WIKIPEDIA ARTICLES PREPROCESSING

Asadullah Mukhtar

L1F20BSCS0522

#Semester Project

(Programming for Big Data)

Presented to

Prof. Saeed Iqbal khattak

Step towards Project Compeletion

(1) Importing files

(2) Cleaning

- ↳ Puntuations & symbols Removing
- ↳ Stemming
- ↳ Lemitization.
- ↳ Stop words Removal

(3) Tokenization

(4) K Means Clustering

Cleaning:-

Removing Punctuation and Symbols from Data Set

→ " ! ( ) - [ ] { } ; : ' " \ < > etc.

These punctuations will be removed from Dataset.

For Example:-

Hi!!! , How are you?

↓

Hi How are you

Tokenization :-

The braking down of Sentences is called tokenization.

For example:-

My name is Asad.

↓

'My' , 'mame' , 'is' , 'Asad'

Stop Words Removal:-

In this step the stopwords like this, in, the, a, that, I, He, She etc are removed.

For Examples:-

I am a Student

↓

Student

Stemming:-

Chope the word.

→ It based on looking of word.

For Example:-

males → male

playing → Play

It also have some disadvantages.

·Ring → R

↑

Notmeaning ful

So we prefer Lemmitization.

Lemmitization:

is very efficient way to reduce a word into it's root form.

→ Stemming works on looking but lemmitization work on calculation.

In other words, resolving a word into dictionary form. (meaningful form).

For Example:-

cities → city

Languages → Language.

KMeans Clustering:-

As we are dealing with text. the KMeans Clustering somehow not different.

We make Clusters on the basis of relevency of Words.

We can also predict a string, by testing by KMEANS, belongs to what cluster.