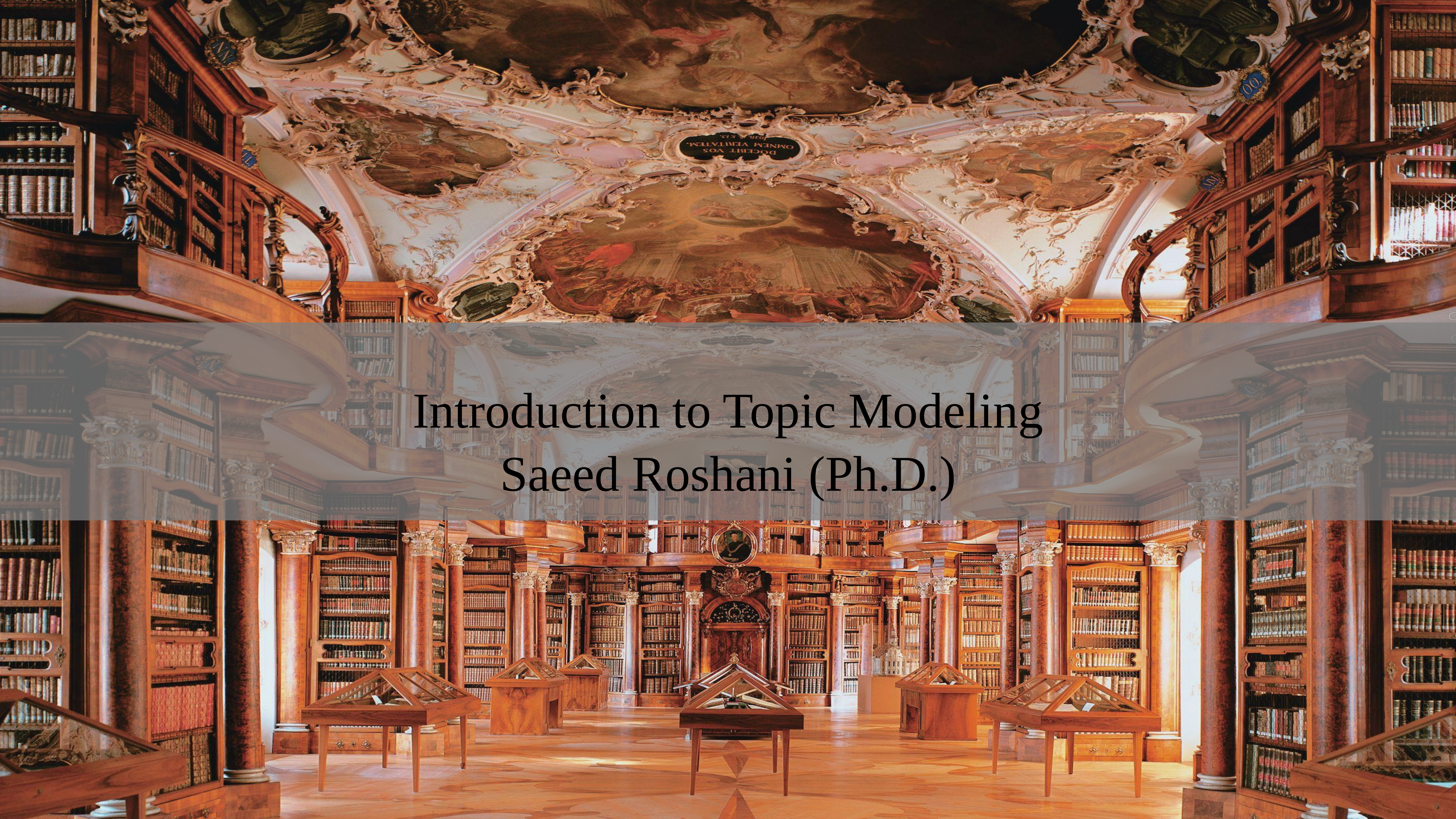
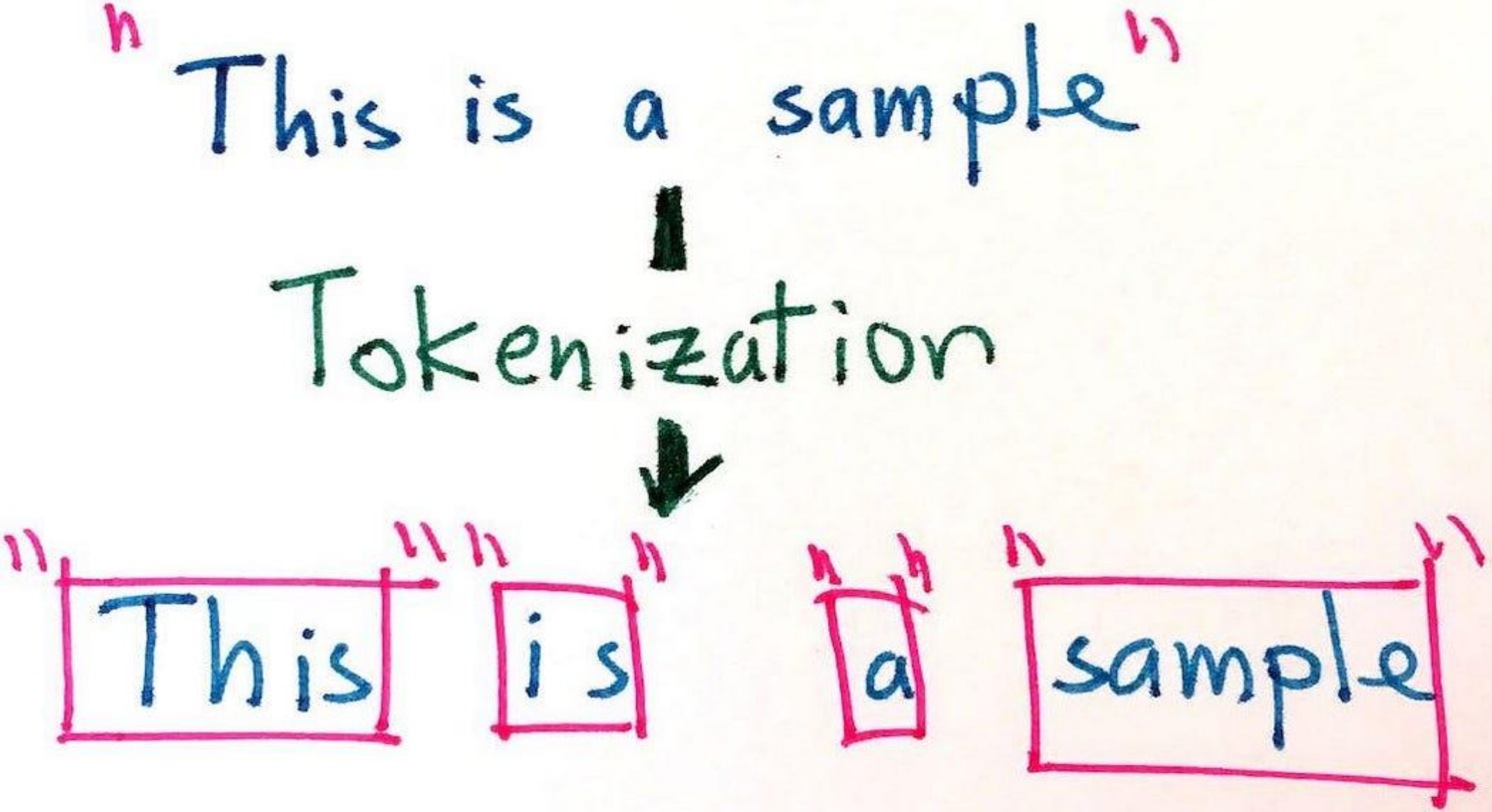


