50 plus article hain or human generated summarizes hain

articles stemmed, segmentized hain

*>>Baki groups b kr rhy text extraction tw yeh different kaise hoga?*

**Hamare pass corpus ha well defined for summazition purpose, hmari preprocessing techniques change hon gae**

**stem , segmentize etc**

**hamra algo different hon gy**

**hmara evaluation method rouge ho ga**

Urdu summary corpus consists of 50 articles collected from various sources. From the original HTML documents only unformatted content text was kept, removing all other things. We provide abstractive summaries of these 50 articles. After normalization, we further applied different NLP tools on the articles to generate part-of-speech tagged, morphologically analyzed, lemmatized and stemmed articles.

Fifty articles for Urdu Summary Corpus (USC) were collected from various online sources mainly news portals and blogs. The sources were chosen based on the following

merits:

1. They contain real text as written by the native speakers.

2. Authors of different backgrounds have written the text.

3. Getting permission from authors to re-distribute the

texts was easy as compared to print

Meetings one ..

**Obstraction based learning:**

The learning on which a person describes the text in his/her on words.

**Extraction based learning:**

The learning on which a person describes the text from the sentence or words taken from text itself.

**Tokenization:**

Like in english tokenization in urdu is not done on the basis of splitting by space.

**Stemming:**

Karachi to kara

**Lemmatization:**

Karachi to karachi

**-> Tokenization from the segmentized form**