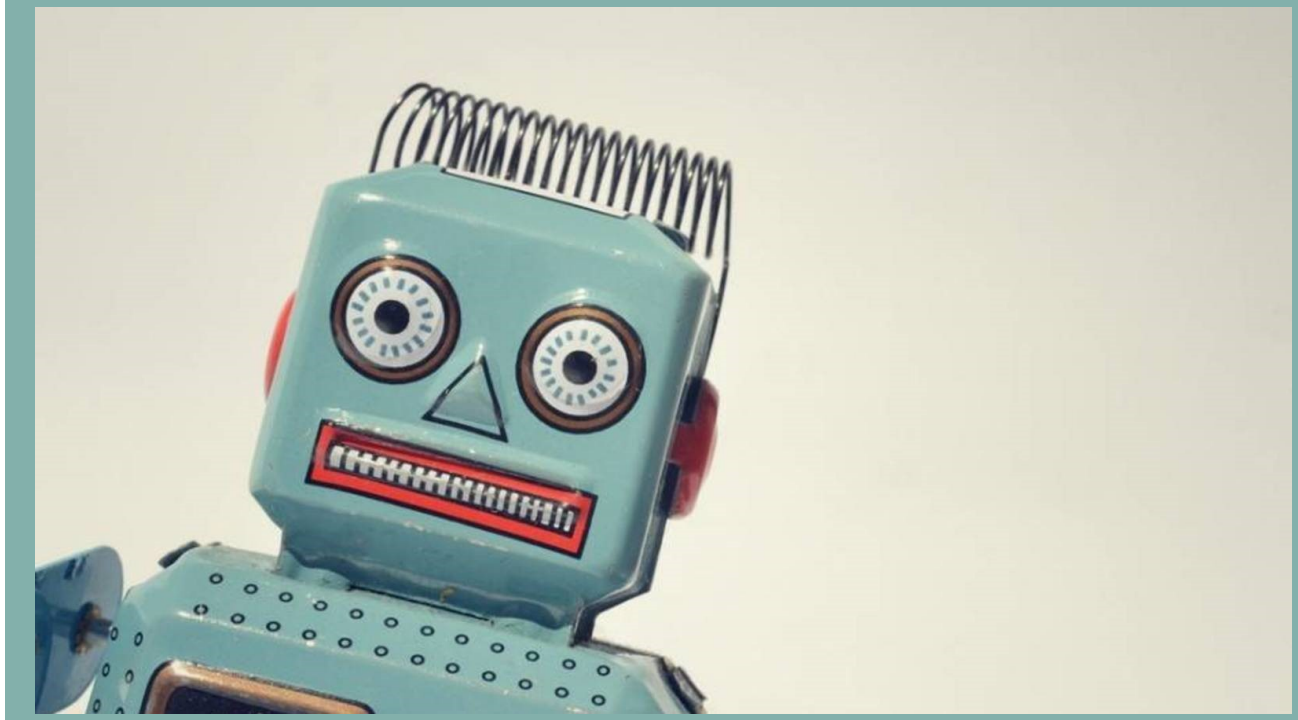
DIGITAL METHODS FOR ANALYSING TEXTS
//

04_Topic modelling



Ana Valdivia
Research Associate
King's College London

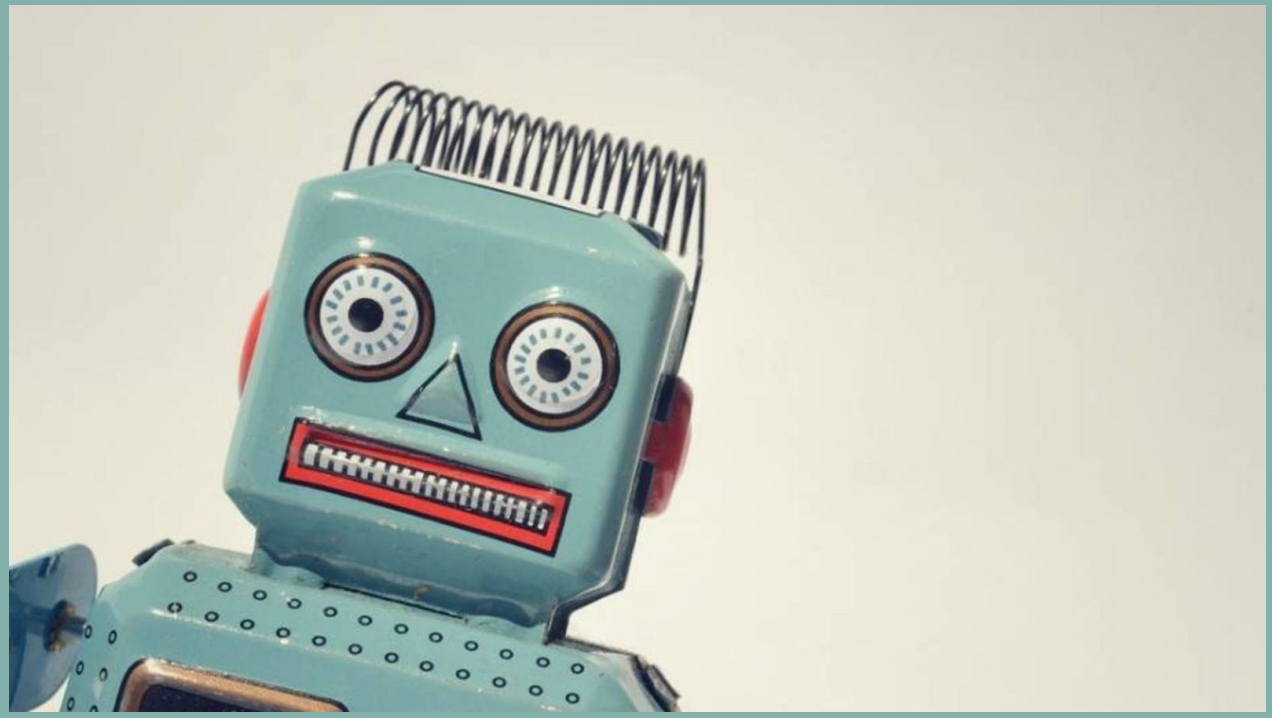
ROAD MAP//

1. INTRODUCTION TO TOPIC MODELLING

2. ALGORITHMS

1. TF-IDF MATRIX + ML CLASSIFIER
2. LATENT DIRICHLET ALLOCATION
3. DOC2VEC + CLUSTERING

INTRODUCTION
//



TOPIC MODELLING//

How would you manually classify documents by topic?

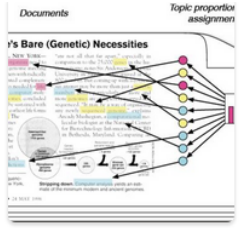
TOPIC MODELLING//

Topic modeling is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents.

Documents are about several **topics** in the same time.
Topics are associated with different **words**.

Topics in the **documents** are express through the **words** that are used.

TOPIC MODELLING//



Topic Models

Edit Task

Miscellaneous • Text Classification

129 papers with code **3** benchmarks **3** datasets

About

Edit

A topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. Topic modeling is a frequently used text-mining tool for the discovery of hidden semantic structures in a text body.

Benchmarks

Add a Result

TREND	DATASET	BEST METHOD	PAPER TITLE	PAPER	CODE	COMPARE
	20 Newsgroups	Bayesian SMM	Learning document embeddings along with their uncertainties			See all
	arXiv	JoSH	Hierarchical Topic Mining via Joint Spherical Tree and Text Embedding			See all
	NYT	JoSH	Hierarchical Topic Mining via Joint Spherical Tree and Text Embedding			See all

