Comparing corpora

Wouter van Atteveldt

June 1, 2016

Comparing corpora

Another useful thing we can do is comparing two corpora: Which words or names are mentioned more in e.g. Bush' speeches than Obama's.

This uses functions from the corpustoools package, which you can install directly from github: (you only need to do this once per computer)

```
install.packages("devtools")
devtools::install_github("kasperwelbers/corpus-tools")
```

For this handout, we will use the State of the Union speeches contained in the `corpustools` package, and create a document term matrix (DTM) from all names and nouns in the speeches by Bush and Obama:

```
library(corpustools)
data(sotu)
dtm = with(subset(sotu.tokens, pos1 %in% c("M", "N")),
           dtm.create(documents=aid, terms=lemma))
```

Now, we can create separate DTMs for Bush and Obama, relying on the headline column in the metadata:

To do this, we split the dtm in separate dtm's for Bush and Obama. For this, we select docment ids using the `headline` column in the metadata from `sotu.meta`, and then use the `dtm.filter` function:

```
head(sotu.meta)
```

id	medium	headline	date
111541965	Speeches	Barack Obama	2013-02-12
111541995	Speeches	Barack Obama	2013-02-12
111542001	Speeches	Barack Obama	2013-02-12
111542006	Speeches	Barack Obama	2013-02-12
111542013	Speeches	Barack Obama	2013-02-12
111542018	Speeches	Barack Obama	2013-02-12

```
obama.docs = sotu.meta$id[sotu.meta$headline == "Barack Obama"]
dtm.obama = dtm.filter(dtm, documents=obama.docs)
bush.docs = sotu.meta$id[sotu.meta$headline == "George W. Bush"]
dtm.bush = dtm.filter(dtm, documents=bush.docs)
```

So how can we check which words are more frequent in Bush' speeches than in Obama's speeches? The function `corpora.compare` provides this functionality, given two document-term matrices:

```
cmp = corpora.compare(dtm.obama, dtm.bush)
cmp = cmp[order(cmp$over), ]
```

```
head(cmp)
```

	term	termfreq.x	termfreq.y	termfreq	relfreq.x	relfreq.y	over	chi
666	terror	1	55	56	0.0001141	0.0058918	0.1616600	48.90822
286	freedom	8	79	87	0.0009130	0.0084628	0.2021641	53.84623
230	enemy	4	52	56	0.0004565	0.0055704	0.2216774	38.31624
668	terrorist	10	73	83	0.0011413	0.0078200	0.2427760	44.15959
366	Iraq	15	94	109	0.0017119	0.0100696	0.2449890	52.73362
794	Saddam	0	26	26	0.0000000	0.0027852	0.2641857	24.43918

For each term, this data frame contains the frequency in the 'x' and 'y' corpora (here, Obama and Bush). Also, it gives the relative frequency in these corpora (normalizing for total corpus size) and the overrepresentation in the 'x' corpus and the chi-squared value for that overrepresentation. So, Bush used the word terrorist 105 times, while Obama used it only 13 times, and in relative terms Bush used it about four times as often, which is highly significant.

Which words did Obama use most compared to Bush?

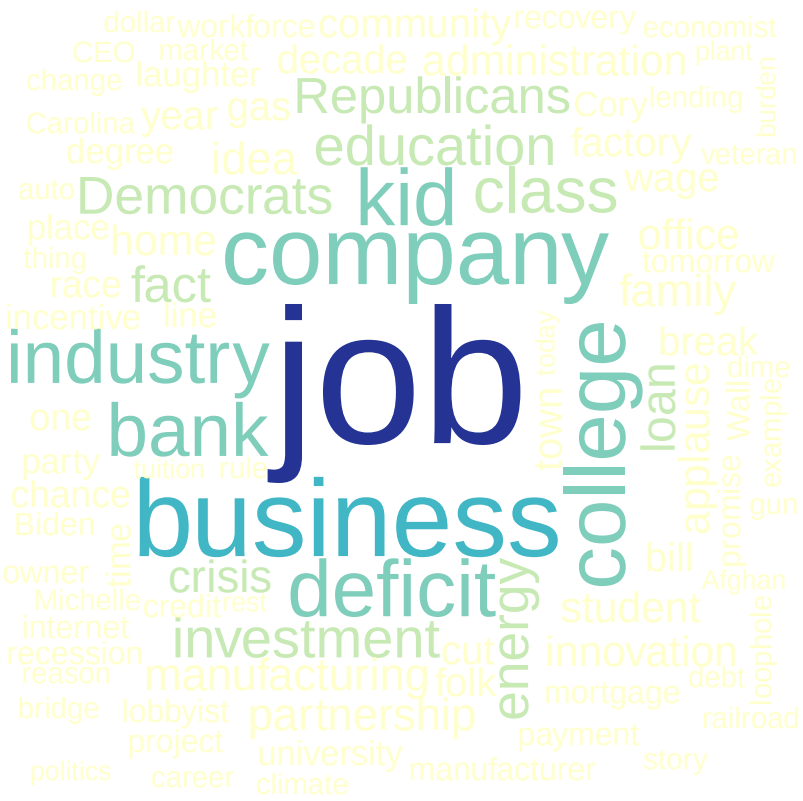
```
cmp = cmp[order(cmp$over, decreasing=T), ]
head(cmp)
```

	term	termfreq.x	termfreq.y	termfreq	relfreq.x	relfreq.y	over	chi
377	kid	31	0	31	0.0035380	0.0000000	4.538005	33.08395
130	company	54	6	60	0.0061630	0.0006427	4.360377	41.67961
52	bank	29	0	29	0.0033097	0.0000000	4.309747	30.94608
350	industry	32	1	33	0.0036521	0.0001071	4.202000	31.20551
124	college	55	9	64	0.0062771	0.0009641	3.705033	36.20383
117	class	26	1	27	0.0029674	0.0001071	3.583483	24.82023

So, while Bush talks about freedom, war, and terror, Obama talks more about industry, banks and education.

Let's make a word cloud of Obama' words, with size indicating chi-square overrepresentation:

```
obama = cmp[cmp$over > 1,]
dtm.wordcloud(terms = obama$term, freqs = obama$chi)
```



And the same for bush: (output not shown)

```
bush = cmp[cmp$over < 1,]
dtm.wordcloud(terms = bush$term, freqs = bush$chi)
```

Note that the warnings given by these commands are relatively harmless: it means that some words are skipped because it couldn't find a good place for them in the word cloud.

Contrast plots

Finally, we can use `plotWords` in the `corpustools` package to make a wordcloud-style plot with Obama's word on the right and Bush' words on the left:

```
with(arrange(cmp, -chi)[1:100, ],
plotWords(x=log(over), words = term, wordfreq = chi, random.y = T))
```