REUTERS

INSIGHTS

CONTENT

GOALS & OBJECTIVES

To investigate the Reuters dataset, break it down and begin drilling into the various questions that you may have.

INVESTIGATION

Develop a methodology to save time and classify dataset from Reuters, that could be used for chatbots.

TIME

Demonstrate the effectiveness of such techniques and also examine alternatives.

IMPACT

METHODOLOGY

With limited time, a fast and impactful way to derive insights needs to be done quantitatively.

Data Cleaning of the Reuters dataset.This is done by fixing indexes within the Titles of the dataset and also by removing filler words that are not the topic of the articles.

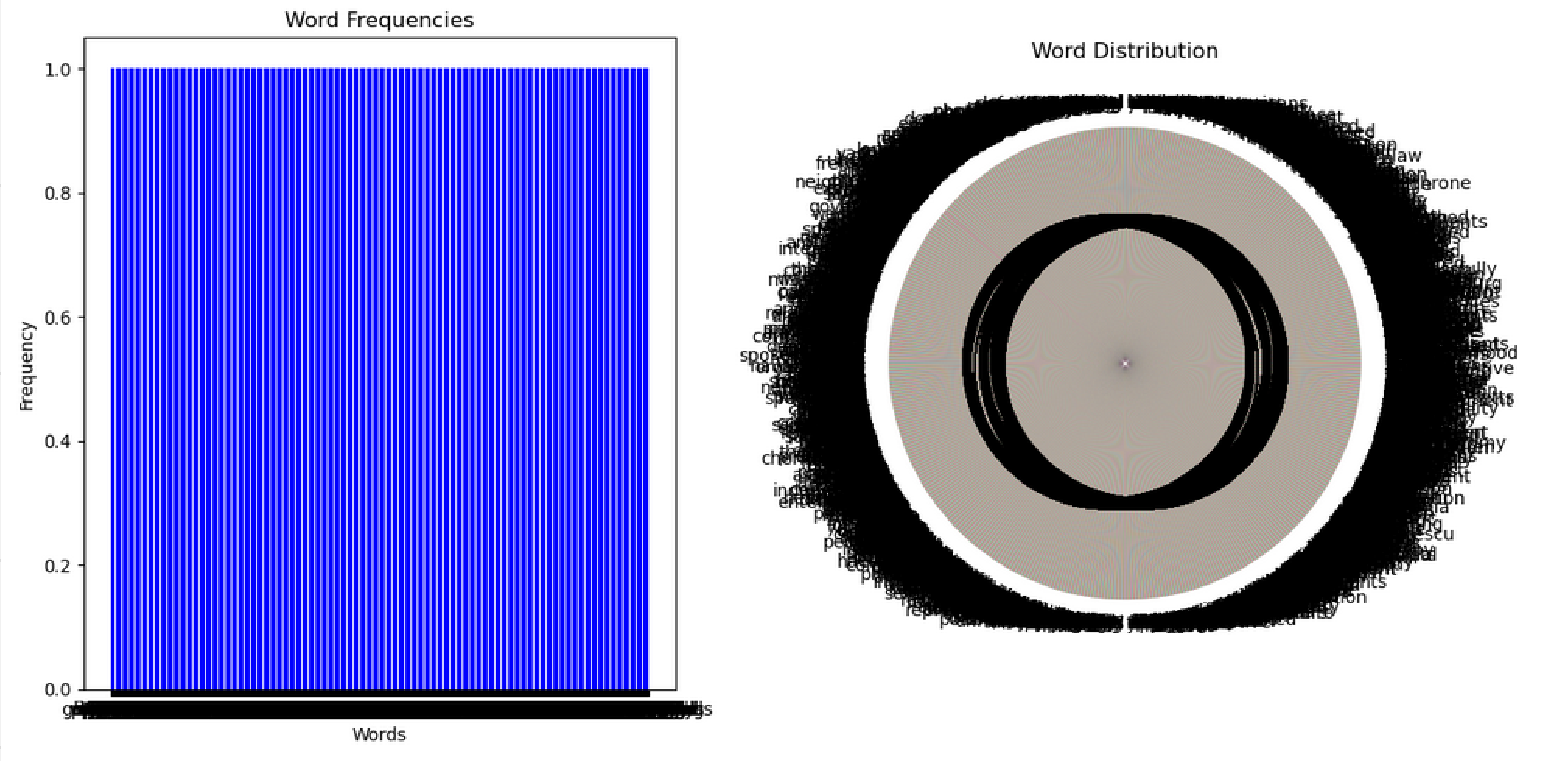
DATA CLEANING

We will be using two models, first is Latent Drichlet Allocation followed by Latent Semantic Analysis and comparing the two models to see which is preferred.

MODEL TRAINING X2

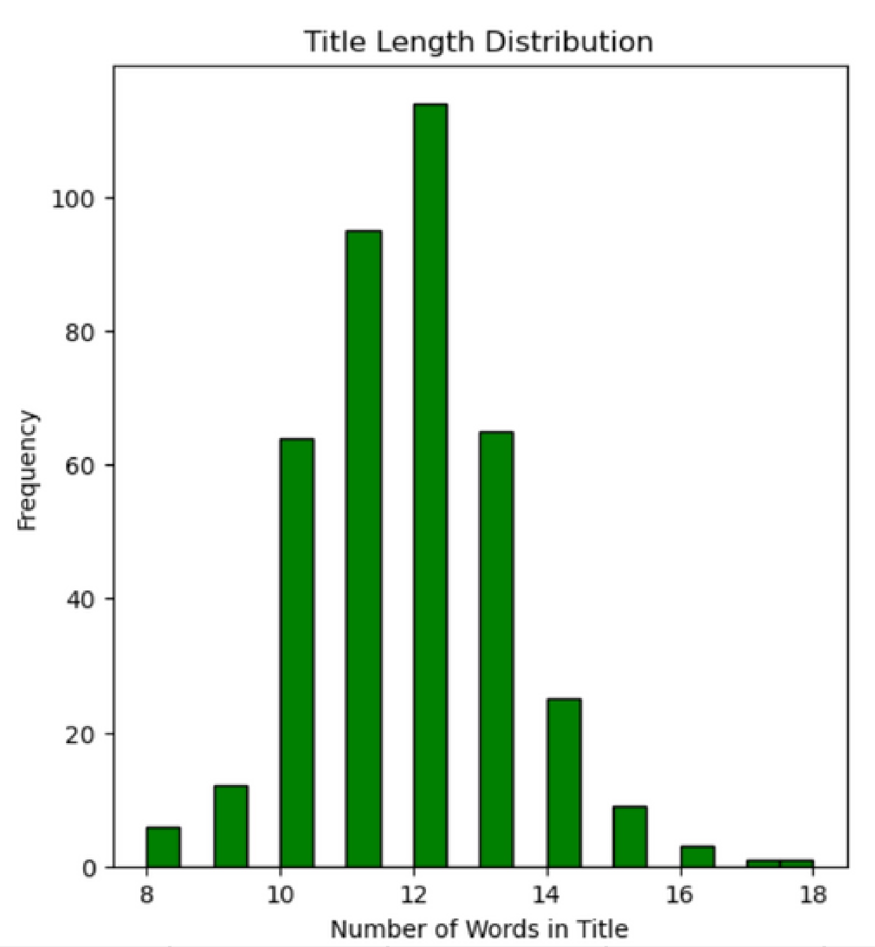
DATA CLEANING & VISUALIZATION

INSIGHTS AT A GLANCE



The plot on the right gives a quick overview into the dataset.

- Too numerous of a word amount to be able to plot.
- All of the word frequencies within the Vocabulary dataset for Reuters are 1 because they are all unique.
- For title, there is a normal distribution on the reuter titles within the dataset.



DATA CLEANING.