Introduction to Topic Modeling

Saeed Roshani (Ph.D.)

The goal of this study is to develop a technology analysis of the evolution of some critical quantum technologies to improve technological forecasting. Methodology applies S-shaped curves based on patent data to analyze the evolutionary phases of quantum technologies over the course of time.


"This is a sample"
Tokenization
"This" "is" "a" "sample"

The goal of this study is to develop a technology analysis of the evolution of some critical quantum technologies to improve technological forecasting. Methodology applies S-shaped curves based on patent data to analyze the evolutionary phases of quantum technologies over the course of time.

```
# Download necessary data and models
import nltk
nltk.download('stopwords')
nlp = spacy.load('en_core_web_sm')
```

The goal of this study is to develop a technology analysis of the evolution of some critical quantum technologies to improve technological forecasting. Methodology applies S-shaped curves based on patent data to analyze the evolutionary phases of quantum technologies over the course of time.

```
# Define stop words
stop_words = stopwords.words('english')
stop_words.extend(['research','use','new','article','study','model','define','background','examine','paper',
                   'eaxamination','ivestigate','aim','goal','objective','US','china','result',
                   'measure','measurment','quality','analysis','approach','method','propose','investigate'])
```

technology analysis

evolution quantum technologies

technological forecasting. S-shaped curves

patent evolutionary phases quantum

technologies .

```
def make_bigrams(texts):
    bigram = Phrases(texts, min_count=10, threshold=50)
    bigram_phraser = Phraser(bigram)
    trigram = Phrases(bigram_phraser[texts], min_count=10, threshold=50)
    trigram_phraser = Phraser(trigram)
    return [trigram_phraser[bigram_phraser[doc]] for doc in texts]
```

technology analysis
evolution quantum technolog(y)ies
technological forecasting. applies S-shaped curves
patent evolutionary phases quantum
technolog(y)ies .

```
def lemmatize(text, allowed_postags=['NOUN', 'VERB', 'ADJ', 'ADV']):
    doc = nlp(text)
    return [token.lemma_ for token in doc if token.pos_ in allowed_postags and token.lemma_ not in stop_words]
```

technolog analysi evolution quantum technolog technolog forecast . S-shape curve patent evolutionary phase quantum technolog .

```
def stem(text):
    p = PorterStemmer()
    return [p.stem(token) for token in text]
```

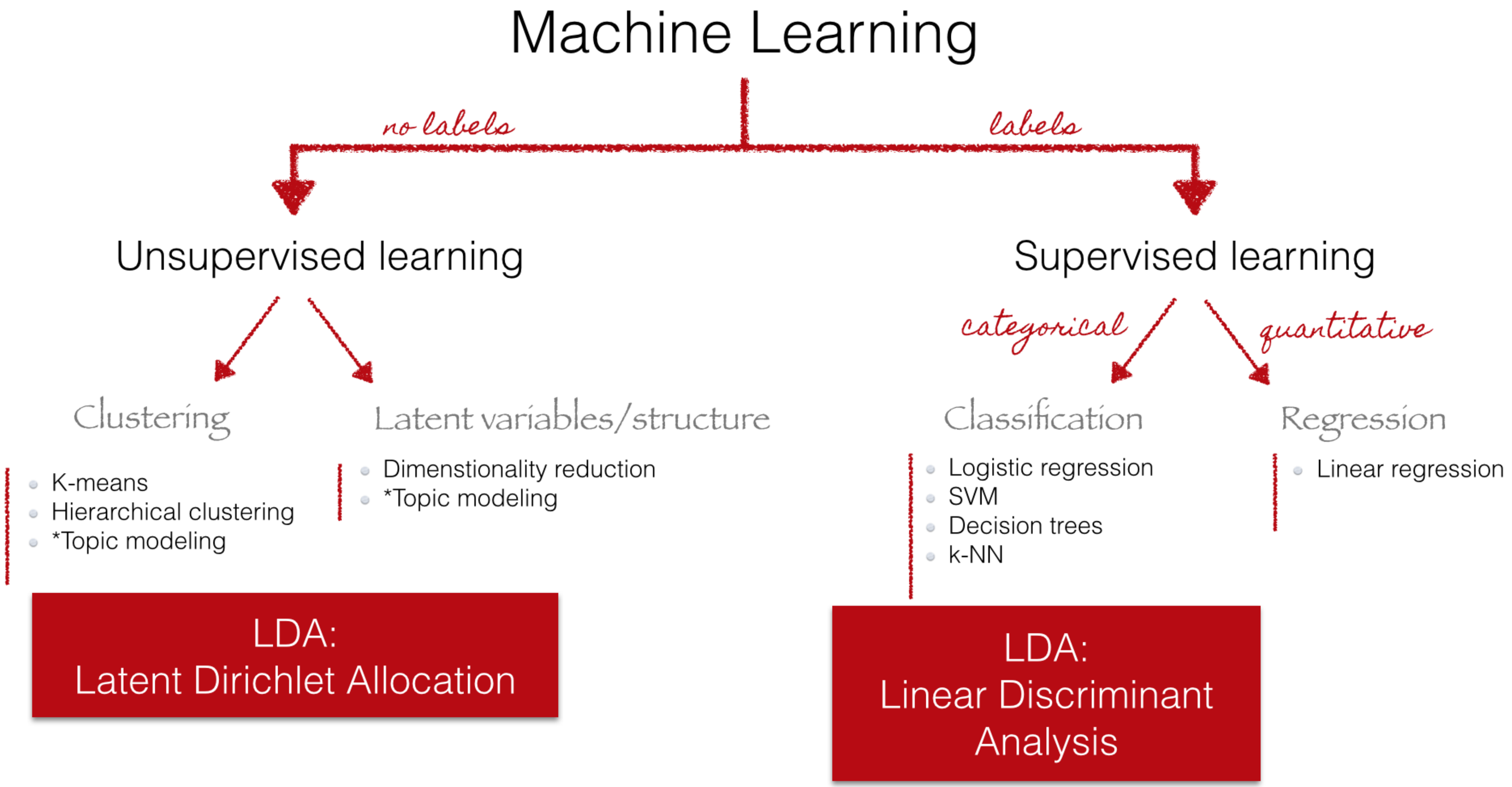
```python
import gensim

# Create a list of lists of the cleaned tokens
text = df['stemmed'].tolist()

# Create a dictionary of the cleaned tokens
dictionary = gensim.corpora.Dictionary(text)
print('Total Vocabulary Size:', len(dictionary))

# Filter out words that occur less than 5 documents, or more than 90% of the documents.
dictionary.filter_extremes(no_below=2, no_above=0.7)
print('Total Vocabulary Size after removing:', len(dictionary))

# Create a bag-of-words corpus of the cleaned tokens
corpus = [dictionary.doc2bow(t) for t in text]
```


