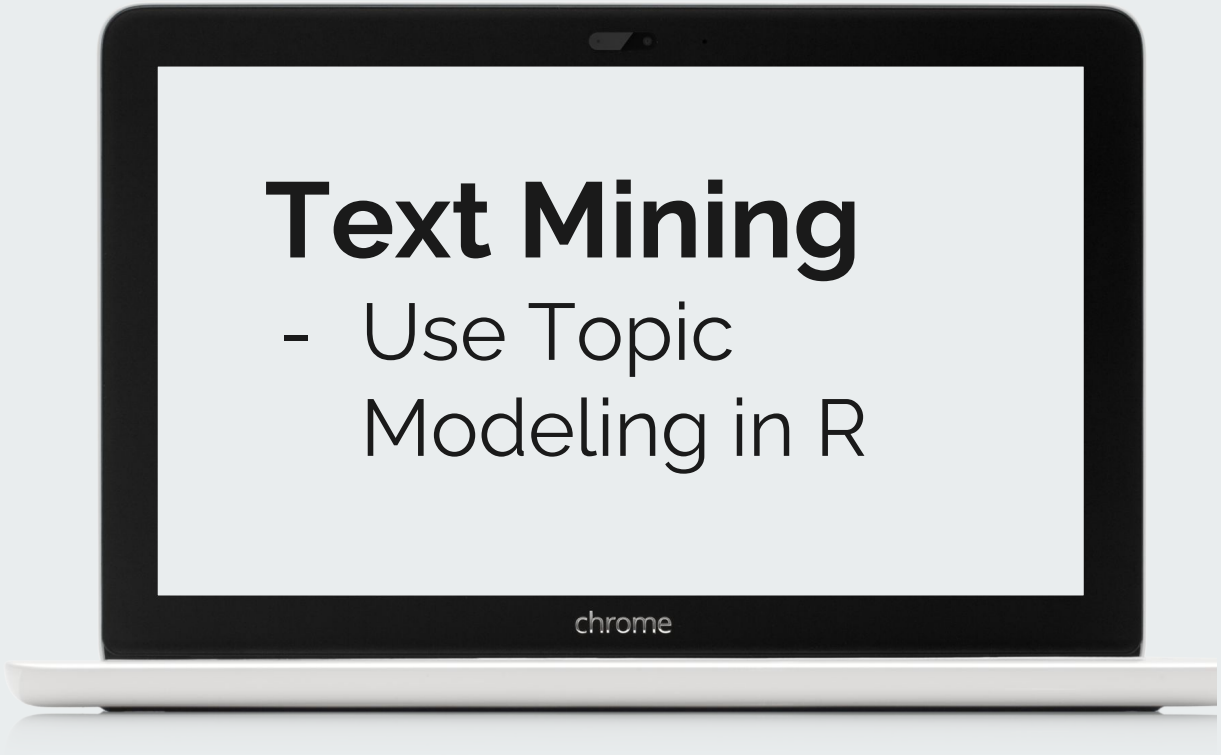

Text Mining
- Use Topic Modeling in R
chrome

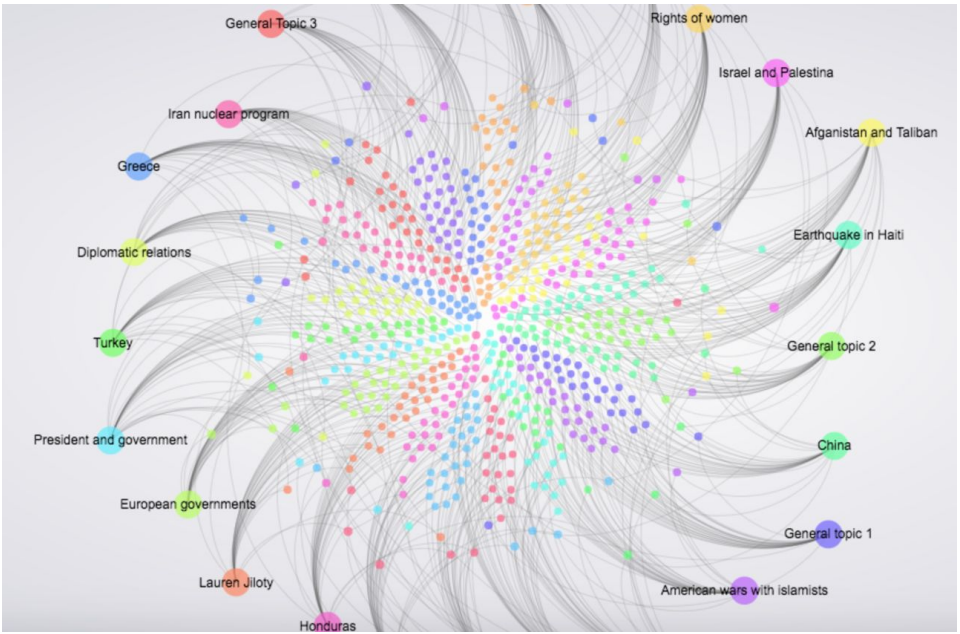
What is a topic model?

- A statistic model for discovering abstract topics in a collection of documents