https://paperswithcode.com/task/topic-models#code

TOPIC MODELLING//

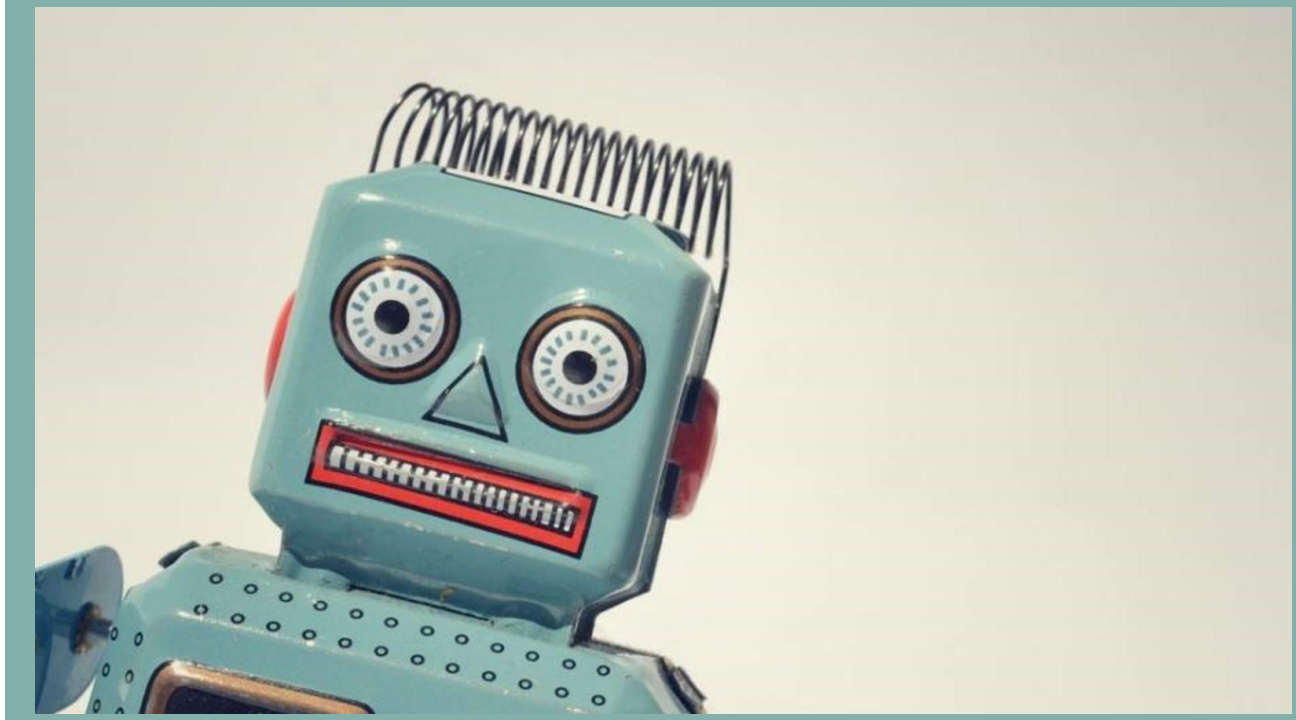
We’ll analyse three approaches:

i) Bag of words + Machine Learning Classifiers

ii) Latent Dirichlet Allocation (LDA)

iii) Word embeddings + Clustering

ALGORITHMS
//



TF-IDF MATRIX + ML//

Bag of words?

Bag of words is an expression used to represent words that appear in documents and sentences.

Like **one-hot-encodings** or **tf-idf matrices.**

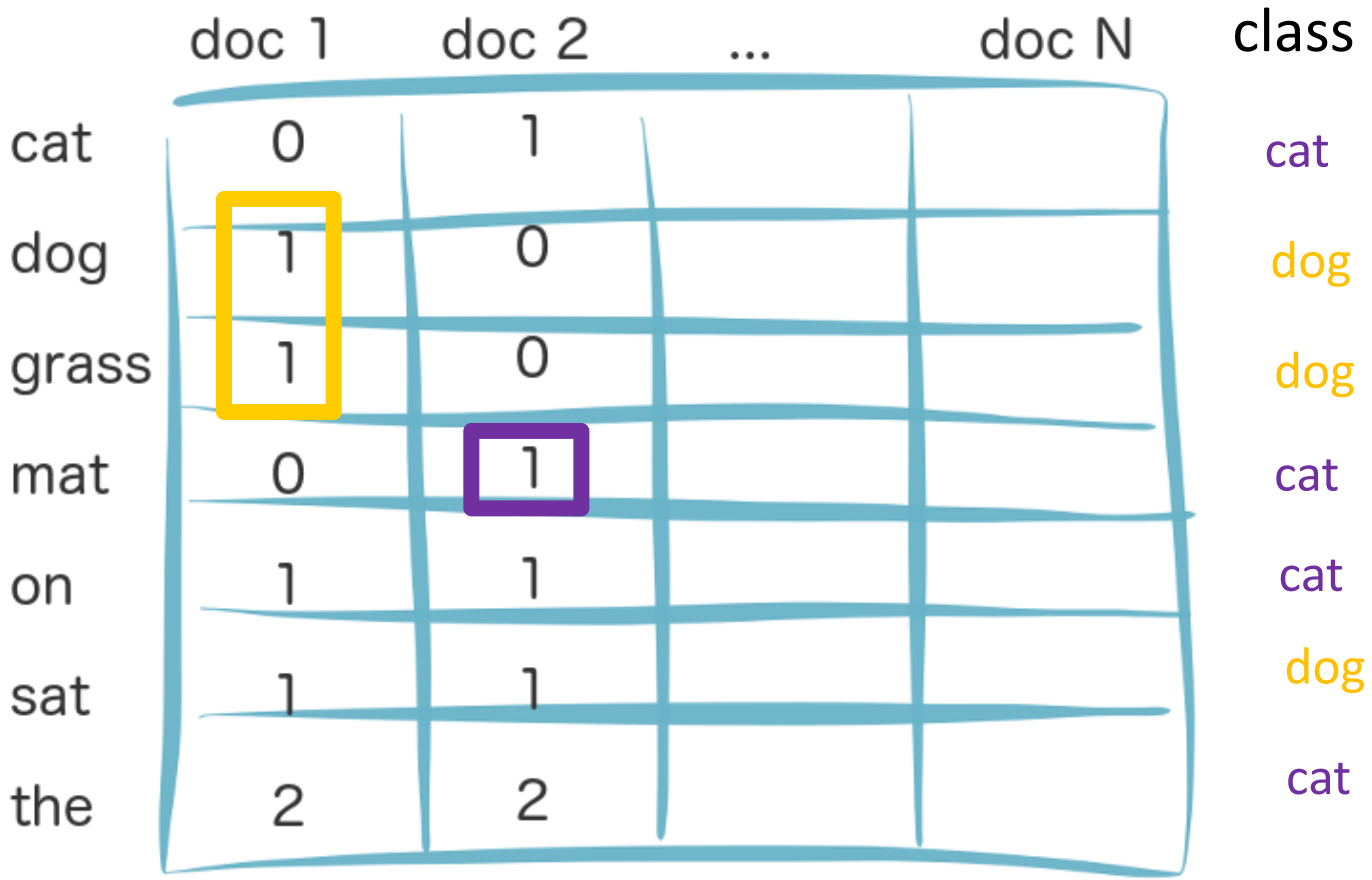
https://medium.com/greyatom/an-introduction-to-bag-of-words-in-nlp-ac967d43b428