Machine Learning
no labels
labels
Unsupervised learning
Supervised learning
categorical
quantitative
Clustering
Latent variables/structure
Classification
Regression
K-means
Hierarchical clustering
*Topic modeling
Dimenstionality reduction
*Topic modeling
Logistic regression
SVM
Decision trees
k-NN
Linear regression
LDA:
Latent Dirichlet Allocation
LDA:
Linear Discriminant Analysis

Definitions

A topic model is a type of statistical model for **discovering** the abstract "**topics**" that occur in a collection of **documents**.

Topic models are a suite of algorithms that uncover the **hidden thematic** structure in document collections. These algorithms help us develop new ways to search, browse and summarize large archives of texts.

Topic models provide a simple way to analyze large volumes of **unlabeled text**. A "topic" consists of a **cluster** of words that **frequently** occur together

Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a **generative statistical model** used for topic modeling, which aims to identify the underlying themes or topics in a collection of documents. In LDA, a document is seen as a mixture of various topics and each topic is represented by a distribution over words in the vocabulary.

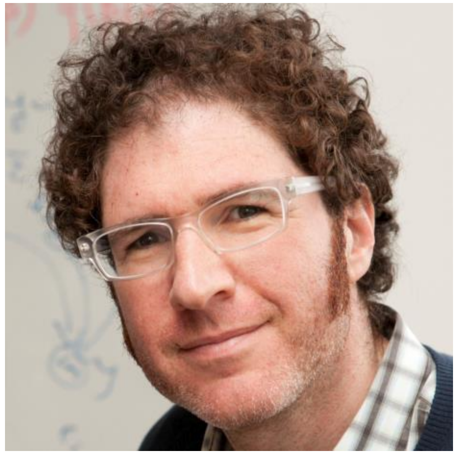
The goal of LDA is to infer the topic distribution in each document and the word distribution in each topic based on the observed words in the corpus. This is done through **Bayesian inference**, where the posterior distribution over the latent variables (i.e., the topic distribution in each document and the word distribution in each topic) is approximated using **Markov chain Monte Carlo** (MCMC).

By simulating this Markov chain for a large number of iterations, we can approximate the posterior distribution over the latent variables. This allows us to estimate the most likely topic and word distributions that generated the observed data, and to make inferences about the relationships between topics and words in the corpus.

