Text Analysis: Supervised and Unsupervised Models

Created	@June 26, 2023 3:45 PM
Tags	Personal Project
Project Type	Text Analysis

About the Project

In this project, I explore supervised and unsupervised models in text analysis. I used built-in US Presidential Inaugural Address Texts dataset from R's quanteda package. Firstly, I used supervised model, using naive bayes classifier, to see whether the model can identify "pre-war" texts. Secondly, I used unsupervised model to see whether the computer could come up with 'topics' and classify the documents based on these topics. This is called topic modelling.

Retrieving the Data

As previously said, I have used a built-in dataset from the quanteda package, called `data_corpus_inaugural`.

R Markdown

RPubs - Text Analysis: Supervised and Unsupervised Models
https://rpubs.com/annisaptr/supervised_unsupervised_models

Analysis

Supervised Models

```
#docvars shows the variables in the dataset
docvars(dtm)
```

Firstly, I used the docvars() function to look at the variables in the dataset. The table below illustrates the first 10 rows of the dataset.

Year	President	First Name	Party
1789	Washington	George	none
1793	Washington	George	none
1797	Adams	John	Federalist
1801	Jefferson	Thomas	Democratic-Republican
1805	Jefferson	Thomas	Democratic-Republican
1809	Madison	James	Democratic-Republican
1813	Madison	James	Democratic-Republican
1817	Monroe	James	Democratic-Republican

1821	Monroe	James	Democratic-Republican
1825	Adams	John Quincy	Democratic-Republican

I then added a new variable titled “pre-war” to indicate whether the document was made before or after the war (i.e. 1945), with TRUE indicating the document was made before the war and FALSE indicating it was made after the war. This then results in the following table, again, showing the first 10 rows.

```
#Adding a new variable, "is_prewar"
docvars(dtm, "is_prewar") ← docvars(dtm, "Year") < 1945
docvars(dtm)
```

Year	President	First Name	Party	Is Pre-war
1789	Washington	George	none	TRUE
1793	Washington	George	none	TRUE
1797	Adams	John	Federalist	TRUE
1801	Jefferson	Thomas	Democratic-Republican	TRUE
1805	Jefferson	Thomas	Democratic-Republican	TRUE
1809	Madison	James	Democratic-Republican	TRUE
1813	Madison	James	Democratic-Republican	TRUE
1817	Monroe	James	Democratic-Republican	TRUE
1821	Monroe	James	Democratic-Republican	TRUE
1825	Adams	John Quincy	Democratic-Republican	TRUE

By creating a training and testing document matrix, and using the naive-bayes classifier text model to classify the data, we get the following prediction:

Year	President Surname	Is Pre-war
1793	Washington	TRUE
1805	Jefferson	TRUE
1829	Jackson	TRUE
1837	VanBuren	TRUE
1849	Taylor	TRUE
1853	Pierce	TRUE
1861	Lincoln	TRUE
1873	Grant	TRUE
1881	Garfield	TRUE
1889	Harrison	TRUE
1897	McKinley	TRUE
1909	Taft	TRUE
1949	Truman	TRUE
1961	Kennedy	FALSE
1989	Bush	FALSE
1993	Clinton	FALSE
1997	Clinton	FALSE
2009	Obama	FALSE
2017	Trump	FALSE

From the table above, we can see that the model predicted correctly for all document except for one. We can also derive this conclusion using a cross tabulation:

Prediction	FALSE	TRUE
FALSE	6	0
TRUE	1	12

From here, we can analyse the following: There is only one false prediction, out of the times that we predict the speech to be pre-war, 11 out of the 12 were predicted correctly. This gives a precision of 91.67%. And out of the times that we predict the speech to not be pre-war, it was all predicted correctly.

Unsupervised Models

This type of text analysis trains a model that does something useful with texts without requiring annotated texts as training data. An example of unsupervised models include topic modelling, whereby the computer uses the dataset we give it, and independently come up with 'topics' and classify the documents accordingly. I will be using the same dataset, however with a focus on the speech texts. I have used the topicmodels package.

After conducting pre-processing, such as removing stopwords, words that are rarely used, and fixing punctuation, I converted the document term matrix in a format that is readable by the topicmodels package.

```
#converting our document term matrix (par_dtm) so it can be used with topicmodels
functionpar_dtm ← convert(par_dtm, to = "topicmodels")
```

I then conducted the unsupervised modelling using the following functions:

```
set.seed(3)lda_model ← topicmodels::LDA(par_dtm, method = "Gibbs", k = 10)
#k indicates the number of topics, Gibbs is the method parameters
terms(lda_model,5)
```

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
may	us	peopl	world	time	shall	law	nation
can	must	nation	new	now	duti	peopl	peac
countri	can	upon	american	year	citizen	govern	war
everi	let	free	america	hope	offic	upon	great
never	work	everi	freedom	one	cohfid	execut	justic

As can be seen, the correlation of words within topics are not very clear - this means that we will need a bigger dataset in order to have a more clearly defined topic models.

Conclusion

In this project, I explored supervised and unsupervised models in text analysis using the US Presidential Inaugural Address Texts dataset. In the supervised model, I used a naive bayes classifier to predict whether a speech was made before or after the war with a high level of accuracy. In the unsupervised model, I used the topicmodels package to conduct topic modelling on the speeches. However, due to the limited number of speeches in the dataset, the correlation of words within topics was not very clear. Overall, this project demonstrates the potential of text analysis in uncovering insights from large datasets, but also highlights the importance of having sufficient data for reliable analysis.