Remove Indexes

Get rid of filler words

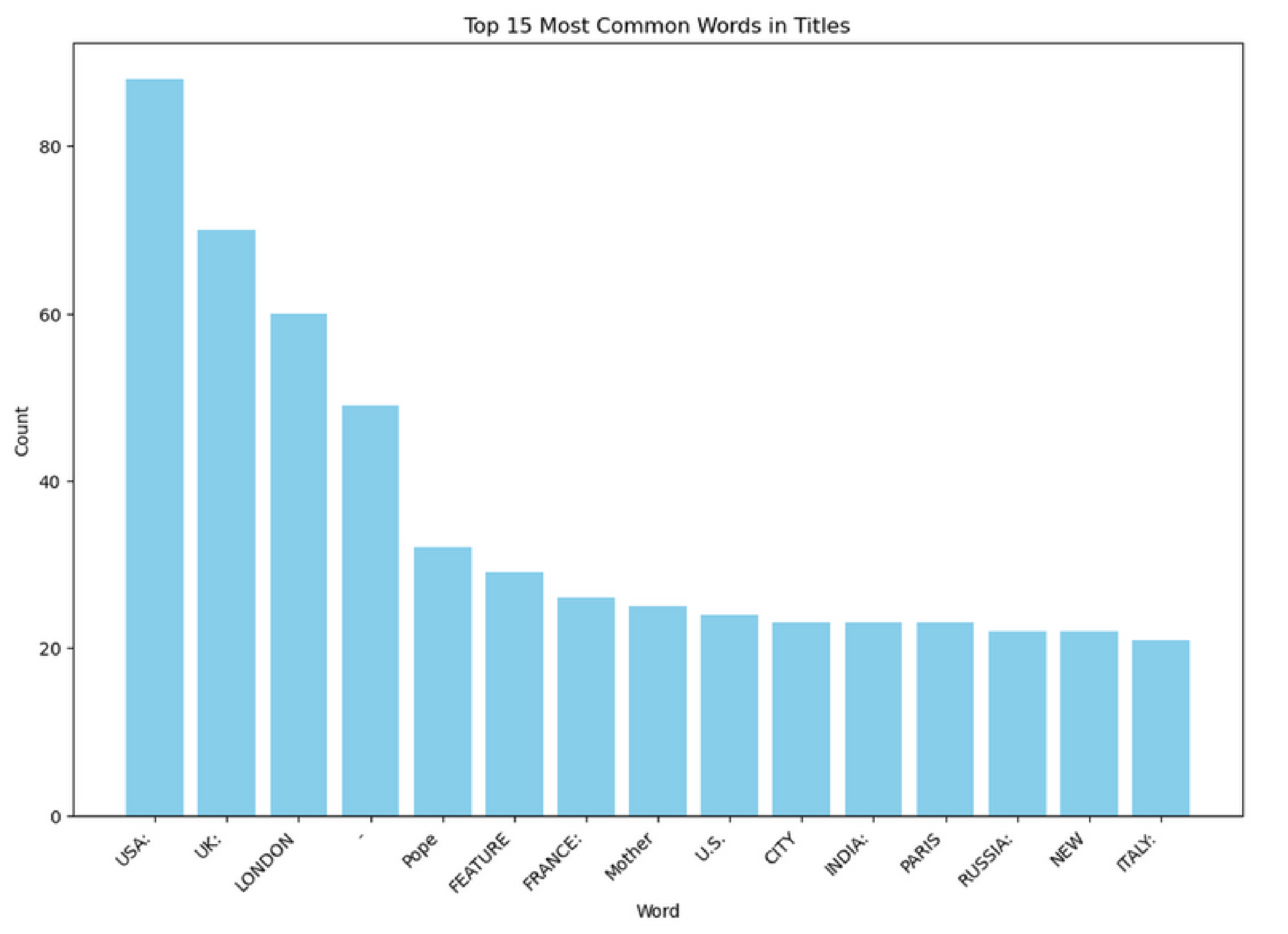
From the chart, we can see that the Titles portion of the dataframe will need some cleaning.

This involves removing some of the fake indexes and resetting them, changing some of the text by removing filler words so that the models have more relevant texts to process.

From the chart on the right:

- Reuter titles are often times focused on the western region. (US, UK, London)
- They report on the Pope more frequently than France.
- The middle east and in fact a number of regions like Africa and South East asia are not in the top 15.

