# Discovering Document Topics



Mark Nunnikhoven AWS COMMUNITY HERO @marknca markn.ca

#### Overview

Examine the concepts behind topic modelling

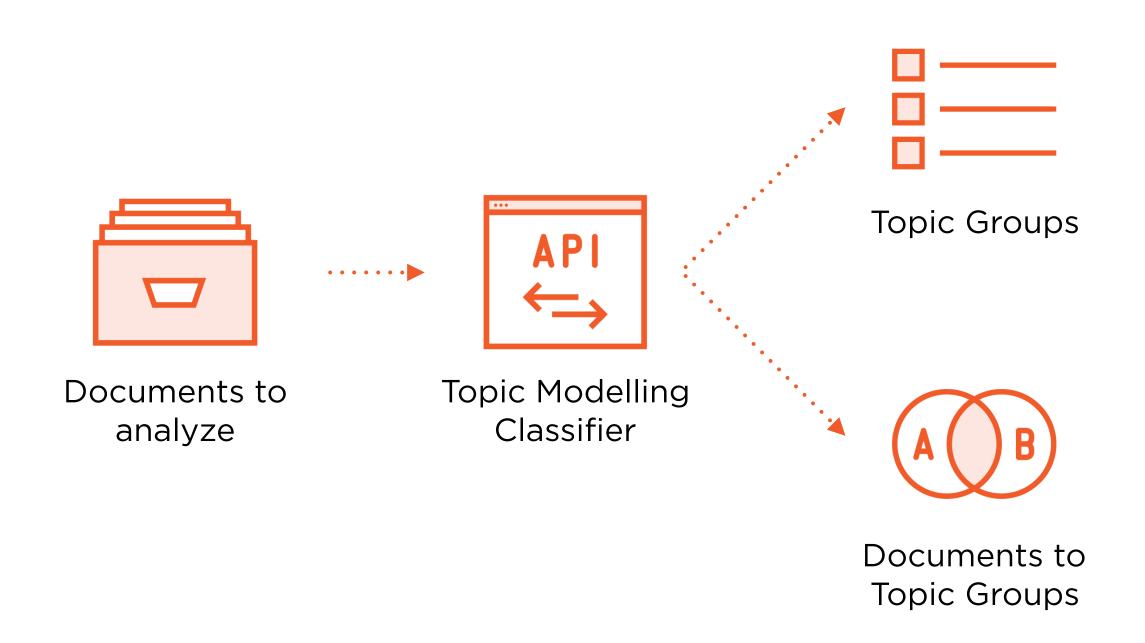
Look at the Latent Dirichlet Allocation method for topic determination

Processing a series of documents to explore the topics presented within

## Topic Modelling

# Identify relevant & common terms and topics in a series of documents

### Topic Modelling Example



# Topic Groups

Topic Group	Keywords	Weight
1	PluralSight	0.97
1	course	0.81
2	API	0.87
2	Job	0.68

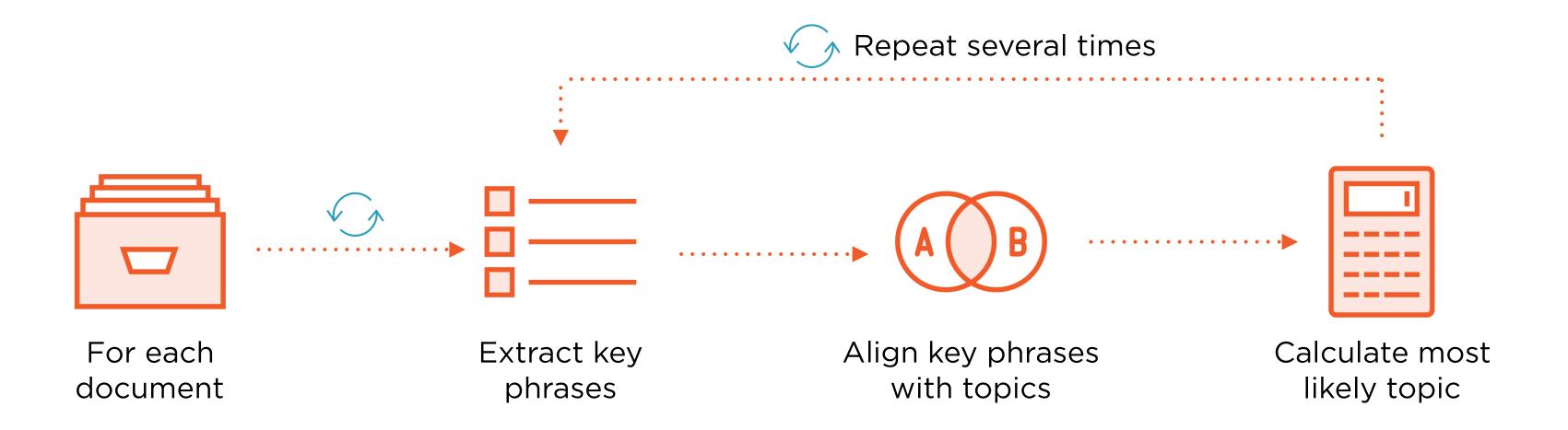
## Documents to Topic Groups

Document Name	Topic Group	Proportion
doc1	1	0.56
doc2	1	0.94
doc3	2	0.81
doc4	2	0.76

