* of categories
- → The weights for this mixture are the true category proportions $P(Z = k) = \theta$

$$P(W) = \sum_{k}^{K} P(W \mid Z = k) P(Z = k)$$

Two applications of the mixture: 1. Naive Bayes

$$P(W) = \sum_{k}^{K} P(W \mid Z = k) P(Z = k)$$

If we we ever observed Z we could learn a lot about $P(W \mid Z = k)$.

Two applications of the mixture: 1. Naive Bayes

$$P(W) = \sum_{k}^{K} P(W \mid Z = k)P(Z = k)$$

If we we ever observed Z we could learn a lot about $P(W \mid Z = k)$.

At the word level we typically do not have access to Zs.

Two applications of the mixture: 1. Naive Bayes

But if we are willing to assume that all the words in a document have the *same* value of Z, then we only need judgments about document content to learn about $P(W \mid Z = k)$.

And if

$$P(W \mid Z = k) = \prod P(W_j \mid Z = k)$$

then we are classifying documents using *Naive Bayes* (Evans et al. 2007)

Two applications of the mixture: 2. Correction

But back to words and their categories…

$$P(W) = \sum_{k}^{K} P(W \mid Z = k)P(Z = k)$$

If we knew about the error process $P(W \mid Z = k)$, we could *back out* the true proportions

A general purpose linear approach

The category proportions are

$$P(W) = \sum_{k}^{K} P(W \mid Z = k) P(Z = k)$$

has the form

$$P = E\theta$$

where P are the coded proportions and E is the $V \times K$ coder error matrix, so

$$\theta = \mathrm{E}^{-1} P$$

Typically we don't exactly know E so the result is an approximation

An application to human sentence coders

Application to Mikhaylov et al.'s subjects coding New Zealand party manifestos.

	true	L	N	R
code	L	430	188	100
	N	254	712	193
	R	41	115	650

Convert to error probabilities

Errors: E

	L	N	R
L	0.59	0.19	0.11
N	0.35	0.70	0.20
R	0.06	0.11	0.69

Convert to error probabilities

Errors: E

	L	N	R
L	0.59	0.19	0.11
N	0.35	0.70	0.20
R	0.06	0.11	0.69

Inverted: E^{-1}

	L	N	R
L	2.00	-0.50	-0.16
N	-1.00	1.75	-0.37
R	0.00	-0.25	1.52

Convert to error probabilities

Errors: E

	L	N	R
L	0.59	0.19	0.11
N	0.35	0.70	0.20
R	0.06	0.11	0.69

Inverted: E^{-1}

	L	N	R
L	2.00	-0.50	-0.16
N	-1.00	1.75	-0.37
R	0.00	-0.25	1.52

$\theta = 0.27, 0.38, 0.35$

Convert to error probabilities

Errors: E

	L	N	R
L	0.59	0.19	0.11
N	0.35	0.70	0.20
R	0.06	0.11	0.69

Inverted: E^{-1}

	L	N	R
L	2.00	-0.50	-0.16
N	-1.00	1.75	-0.37
R	0.00	-0.25	1.52

$\theta = 0.27, 0.38, 0.35$

Implications for derived measurements

If [L, N, R] were [20, 0, 10]

→ true position: -0.33 on our previous left-right scale, ignoring N

Under measurement error we would *expect* to see about [13, 9, 8]

→ an attenuated -0.24 on our previous scale

Correcting this with the correct error matrix would recover the right proportions, and so the right position

Implementations and limitations

Hopkins and King (2010) and King and Lu (2008) implement this strategy

Implementations and limitations

Hopkins and King (2010) and King and Lu (2008) implement this strategy

When coder errors are only *estimated* the result hold in expectation.

The tradeoff is between

- → errors coders make
- → errors we make estimating the errors coders make…

Implementations and limitations

Linearity means we sometimes estimate proportions outside [0,1]

Alternatively we can assign distributions and work with the original structure

$$P(W) = \sum_{k}^{K} P(W \mid Z = k) P(Z = k)$$

and that is exactly what topic models do.

References

King, G. & Lowe, W. (2003). 'An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design'. *International Organization*, *57*(3), 617–642.