	Title	Unique_Word_Count
0	0 UK: Prince Charles spearheads British royal ...	10
1	1 GERMANY: Historic Dresden church rising from...	12
2	2 INDIA: Mother Teresa's condition said still ...	10
3	3 UK: Palace warns British weekly over Charles...	11
4	4 INDIA: Mother Teresa, slightly stronger, ble...	10
...	...	...
390	390 CANADA: FEATURE - French-speaking Quebec c...	12
391	391 BULGARIA: FEATURE - Bulgarian opera stars ...	12
392	392 USA: Fans end Elvis Presley fete with conc...	12
393	393 UK: Volcano buries studio where rock legen...	11
394	394 USA: Joseph Vostal, ex-Kidder muni banker,...	13



MODELLING LDA & LSA

WHAT IS LATENT DIRICHLET ALLOCATION?

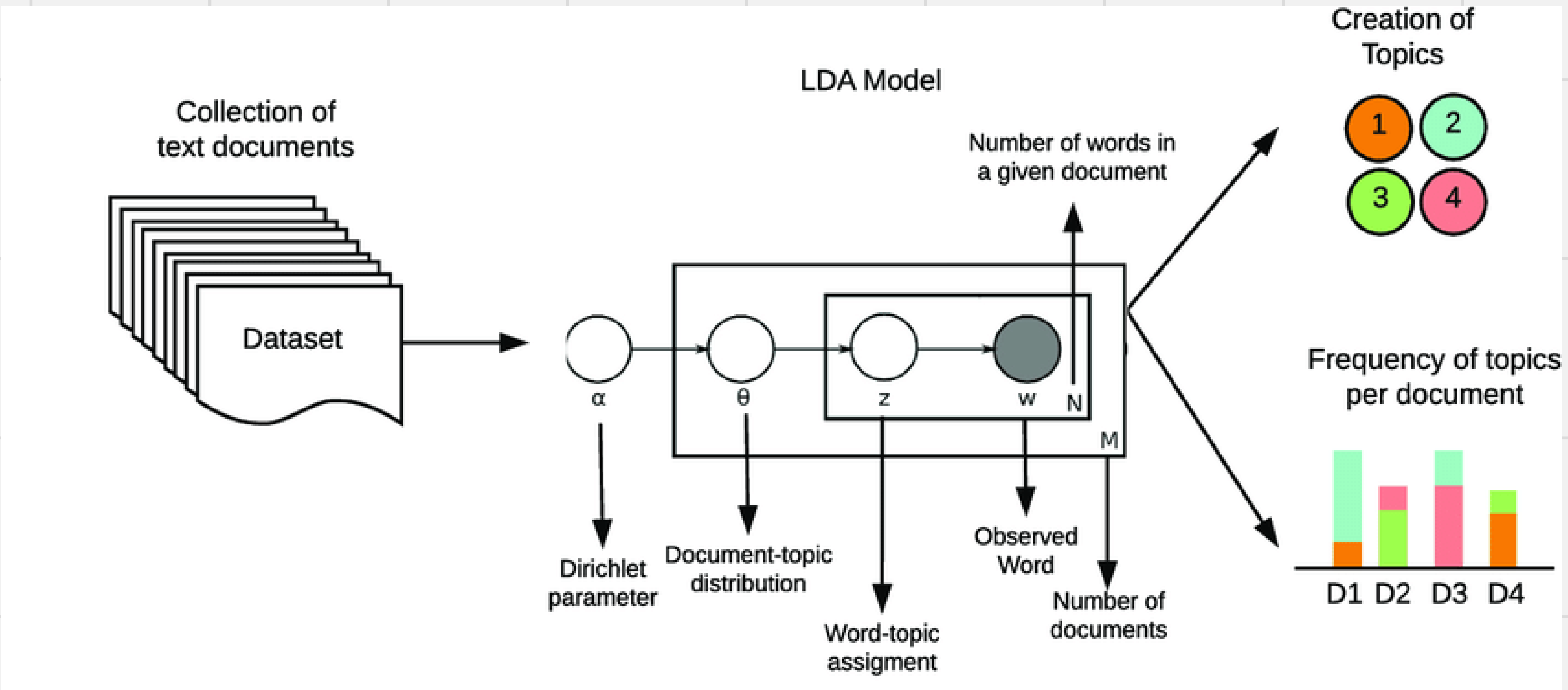
In natural language processing, latent Dirichlet allocation (LDA) is a Bayesian network for modeling automatically extracted topics in textual corpora.

