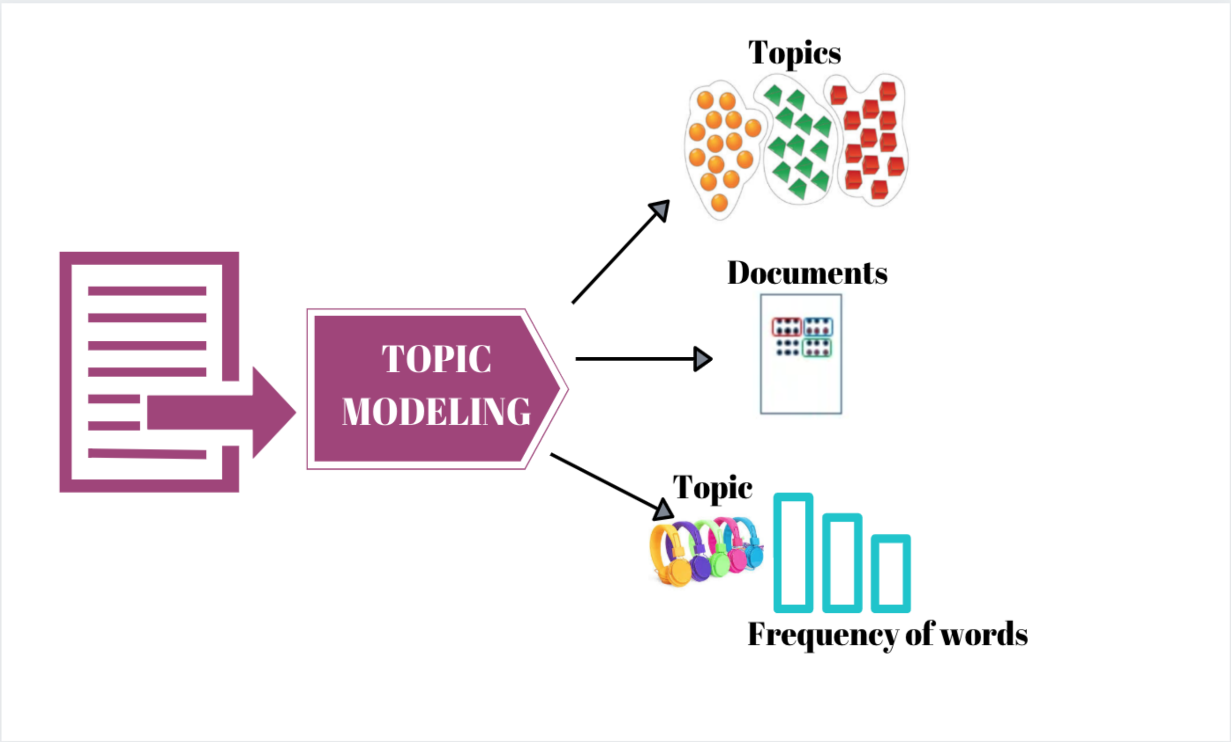
Exploring trends using Topic Modelling

Bachelor’s of Technology Project under the supervision of : **Professor Akhilesh Kumar**
(Department of Industrial and System Engineering)

By: Shubham Sonal (16NA10028)
(Department of Ocean Engineering and Naval Architecture)

INTRODUCTION

Topic modelling is an unsupervised machine learning technique that's capable of scanning a set of documents, detecting word and phrase patterns within them, and automatically clustering word groups and similar expressions that best characterize a set of documents.



TOPIC MODELLING

Topic Modelling methods tries to figure out which topic are present in the documents of the corpus and how strong its presence is.

Topic modelling refers to the process of dividing a corpus of documents in two:

1) A list of topic covered by the documents in the corpus.