- Based on machine learning and natural language processing

What is a topic model?

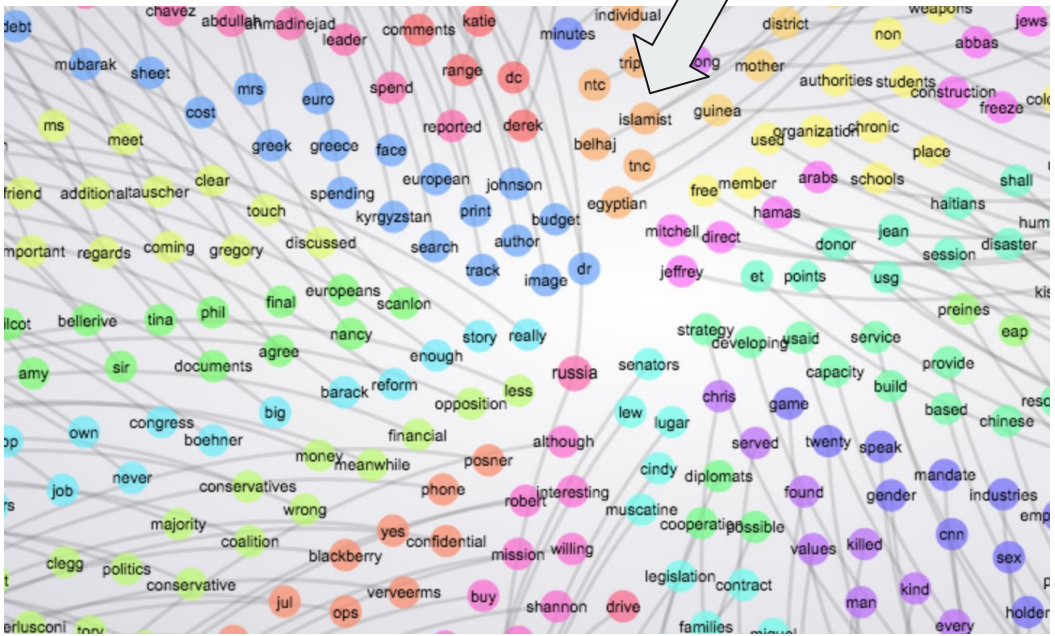
➔ A topic model based on Hillary Clinton's leaked emails:

http://mellain.github.io./topic-term.html



What is a topic model?