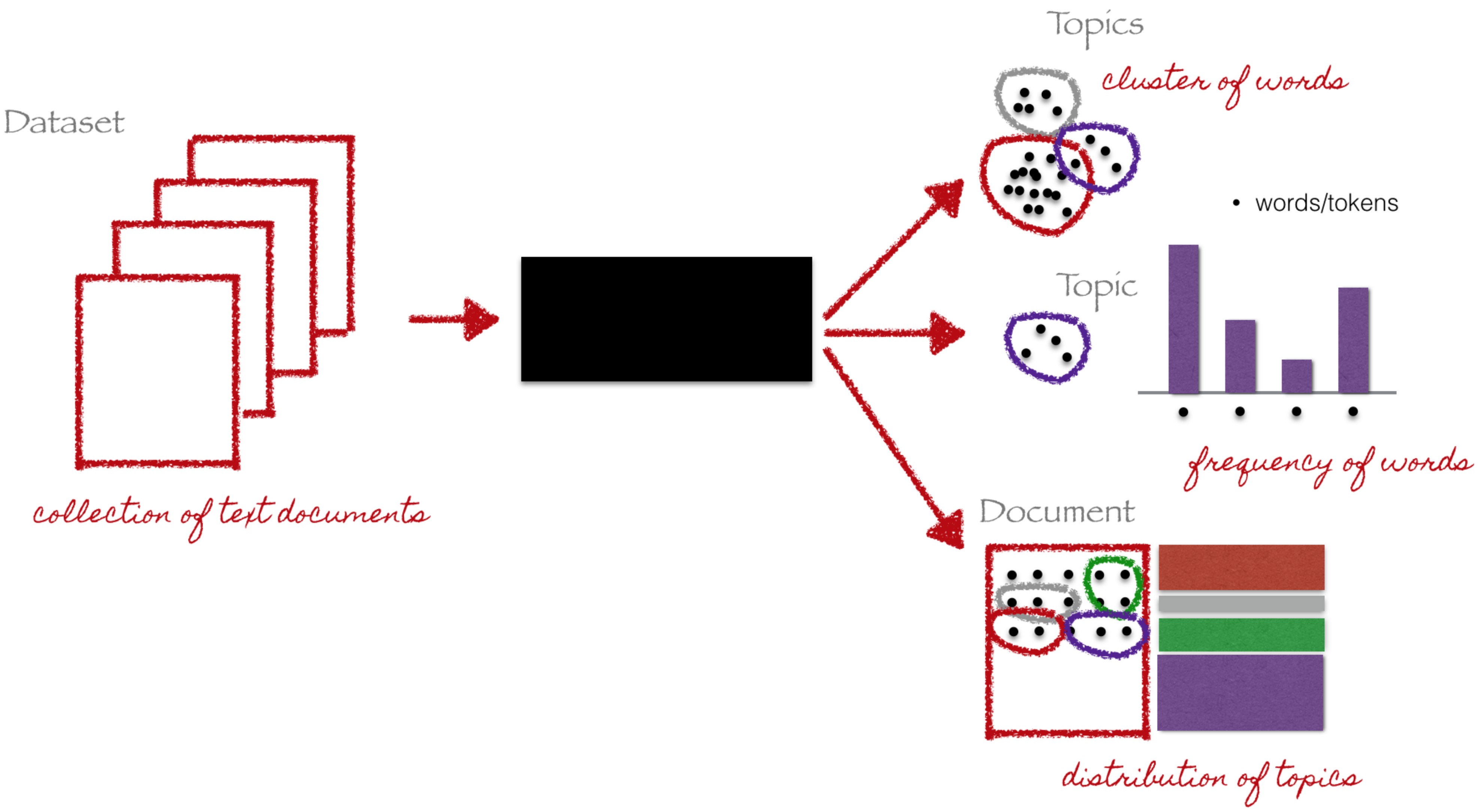
Latent Dirichlet Allocation

David M. Blei, Andrew Y. Ng and **Michael I. Jordan**
University of California, Berkeley
Berkeley, CA 94720

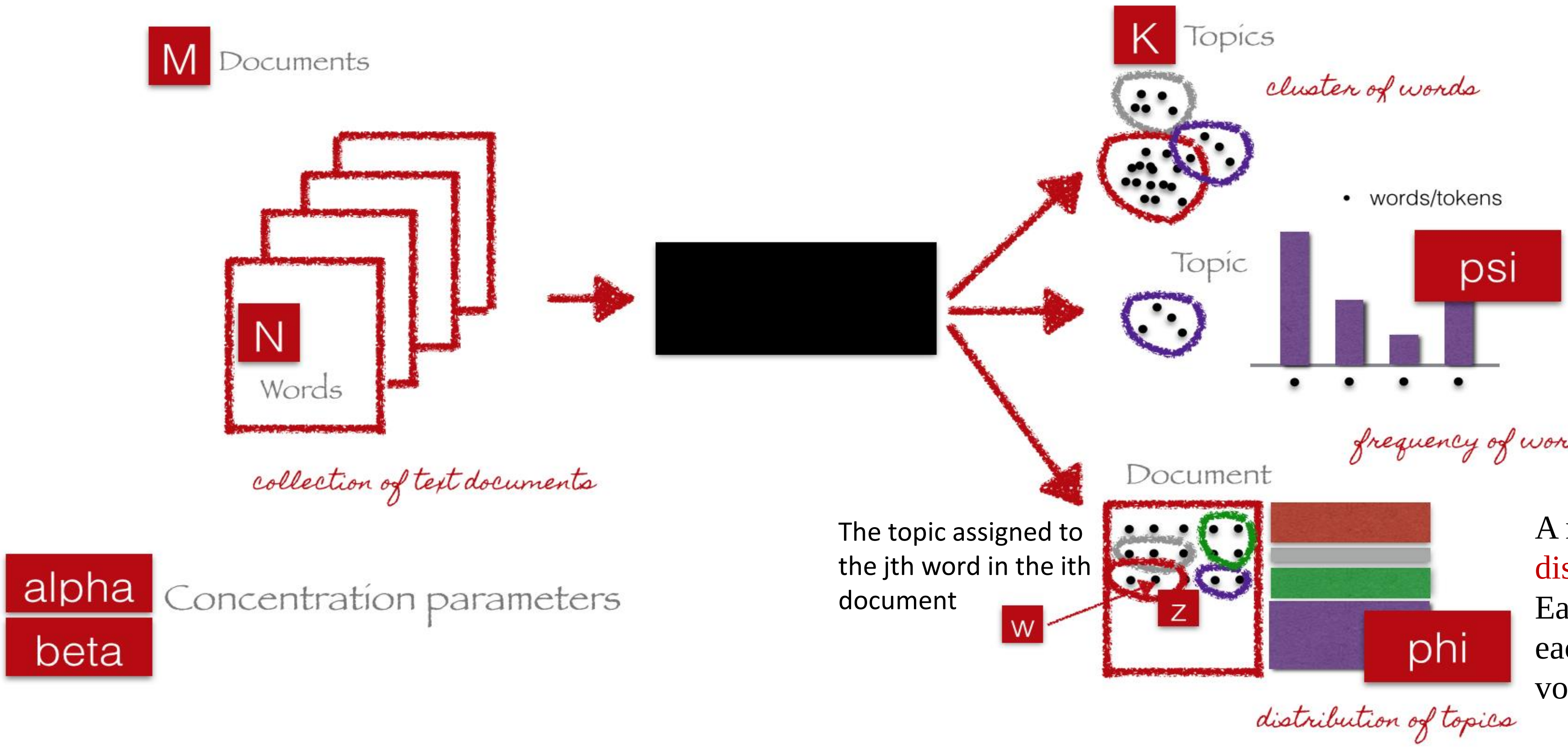
Abstract



We propose a generative model for text and other collections of discrete data that generalizes or improves on several previous models including naive Bayes/unigram, mixture of unigrams [6], and Hofmann's aspect model, also known as probabilistic latent semantic indexing (pLSI) [3]. In the context of text modeling, our model posits that each document is generated as a mixture of topics, where the continuous-valued mixture proportions are distributed as a latent Dirichlet random variable. Inference and learning are carried out efficiently via variational algorithms. We present empirical results on applications of this model to problems in text modeling, collaborative filtering, and text classification.



Dataset
collection of text documents
Topics
cluster of words
• words/tokens
Topic
frequency of words
Document
distribution of topics



A Dirichlet distribution parameter used for the generation of the topic distribution in the model.

A matrix that represents the distribution of words in each topic. Each row represents a topic, and each column represents a word in the vocabulary

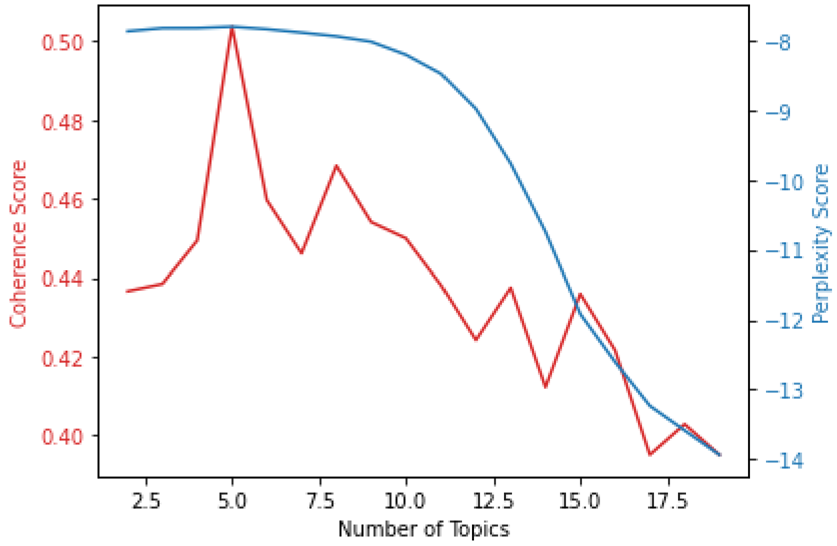
alpha: A hyperparameter that controls the distribution of topics in a document. A higher alpha value results in a document containing more topics.
Beta: A hyperparameter that controls the distribution of words in a topic. A higher beta value results in a topic containing more words.