TF-IDF MATRIX + ML CLASSIFIER//

Pipeline

TF-IDF MATRIX + ML CLASSIFIER//

It’s not MAGIC!!!

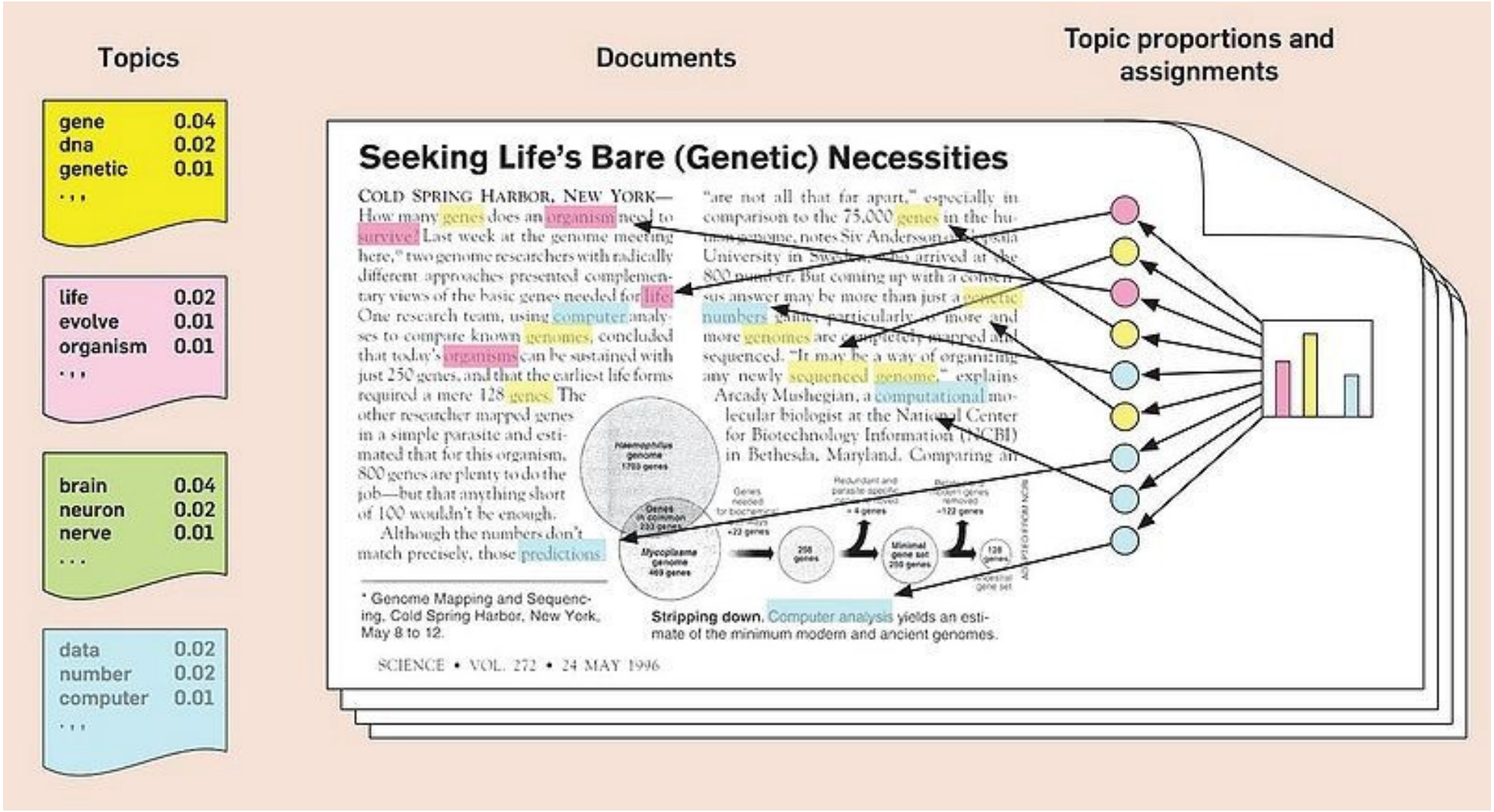


TF-IDF MATRIX + ML CLASSIFIER//

It’s not magic!

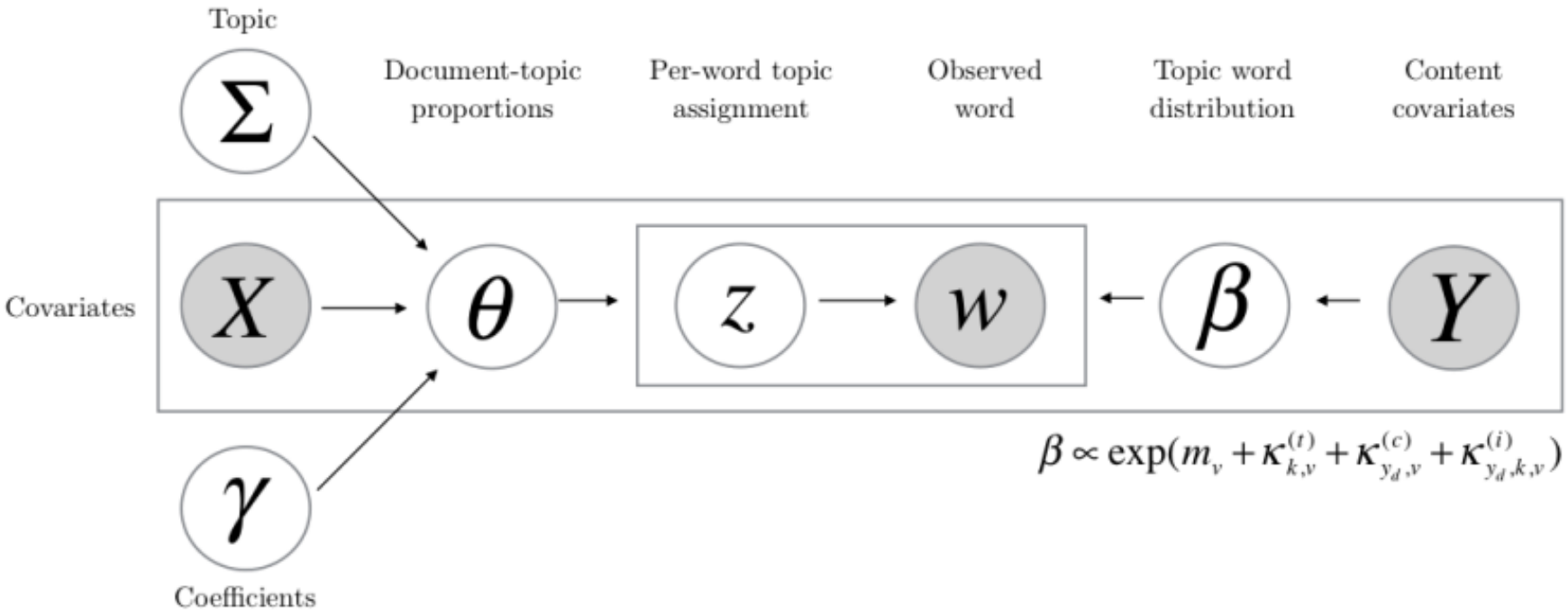
Docs	Drought	Security	Border	Flood	Snowfall	Migrants	Class
Doc1	1	0	0	1	0	0	climatic emergency
Doc2	0	1	0	0	0	1	border security
Doc3	0	0	0	1	1	0	climatic emergency
Doc4	1	0	0	0	0	1	climatic emergency
Doc5	0	1	1	0	0	0	border security
Doc6	0	1	1	0	0	1	border security

