**Generic LNP System and Preprocessing**

**1.3. Generic NLP system**

**Generic NLP System Block Diagram**

|  |
| --- |
|  |

**Pre-processing:**

**Tokenization:**

Tokenization is a process of converting sentence into a chain of words so that processing word by word can be easily performed. Here we use white space character for tokenization.

New York and rock ’n’ roll are individual words despite the fact that they contain spaces, but for many applications we’ll need to separate I’m into the two words I and am.

***New York and rock ’n’ roll are individual words despite the fact that they contain spaces, but for many applications we’ll need to separate I’m into the two words I and am.***

**Stop word removal:**

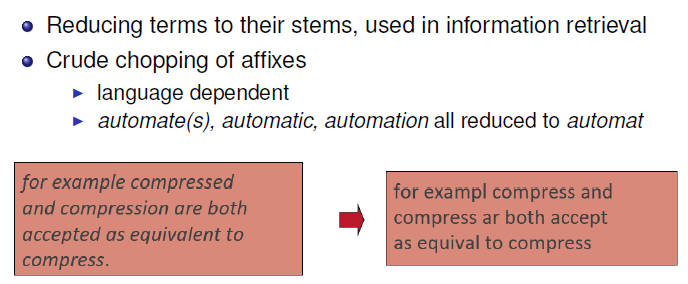
Stop words are the most frequently occurring words which slow down the processing of documents as these

words are irrelevant. Hence we remove the stop words to enhance the speed of searching. A corpus of stop words is used to filter out the stop words from the documents.



'their', 'then', 'not', 'ma', 'here', 'other', 'won', 'up', 'weren', 'being', 'we', 'those', 'an', 'them', 'which', 'him', 'so', 'yourselves', 'what', 'own', 'has', 'should', 'above', 'in', 'myself', 'against', 'that', 'before', 't', 'just', 'into', 'about', 'most', 'd', 'where', 'our', 'or', 'such', 'ours', 'of', 'doesn', 'further', 'needn', 'now', 'some', 'too', 'hasn', 'more', 'the', 'yours', 'her', 'below', 'same', 'how', 'very', 'is', 'did', 'you', 'his', 'when', 'few', 'does', 'down', 'yourself', 'i', 'do', 'both', 'shan', 'have', 'itself', 'shouldn', 'through', 'themselves', 'o', 'didn', 've', 'm', 'off', 'out', 'but', 'and', 'doing', 'any', 'nor', 'over', 'had', 'because', 'himself', 'theirs', 'me', 'by', 'she', 'whom', 'hers', 're', 'hadn', 'who', 'he', 'my', 'if', 'will', 'are', 'why', 'from', 'am', 'with', 'been', 'its', 'ourselves', 'ain', 'couldn', 'a', 'aren', 'under', 'll', 'on', 'y', 'can', 'they', 'than', 'after', 'wouldn', 'each', 'once', 'mightn', 'for', 'this', 'these', 's', 'only', 'haven', 'having', 'all', 'don', 'it', 'there', 'until', 'again', 'to', 'while', 'be', 'no', 'during', 'herself', 'as', 'mustn', 'between', 'was', 'at', 'your', 'were', 'isn', 'wasn'}

**Stemming:**



The stem is not necessarily the linguistic root of the word.

**Filtration & Script validation**

As presence of special characters in Devanagari documents degrades the performance, it needs to be removed. This removal of special characters from Devanagari script is called as filtration of document. Token creation of special characters and its recognition with UTF-8[12] is time consuming which leads to memory wastage. The special characters such as “ ” ‘ ’ , . / ? [ ] { } : ; \ | ~ ! @ # $ % ^ & \* ( ) \_ - = + < > are frequently used in many language scripts.

Script Validation is concerned with removal of non-Devanagari characters which is done by comparing with the UTF-8 list[12]. These characters will not contribute towards final result. To perform filtration and script validation operation we have used Unicode values called UTF-8 for Devanagari script document. We compared UTF-8 list with each character of each token, if match found the character is valid and allowed otherwise removed from the document. The aim of this phase is to maintain pure Devanagari script document as input to Morphological Analyzer.

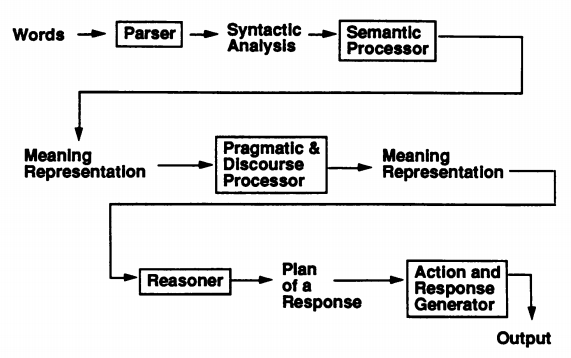
**Lemmatization:**

Lemmatization reduces words to their base word, which is linguistically correct lemmas. It transforms root word with the use of vocabulary and morphological analysis. Lemmatization is usually more sophisticated than stemming. Stemmer works on an individual word without knowledge of the context. For example, The word "better" has "good" as its lemma. This thing will miss by stemming because it requires a dictionary look-up.

Reduce inflections or variant forms to base form:

am, are, is ! be

car, cars, car’s, cars’ ! car



**Explain each steps in one or two more lines[covered in stages of NLP].**