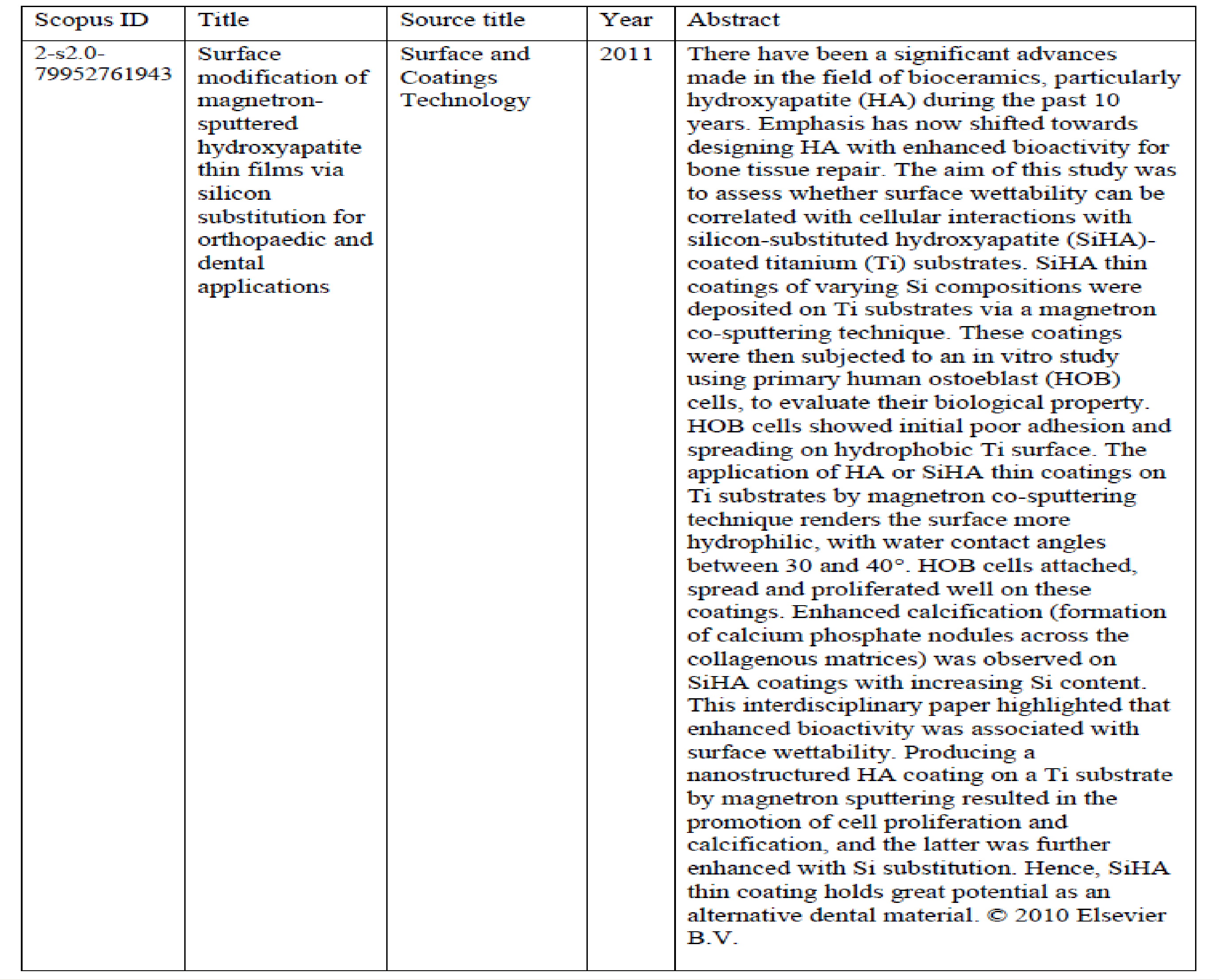
2) Several sets of documents from the corpus grouped by the topic they cover.

DATASET:

Data for this project is extracted using Scopus API. We have extracted data of 20,000 research papers from different journals.