LDA is an example of a Bayesian topic model.

In this, observations (e.g., words) are collected into documents, and each word's presence is attributable to one of the document's topics. Each document will contain a small number of topics.

Within Machine Learning, this is used for topic discovery especially on unsupervised text or large collections of documents.

For example, in a document collection related to pet animals, the terms dog, spaniel, beagle, golden retriever, puppy, bark, and woof would suggest a DOG_related theme, while the terms cat, siamese, Maine coon, tabby, manx, meow, purr, and kitten would suggest a CAT_related theme.



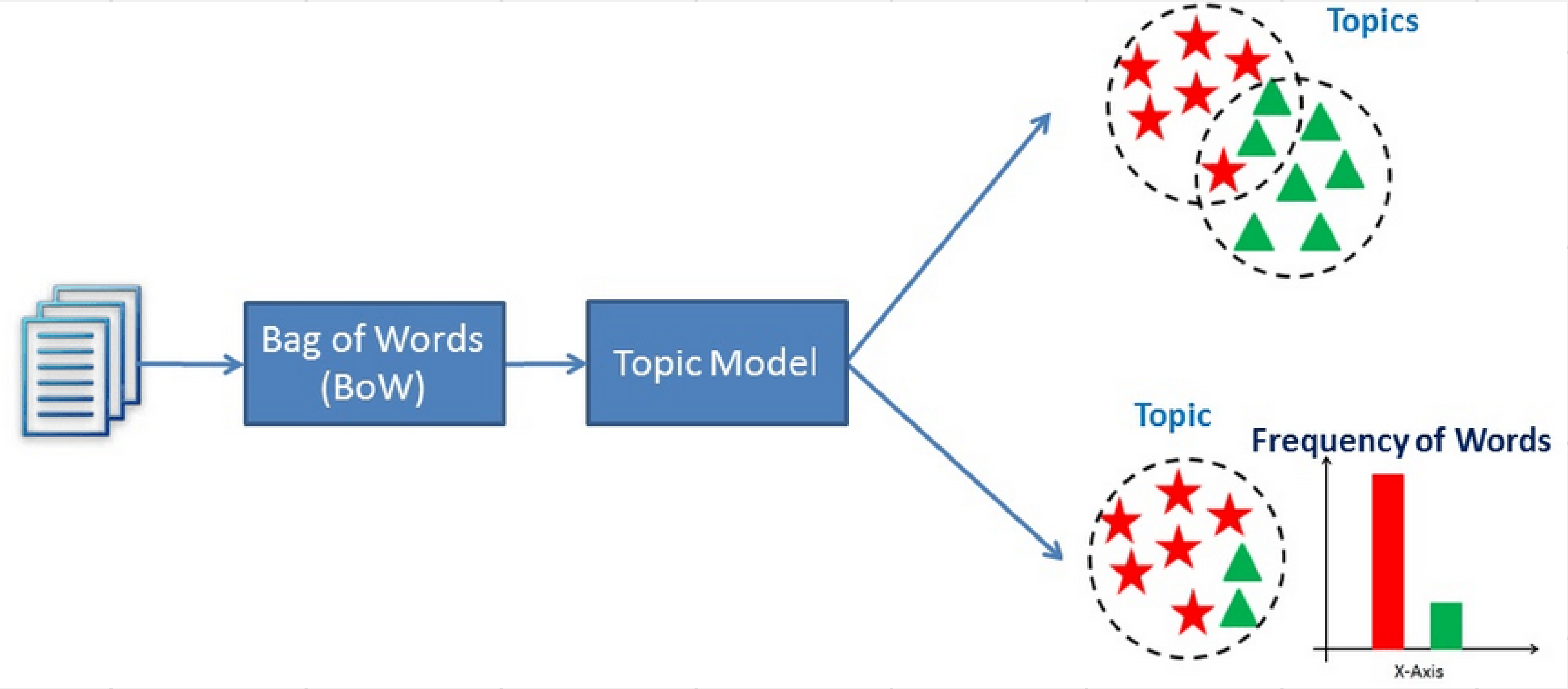
WHAT IS LATENT SEMANTIC ANALYSIS?