How to find optimal number of topics?

Coherence score and **perplexity** are two common evaluation metrics used in topic modeling.

Coherence score measures how well the words in a topic relate to each other and how distinct they are from the words in other topics.

Perplexity measures the ability of the model to predict new documents. It calculates how well the model can predict the probability of unseen words in a held-out test set. A lower perplexity score indicates that the model is better at predicting unseen data.



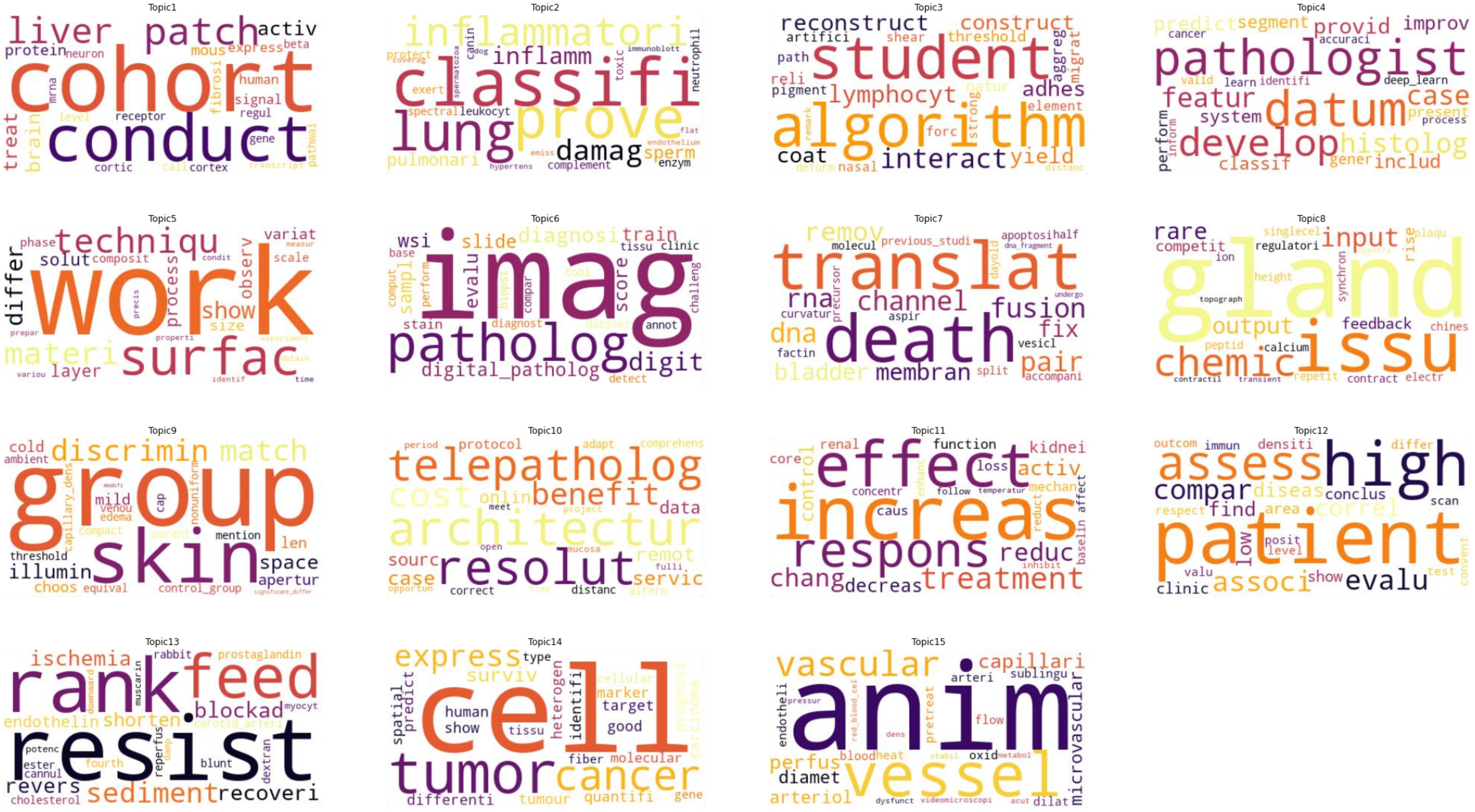
Both coherence score and perplexity are useful in evaluating the performance of a topic model. However, it's important to keep in mind that they don't necessarily reflect the real-world usefulness or relevance of the topics generated by the model. Therefore, it's always recommended to supplement quantitative evaluations with qualitative assessments by domain experts.