6 Variables in the dataset are:

1) Scopus ID
2) Title
3) Source title
4) Year
5) Abstract



Scopus ID	Title	Source title	Year	Abstract
2-s2.0-79952761943	Surface modification of magnetron-sputtered hydroxyapatite thin films via silicon substitution for orthopaedic and dental applications	Surface and Coatings Technology	2011	There have been a significant advances made in the field of bioceramics, particularly hydroxyapatite (HA) during the past 10 years. Emphasis has now shifted towards designing HA with enhanced bioactivity for bone tissue repair. The aim of this study was to assess whether surface wettability can be correlated with cellular interactions with silicon-substituted hydroxyapatite (SiHA)-coated titanium (Ti) substrates. SiHA thin coatings of varying Si compositions were deposited on Ti substrates via a magnetron co-sputtering technique. These coatings were then subjected to an in vitro study using primary human ostoeblast (HOB) cells, to evaluate their biological property. HOB cells showed initial poor adhesion and spreading on hydrophobic Ti surface. The application of HA or SiHA thin coatings on Ti substrates by magnetron co-sputtering technique renders the surface more hydrophilic, with water contact angles between 30 and 40°. HOB cells attached, spread and proliferated well on these coatings. Enhanced calcification (formation of calcium phosphate nodules across the collagenous matrices) was observed on SiHA coatings with increasing Si content. This interdisciplinary paper highlighted that enhanced bioactivity was associated with surface wettability. Producing a nanostructured HA coating on a Ti substrate by magnetron sputtering resulted in the promotion of cell proliferation and calcification, and the latter was further enhanced with Si substitution. Hence, SiHA thin coating holds great potential as an alternative dental material. © 2010 Elsevier B.V.

Pre- Processing:

- Lowercase all the alphabets in the abstract
- Removal of punctuations {@ # $ % & ( ) > < ? / ; : “ ‘ | \ = -}
- Removal of stop words.
- Remove all non-keyboard characters.
- Divide the text data into sentence tokens. And those sentence tokens will further be divided into word tokens.
- Lemmatize the token

Feature Extraction:

1) Bag of words (Bow)

2) N-grams:

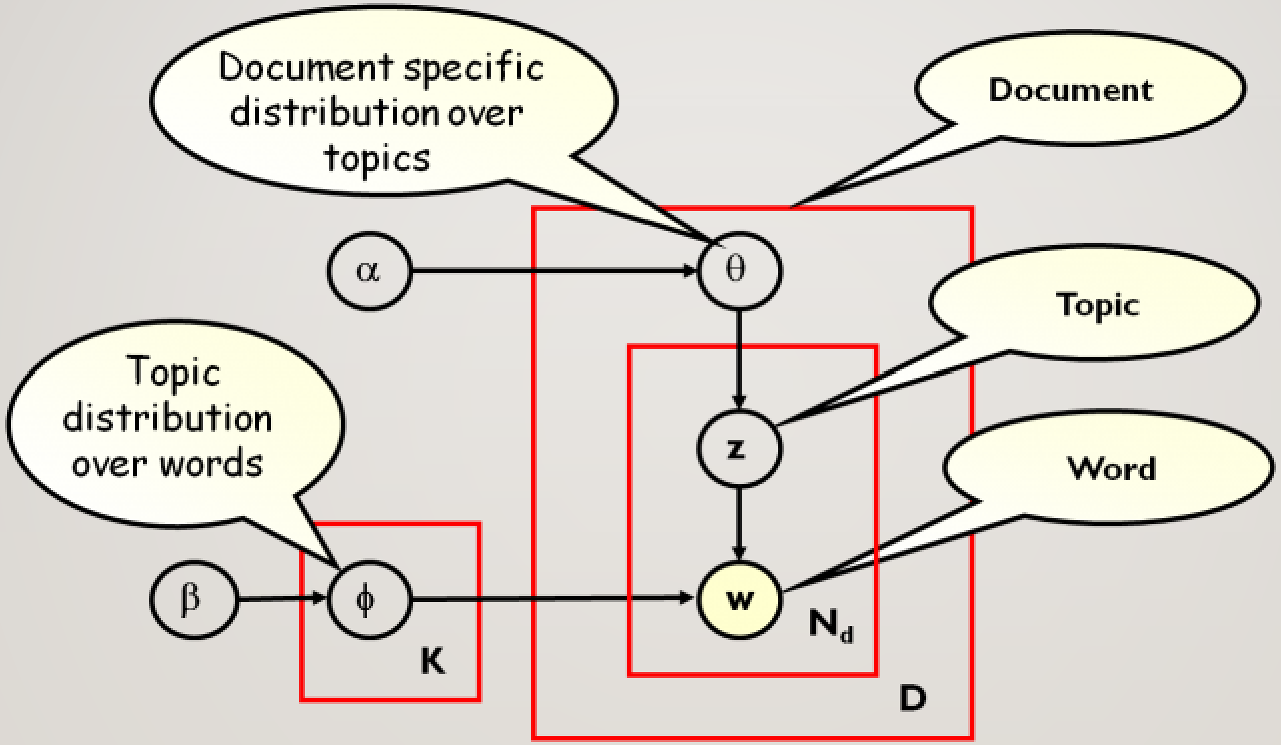
   Means sequence of n-words
   1-gram for tokens
   2-gram for token pairs

3) TF-IDF: Term frequency and Inverse Document frequency.