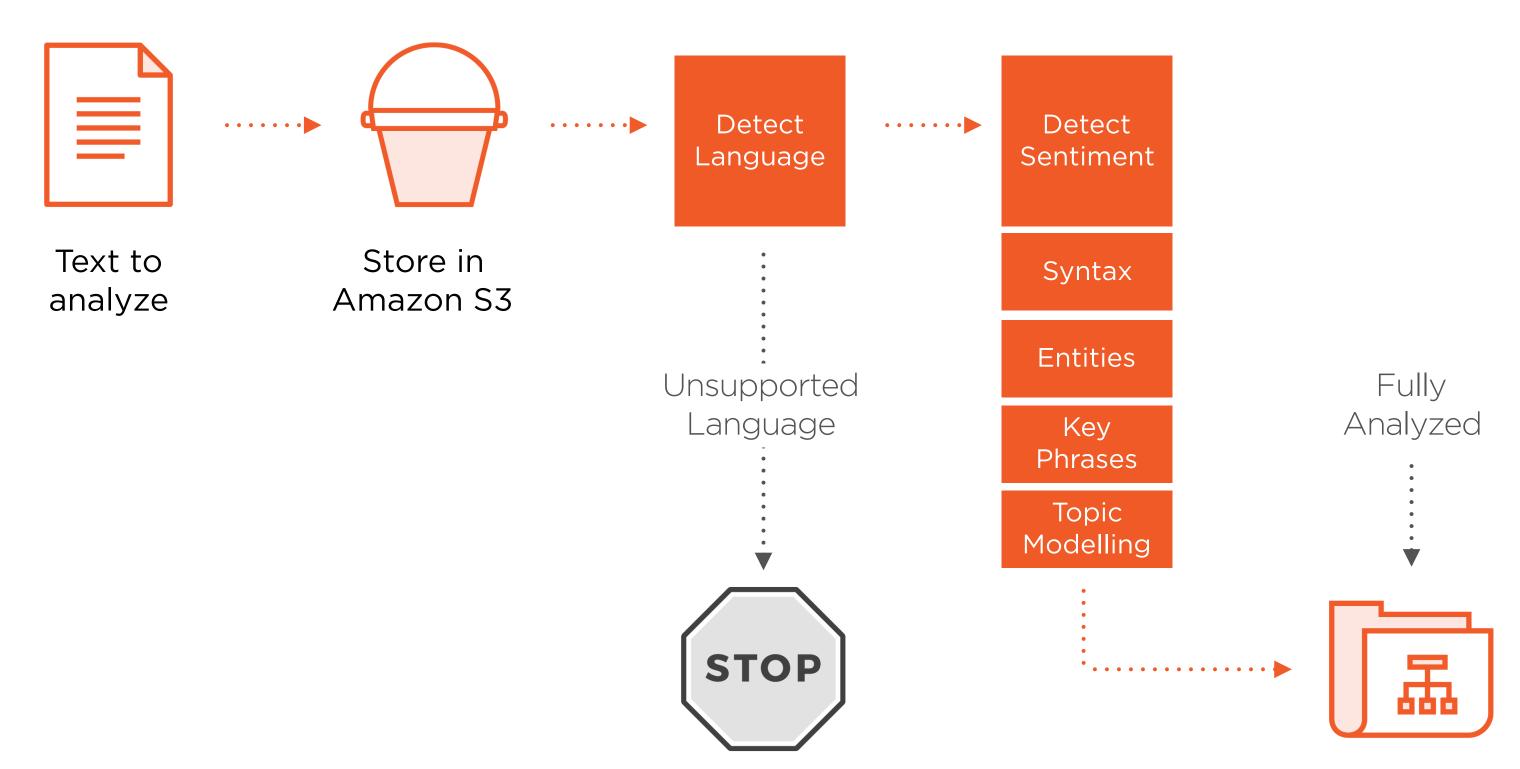
## Latent Dirichlet Allocation

LDA is a topic modelling technique that uses a probabilistic model where each topic is a distribution of words in the sampled documents

#### Latent Dirichlet Allocation



#### Document Analysis



# StartTopicsDetectionJob

An asynchronous API call that locates key phrases within a specified document stored in Amazon S3

```
{
    ...
    "NumberOfTopics": number,
    ...
}
```

#### Request Format

Standard format for "Start\_\_\_Job" requests

"NumberOfTopics" is an optional parameter with a value of 1—100 and helps to narrow the LDA algorithm's modelling

#### Demo

Model topics for a series of documents

Poll the API for the job results

Review the results when the job is complete

#### Review



Topic modelling works of a group of documents to highlight various topics and their rate occurrence within those documents



Amazon Comprehend uses Latent Dirichlet Allocation as a technique to determine topics from the prevalence of key phrases within the text



Topic modelling works best with a large volume of similar documents. This increases topic weights and proportion assessments