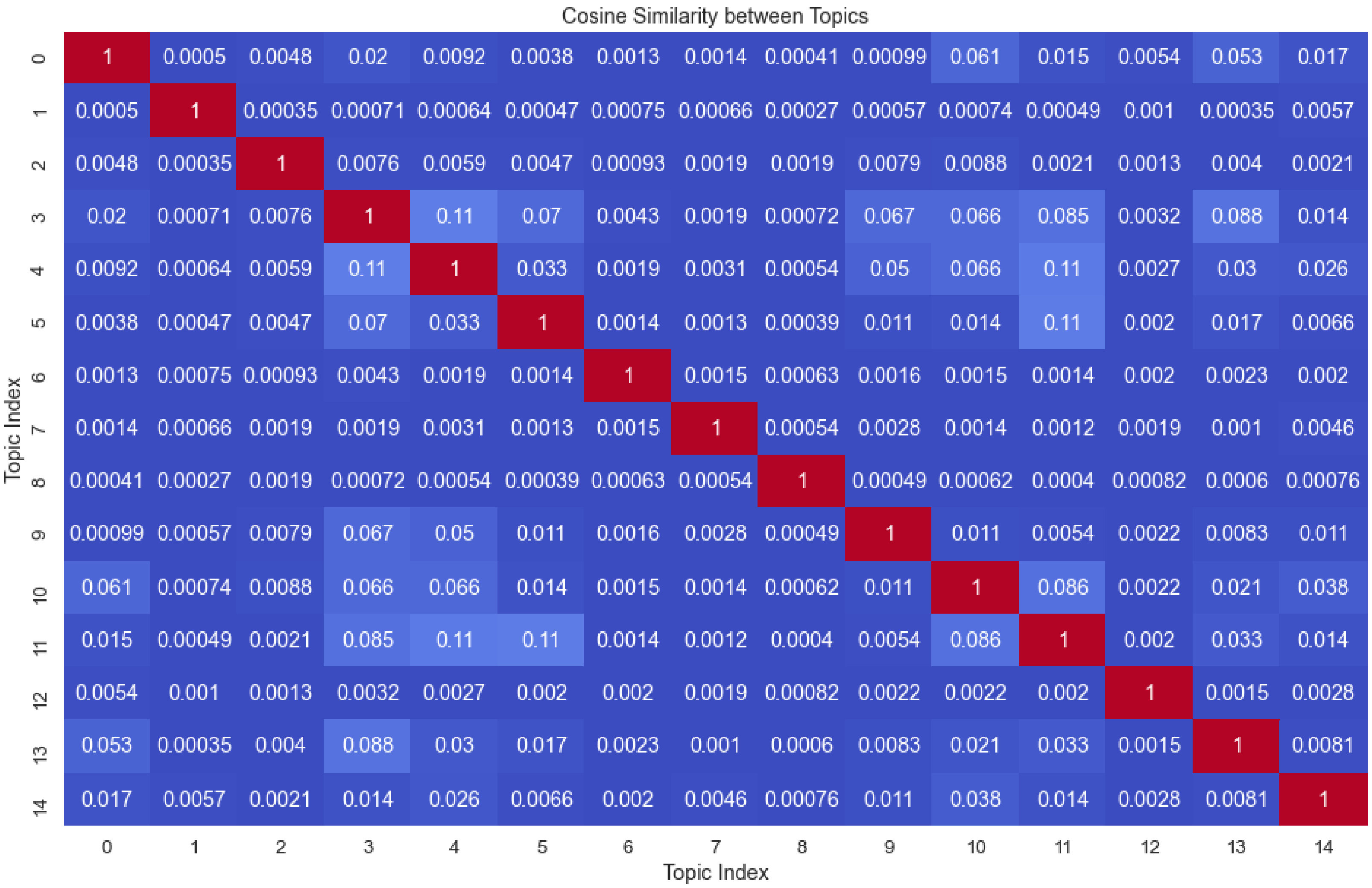
```python
# Train an LDA model with the optimal number of topics
lda_model = LdaModel(corpus=corpus,
                     id2word=dictionary,
                     num_topics=15,
                     chunksize=100,
                     passes=10,
                     alpha='auto',
                     per_word_topics=True)
for topic in lda_model.print_topics():
    print(topic)
```

id2word:mapping of the vocabulary IDs to the actual words in the corpus

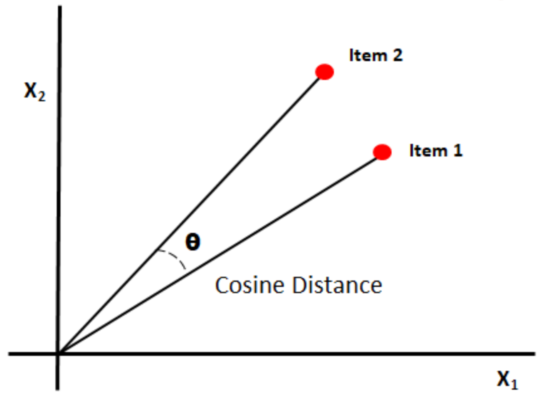
specifies the number of documents to be used in each training chunk

specifies the number of times the corpus will be processed during training


Topic1
cohort conduct liver patch activ protein neuron mous express beta human signal regul fibrosi mrna treat brain level receptor gene pathwai cortic cell cortex transcript
Topic2
inflammatori classifi prove lung inflamm damag protect coverag exert spermatozoa canin dog immunoblott toxic neutrophil spectral leukocyt pulmonari sperm endothelium flat emiss hypertens complement enzym
Topic3
reconstruct construct student algorithm artifici shear threshold path reli pigment lymphocyt natur adhes aggreg migrat element remark strong forc coat interact yield deform nasal distanc
Topic4
predict segment provid improv cancer accuraci pathologist featur datum case valid learn identifi deep_learn system present process perform inform develop histolog classif gener includ
Topic5
work techniqu surfac phase solut composit variat measur condit scale differ process observ show size precis prepar properti experiment materi layer obtain variou identif time
Topic6
imag patholog wsi slide diagnosi train tissu clinic base evalu score comput sampl perform biopsi compar tool challeng stain diagnost dataset annot digital_patholog digit detect
Topic7
translat death remov molecul previous_studi apoptosi half dna_fragment dayold rna channel fusion precursor curvatur aspir undergo dna fix factin vesicl bladder membran split pair accompani
Topic8
gland issu rare competit input regulatori singlecel plaqu rise ion spectra height synchron topograph output feedback peptid calcium chines chemic contractil transient repetit contract electr
Topic9
group skin cold discrimin match ambient capillary_dens modifi mild venou edema nonuniform cap compact parent mention threshold len illumin space apertur choos equival control_group significant_differ
Topic10
telepatholog cost architectur resolut period protocol adapt comprehens onlin benefit meet project data open mucosa remot sourc case fulli servic opportun correct time distanc altern
Topic11
effect increas respons renal function kidnei core control loss activ mechan concentr enhanc follow temperatur affect caus reduct reduc baselin inhibit chang decreas treatment
Topic12
assess high patient outcom immun densiti differ compar diseas conclus scan respect find area correl posit level low valu show test convent clinic associ evalu
Topic13
rank feed resist ischemia rabbit prostaglandin muscarin downward blockad myocyt endothelin shorten carotid_arteri potenc ester cannul fourth damp reperfus blunt dextran revers sediment recoveri cholesterol
Topic14
cell tumor express type surviv cellular prognosi marker target spatial predict human show tissu heterogen identifi good carcinoma fiber molecular cancer differenti tumour quantifi gene
Topic15
anim vessel vascular capillari arteri sublingu endotheli pressur red_blood_cel pretreat flow microvascular dens perfus bloodheat stabil oxid metabol diamet arteriol dysfunct videomicroscopi acut dilat



