**Natural Language Processing**

**NatGeo Magnum Opus**

**Hiring Ms.Lakshmi Sankaran**

**Team – 2**

**ISHAAN OHRI  
18BCE0265**

**Submitted to**

**Prof. Sharmila Banu K**

Problem Statement:

Assume you are a part of the NLP Tech team that works for a Publishing House. There is a shortlisted applicant (with her writing samples) for the Editor-in-chief position. How can you help the publishing house with the decision on hiring this applicant?

Pipeline:

* Pre-processing of the articles:
  + Stop word removal
  + Tokenization
  + Lemmatization
* Creating a corpus from all articles
* Bag-of-words creation
* Computing TF for all terms in all docs
* Computing IDF vector for all terms
* Computing TF-IDF
* Normalizing the TF-IDF
* Finding the Cosine similarity

Overview of all steps:

* Pre-processing of the articles:

The readArticle() function opens all the 4 articles and tokenizes the article using RegexpTokenizer, carries out WordNetLemmatizer and removes all stop words.

* Creating a corpus from all articles

Merge all words in all articles to form the corpus.

* Bag-of-words creation

Construct the bad of words for all documents and represent as a data frame.

* Computing TF for all terms in all docs

Compute the count of all words in all docs are represent the count in the form of a dictionary. Calculate the term frequency (TF) for all words in all docs using the function computeTF() and represent as a data frame.

* Computing IDF vector for all terms

Find the Inverse Document Frequency (IDF) of all words and represent as a data frame.

* Computing TF-IDF

Calculate the TF-IDF by multiplying the TF of word in documents with the corresponding IDF value and represent as a data frame.

* Normalizing the TF-IDF

Divide the TF-IDF value with the count of number of words in the document and represent as a data frame.

* Finding the Cosine similarity

Find the cosine similarity inorder to find the similarity of all documents. More is the value of cosine similarity, more is the similarity between the documents.