LATENT DIRICHLET ALLOCATION//



LATENT DIRICHLET ALLOCATION//

Latent Dirichlet allocation (**LDA**) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar.



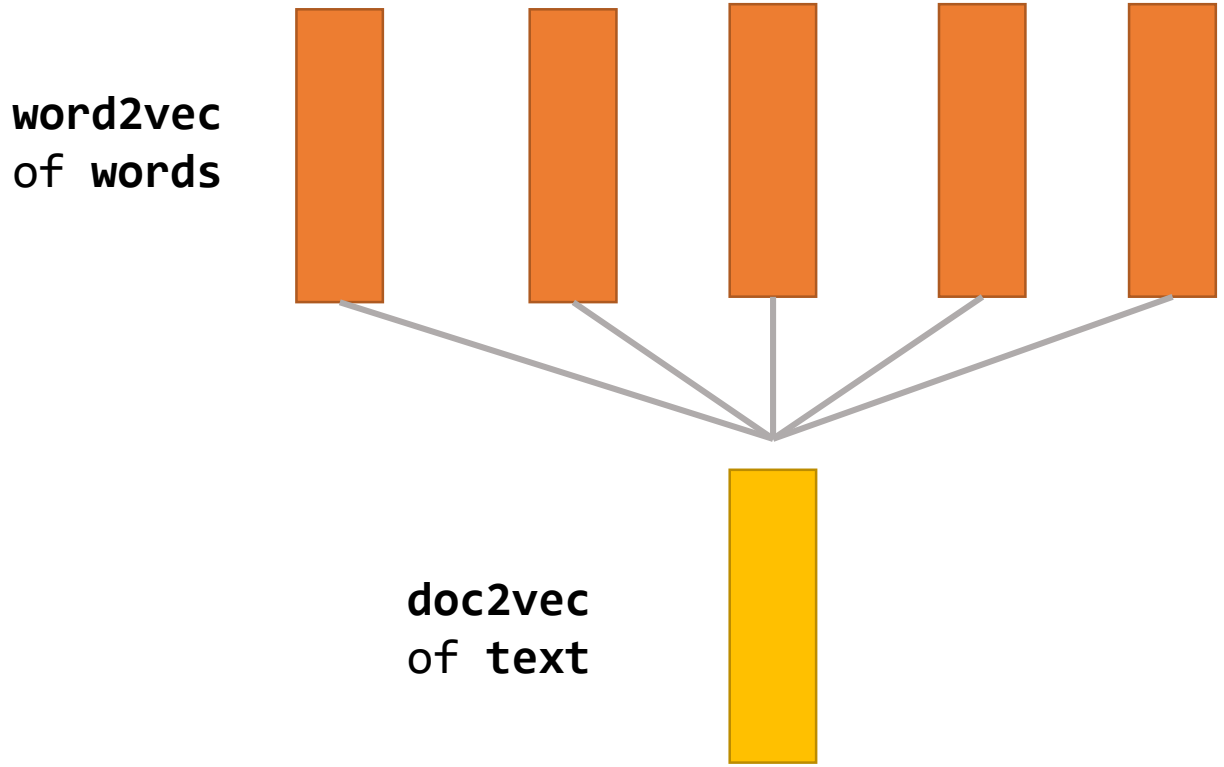
DOC2VEC + CLUSTERING//

DOC2VEC

Document embeddings (doc2vec) are vector representations of documents. Same idea of word embeddings but with documents.