   TF-IDF is the measure of how important a term is.
   Term frequency TF = (No. of term t in the document)/ (No. of term in the document)
   IDF = log(number of documents/number of documents with term t)

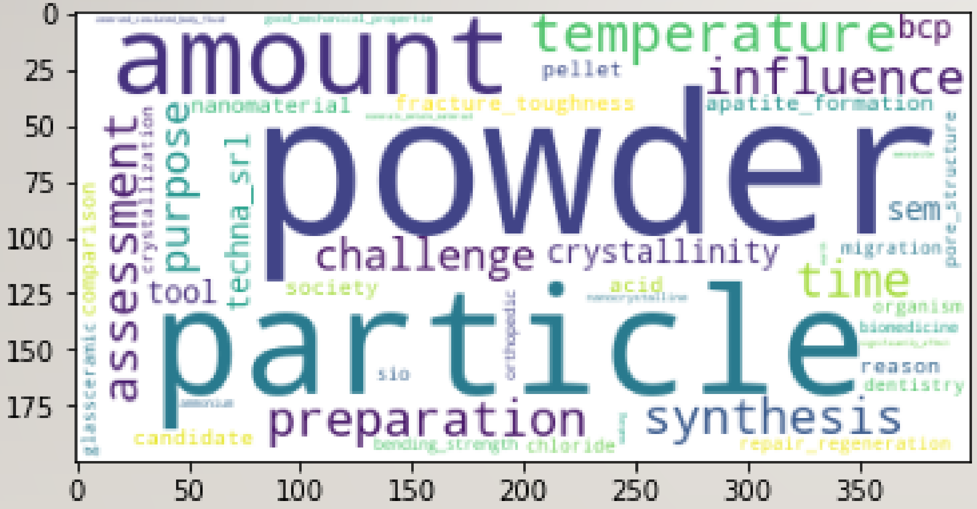
   TF-IDF = TF * IDF

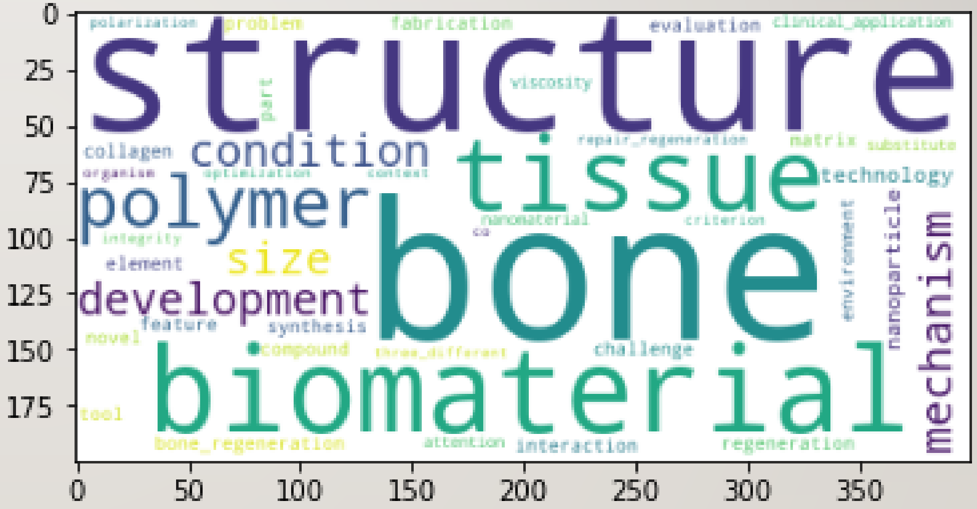
LDA- Latent Dirichlet Allocation model:


Document specific distribution over topics
Document
α
θ
Topic
Topic distribution over words
z
Word
β
ϕ
w
N_d
K
D

RESULT:

```
Perplexity: -14.751029581005083
Coherence Score: 0.4026414013401591
```





THANKYOU