Agenda:

- ***Pre-processing***
- ***Convert-tokens***
- ***Model***
- ***Visualization***

Pre-Processing:

- **We have the data set that contains (id, author, url, content), so cause that topic modeling is a clustering problem so we need only the content column from the data set and drop any duplicates**

```
data_=data["content"].drop_duplicates()
```

- **After that we *remove stop words* from the data set that we collect**

```
def remove_stopwords(text : str):
    textArr = tokenizer.tokenize(text)
    rem_text = " ".join([word for word in textArr if word.lower() not in
stop_words ])
    return rem_text
```

- **After removing stop words we apply lemmatization (*The process of lemmatization aims to convert different inflected forms of a word into a single canonical form to facilitate analysis and understanding of the text*) on the tokens**

```
def lemmatization(texts,allowed_postags=['NOUN', 'ADJ']):
       output = []
       for sent in texts:
             doc = nlp(sent)
```

```
            output.append([token.lemma_ for token in doc if token.pos_ in
allowed_postags ])       return output
```

Convert-tokens:

- **After pre-processing the data we want to convert tokens to thing that computer and model can understand and train on it so we use a bag of words(*It represents a text document as an unordered collection, or "bag," of words, disregarding grammar and word order but considering their frequencies)***

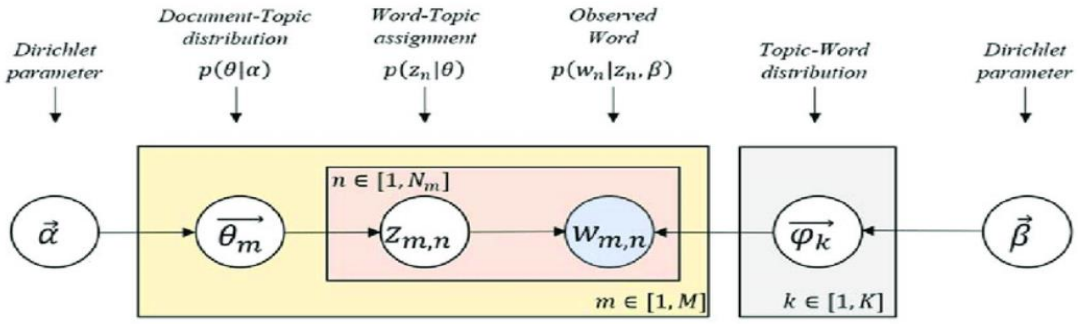
```
# Create a dictionary from the preprocessed data
dictionary = corpora.Dictionary(data_lemma)
# bag of words
corpus = [dictionary.doc2bow(doc) for doc in data_lemma]
```

Model:

- **After convert and encode the data, we split the *corpus* into train data and test data**

```
train_data,test_data=train_test_split(corpus, test_size=0.3,
random_state=42)
```

- **Then we take the train data and we will train the model, the model that we use *LDA (Latent diriclet Allocation) model***



Evaluations of the models

Model	*Coherence metric*	*Perplexity*
LDA with 30 topic	*55*	*-8.5*
LDA with 25 topic	*53*	*-8.3*
HDP	*42*	*-*

Visualization