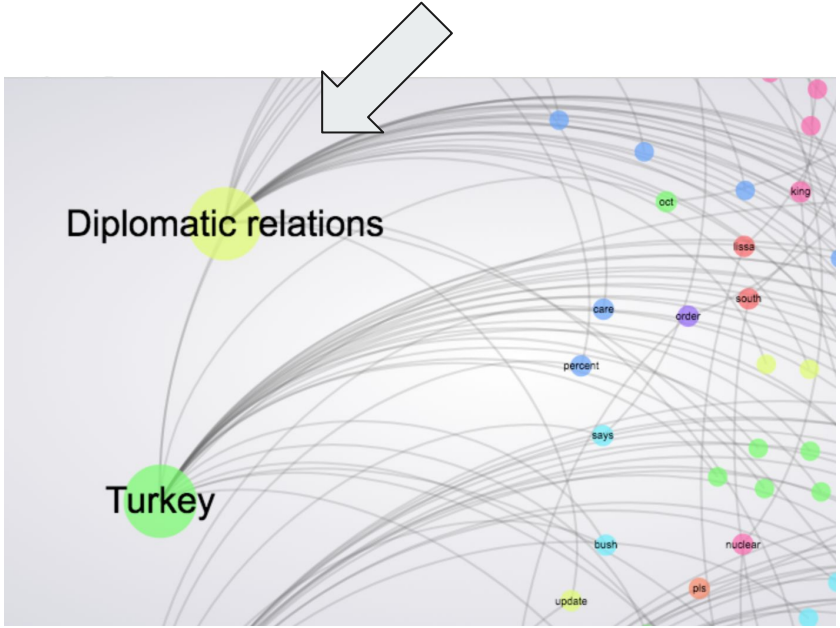
- Based on semantic patterns in text
- *Assumption*: Words belonging to a similar topic tend to co-occur in the same document.



What is a topic model?

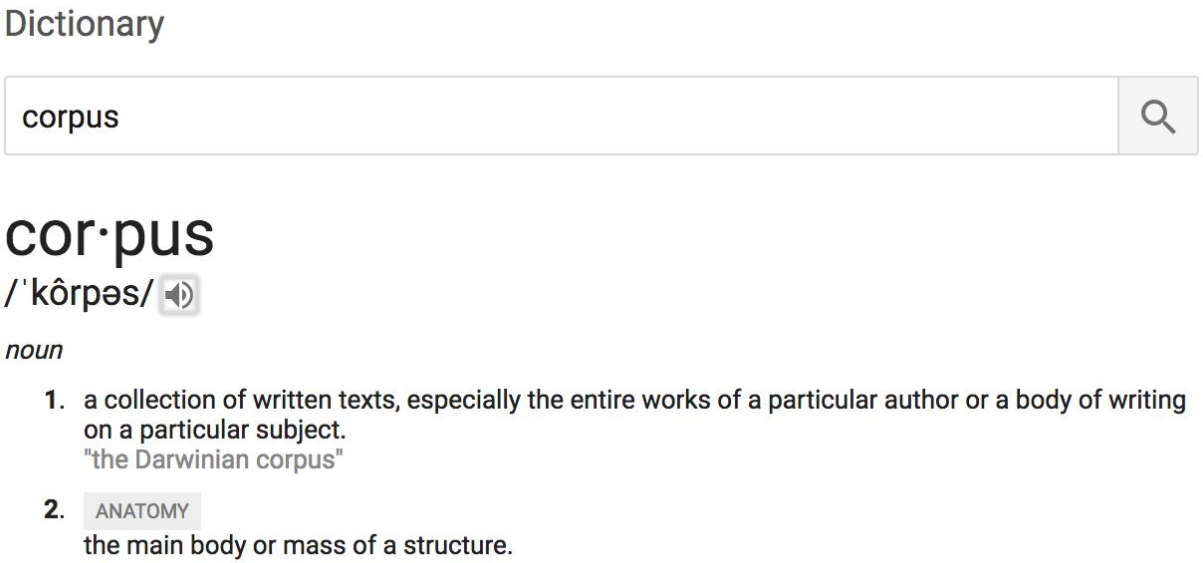
Topics

- Based on semantic patterns in text
- *Assumption*: Words belonging to a similar topic tend to co-occur in the same document.