Latent semantic analysis (LSA) is a technique in natural language processing, in particular distributional semantics, of analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms.

LSA assumes that words that are close in meaning will occur in similar pieces of text (the distributional hypothesis).

A matrix containing word counts per document (rows represent unique words and columns represent each document) is constructed from a large piece of text and a mathematical technique called singular value decomposition (SVD) is used to reduce the number of rows while preserving the similarity structure among columns.

Documents are then compared by cosine similarity between any two columns. Values close to 1 represent very similar documents while values close to 0 represent very dissimilar documents



MAIN POINTERS FROM EACH PLOT

```
Topic 0: british churchill sale million major letters west
Topic 1: church government political country state people party
Topic 2: elvis king fans presley life concert young
Topic 3: yeltsin russian russia president kremlin moscow michael
Topic 4: pope vatican paul john surgery hospital pontiff
Topic 5: family funeral police miami versace cunanan city
Topic 6: simpson former years court president wife south
Topic 7: order mother successor election nuns church nirmala
Topic 8: charles prince diana royal king queen parker
Topic 9: film french france against bardot paris poster
Topic 10: germany german war nazi letter christian book
Topic 11: east peace prize award timor quebec belo
Topic 12: n't life show told very love television
Topic 13: years year time last church world people
Topic 14: mother teresa heart calcutta charity nun hospital
Topic 15: city salonika capital buddhist cultural vietnam byzantine
Topic 16: music tour opera singer israel people film
Topic 17: church catholic bernardin cardinal bishop wright death
Topic 18: harriman clinton u.s ambassador paris president churchill
Topic 19: city museum art exhibition century million churches
```

Latent Dirichlet Allocation

```
LSA Topic 0: divorce, spokesman, minister, take, including, week, saying, american
LSA Topic 1: divorce, minister, saying, american, take, week, spokesman, including
LSA Topic 2: last, first, world, miami, mother, catholic, orthodox, charles
LSA Topic 3: black, published, tour, child, white, last, first, world
LSA Topic 4: archbishop, u.s, small, president, during, against, year, life
LSA Topic 5: give, pontiff, go, london, help, foreign, outside, earlier
LSA Topic 6: television, members, held, prime, never, mass, following, pontiff
LSA Topic 7: brought, appendix, big, sent, stay, jews, cancer, christmas
LSA Topic 8: took, part, great, expected, early, born, wife, taken
LSA Topic 9: expected, born, taken, white, wife, italian, although, england
LSA Topic 10: queen, son, house, children, next, great, part, took
LSA Topic 11: want, britain, public, clinton, death, part, great, time
LSA Topic 12: took, great, part, order, time, capital, since, leader
LSA Topic 13: want, economic, britain, public, exhibition, wife, white, asked
LSA Topic 14: rights, days, capital, friday, popular, mark, exhibition, month
LSA Topic 15: rights, went, asked, since, saturday, art, time, wife
LSA Topic 16: economic, rights, expected, taken, born, time, south, since
LSA Topic 17: royal, clinton, death, opinion, exhibition, woman, bishop, us
LSA Topic 18: white, economic, death, clinton, italian, capital, day, political
LSA Topic 19: asked, saturday, prize, rome, want, national, britain, winston
```

Latent Semantic Analysis

CONCLUSION

Latent Dirichlet Allocation

- GENERAL PROBABILISTIC MODEL
- USES BAYESIAN INFERENCE TO FIND UNDERLYING TOPICS

Latent Semantic Analysis

- LEVERAGES SINGULAR VECTOR DECOMPOSITION TO REDUCE DIMENSIONALITY
- CAPTURES UNDERLYING RELATIONSHIPS BETWEEN WORDS

THANK YOU