doc2vec

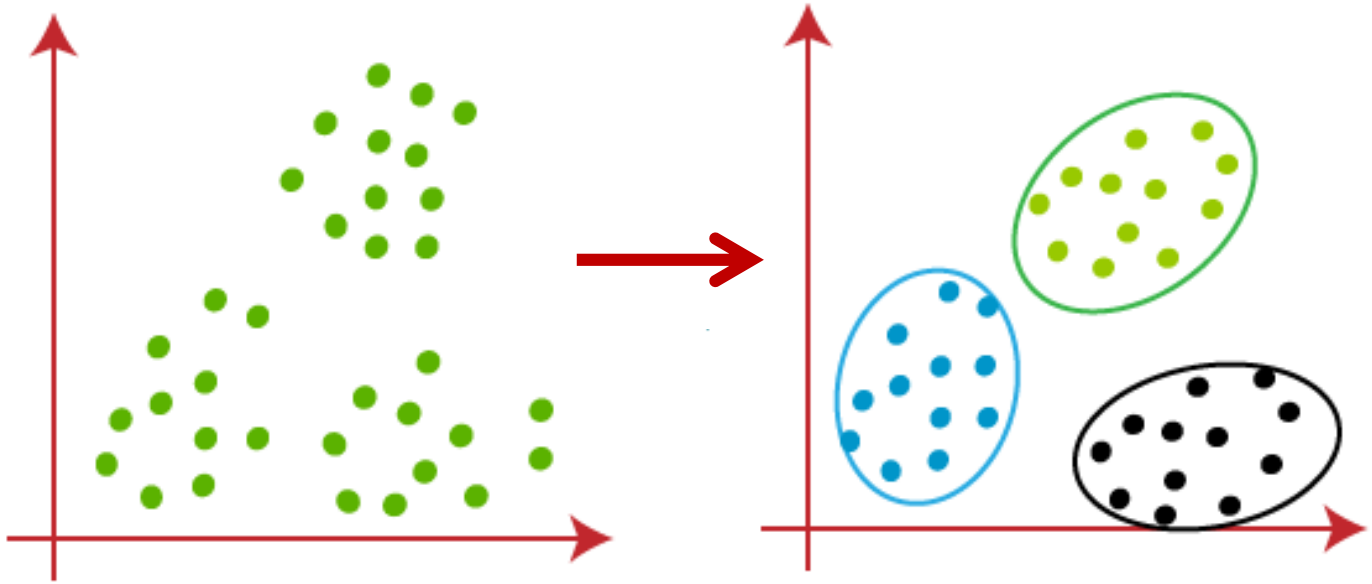
Mean of all word2vec of words within a text.



DOC2VEC + CLUSTERING//

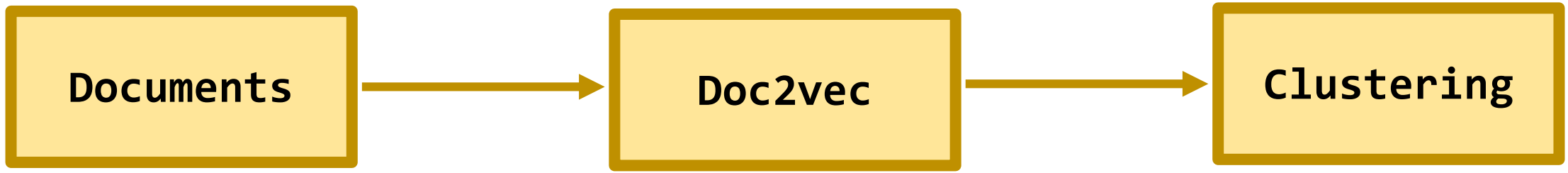
Clustering

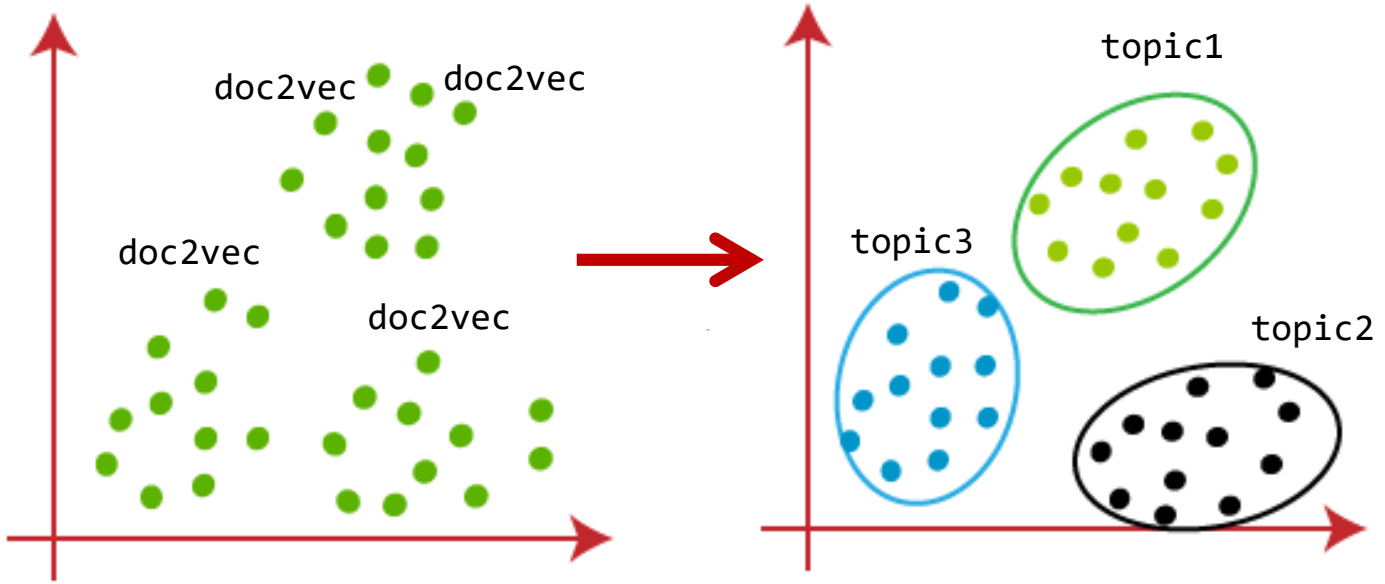
Cluster analysis or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense) to each other than to those in other groups (clusters).