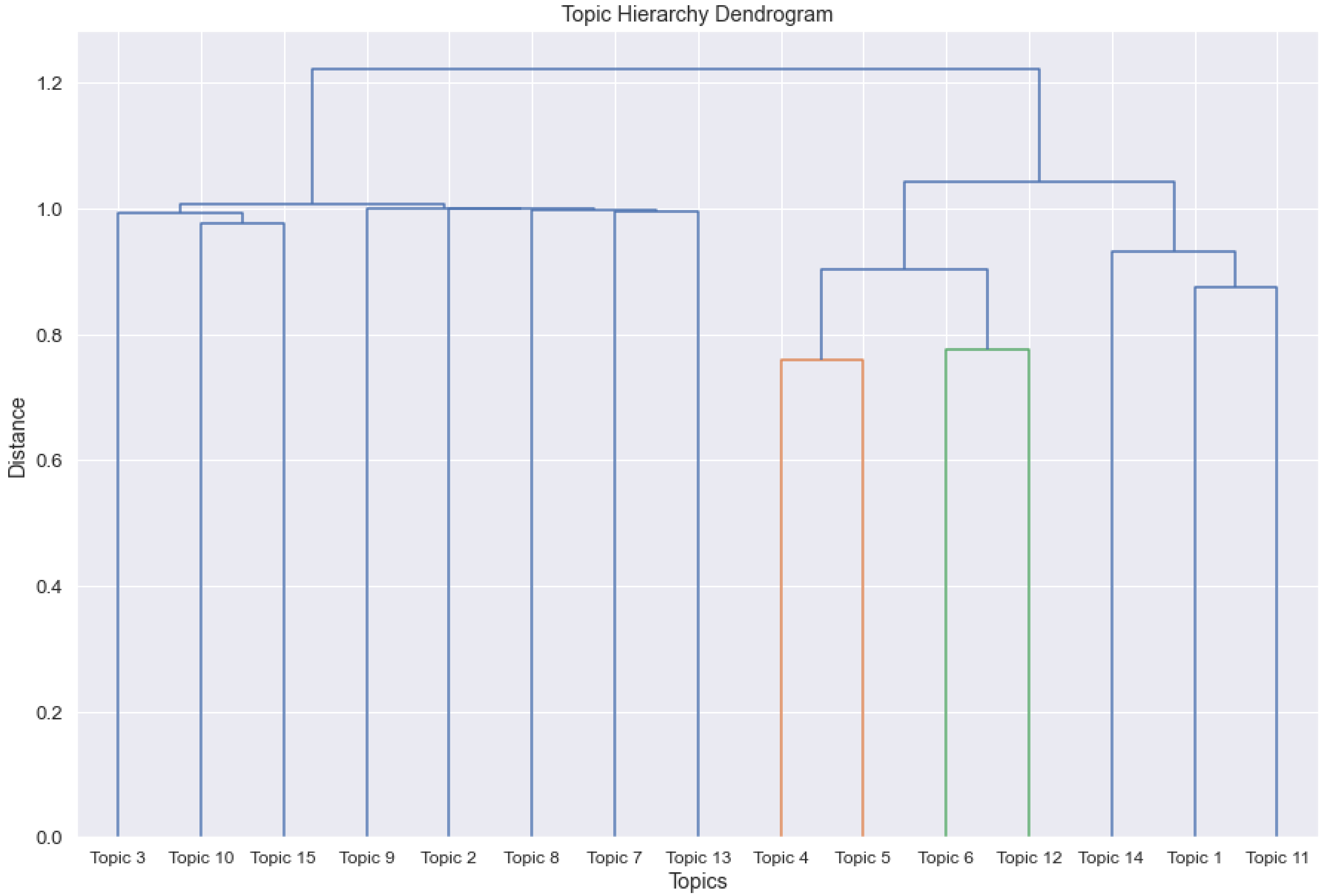
Cosine Distance/Similarity



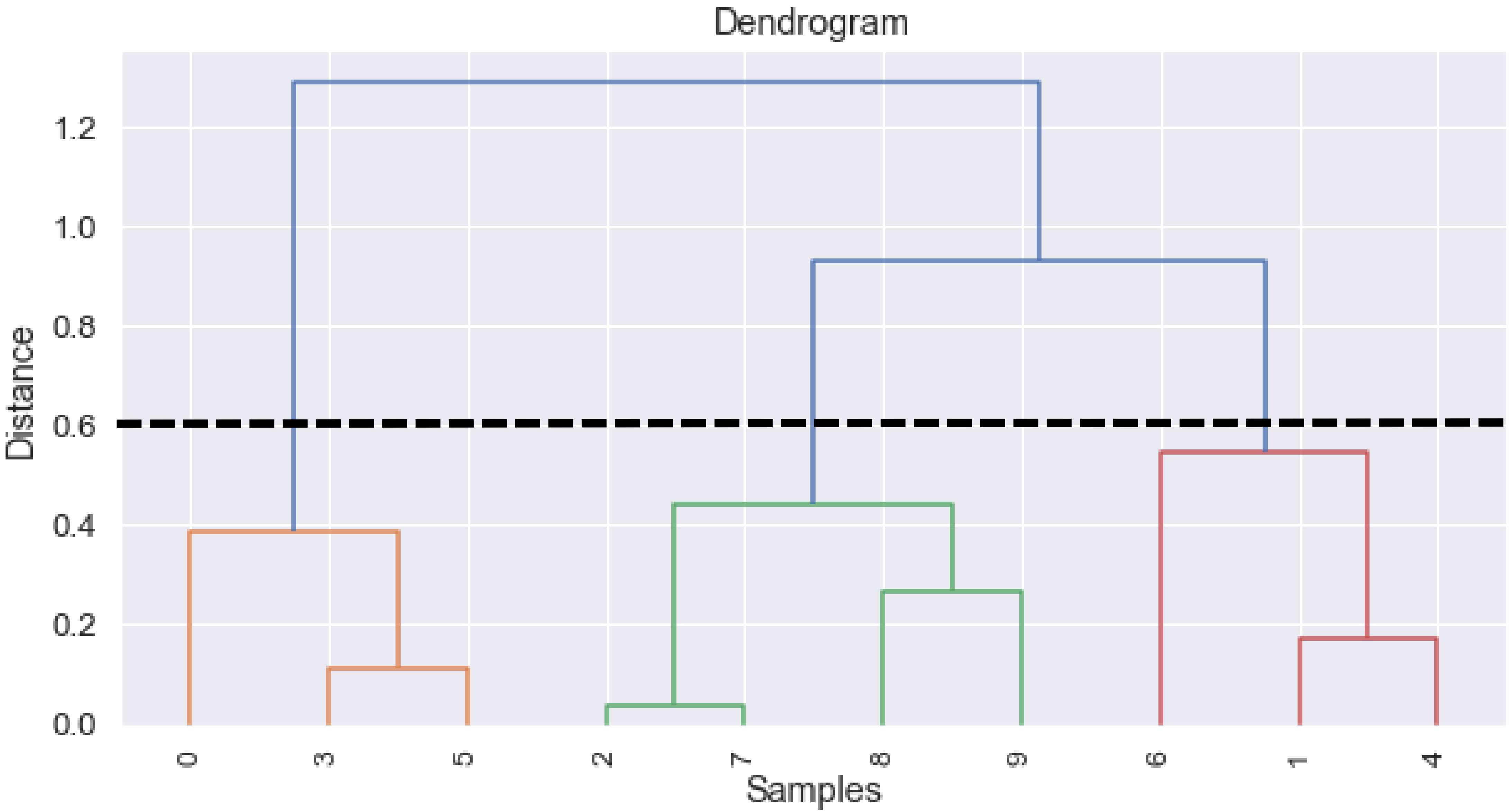
$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