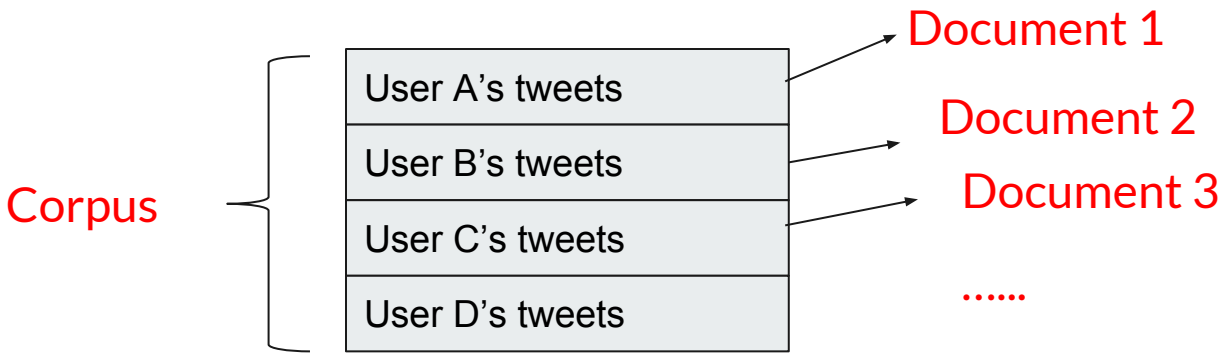
Mind the jargons!

- ➔ Corpus
- ➔ Documents



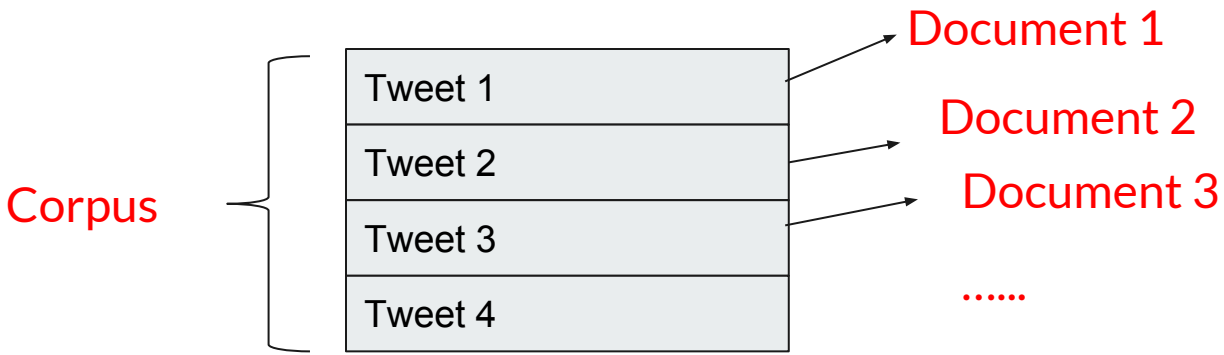
Mind the jargons!

➔ Corpus is a collection of documents



Mind the jargons!

➔ Another example of corpus



More jargons!

➔ DTM (Document-Term Matrix)

Docs	Terms writing,	wrote	yes,	yes.	yet	york.	you	you're	you,	your
1	1	2	1	3	2	1	27	1	1	2
2	0	0	0	1	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0
6	0	0	0	1	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0

>

More jargons!

➔ Stop words

a	ourselves
about	out
above	over
after	own
again	same
against	shan't
all	she
am	she'd
an	she'll
and	she's
any	should
are	shouldn't
aren't	so
as	some
at	such
be	than
because	that
been	that's
before	the
being	their
below	theirs
between	them

More jargons! But a very important one

➔ K (the number of topics in a topic model)

You need to tell algorithm how many topics to look for when generating a topic model.

Back-engineering, let's look at what we will produce based on the tutorial