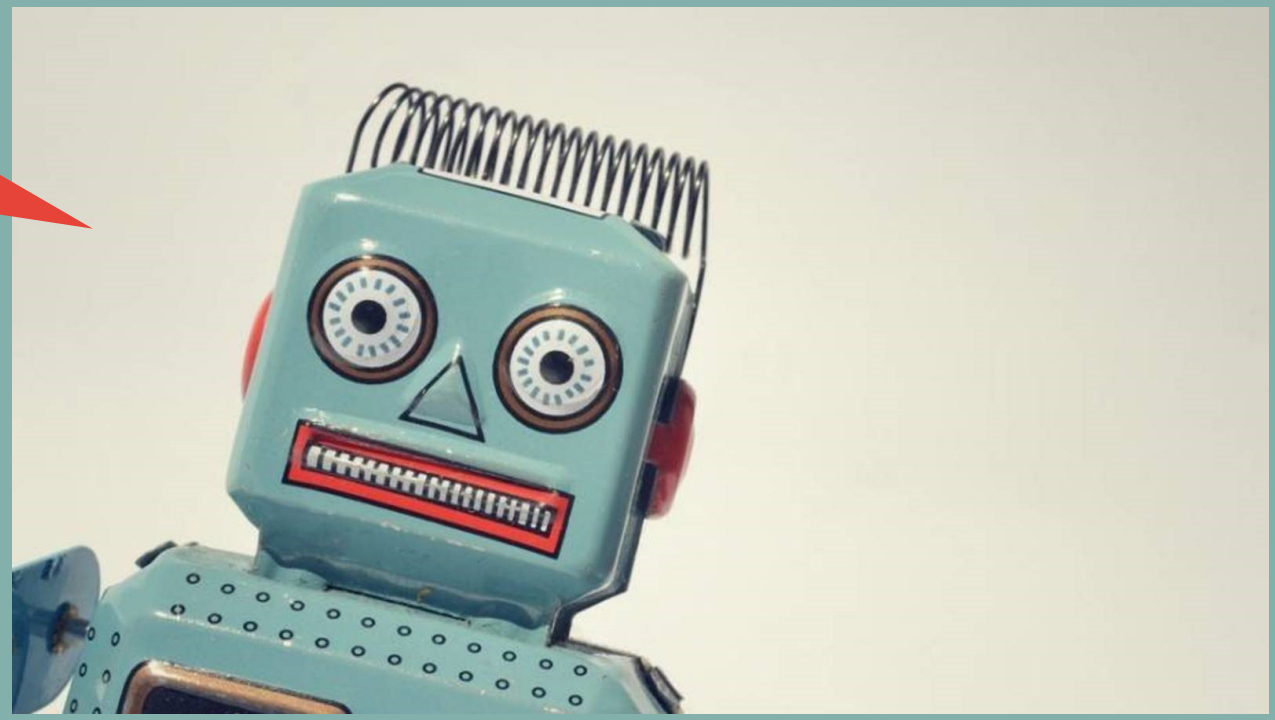
DOC2VEC + CLUSTERING//

Doc2vec + clustering





LET'S CODE!



WE'LL BE BACK IN 15 MIN...