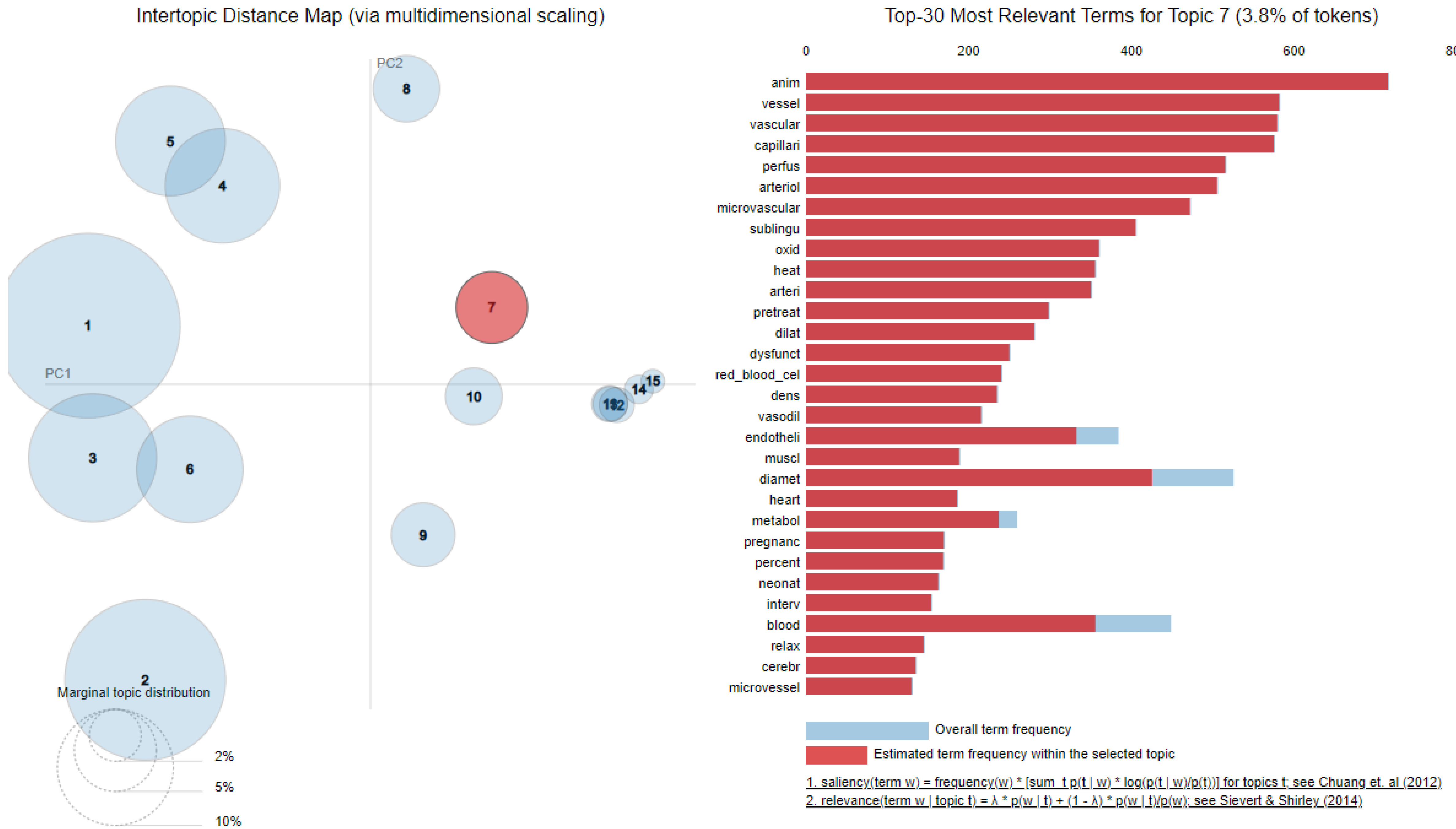
Topic 1: 0.3 * word1 + 0.4 * word2 + 0.3 * word3

the value 0.3 for word 1 in topic 1 represents the probability of word 1 being generated from topic 1. In other words, if we randomly select a word from a document that belongs to topic 1, there is a 30% chance that the word will be word 1.



The result of the hierarchical clustering analysis shows how the topics are related to each other based on their similarity.
The dendrogram visualizes the hierarchical structure of the topics, with each **node** representing a cluster of topics.
The horizontal lines in the dendrogram represent the distances between the clusters, with longer lines indicating greater distance. The vertical lines represent the topics or clusters. The height at which a vertical line intersects with a horizontal line represents the level of similarity between the topics or clusters.
The colors of the vertical lines represent the different clusters identified by the algorithm. The topics that are most similar to each other are grouped together in the same cluster. By analyzing the dendrogram, you can identify the different levels of similarity between the topics and how they are organized in a hierarchical structure.


Dendrogram
Distance
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
3
5
2
7
8
9
6
1
4
Samples


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
8
5
4
1
7
10
14
15
3
6
9
2
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 7 (3.8% of tokens)
0
200
400
600
800
anim
vessel
vascular
capillari
perfus
arteriol
microvascular
sublingu
oxid
heat
arteri
pretreat
dilat
dysfunct
red_blood_cel
dens
vasodil
endotheli
muscl
diamet
heart
metabol
pregnanc
percent
neonat
interv
blood
relax
cerebr
microvessel
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)