Kannada article Summarizer

Team 39 - SammuraiZers:

Teammates

1. Virupaksha Swamy (USCID - 8970 8523 14) vswamy@usc.edu
2. Swaroop Venu (USCID - 2504 7966 30) venu@usc.edu
3. Nitish Mohan Shandilya (USCID - 9604 8761 22) nshandil@usc.edu
4. Achyut Kulkarni (USCID - 9589 8079 26) agkulkar@usc.edu

Project Source : https://github.com/vswamy/summarizer/
Annotator Tool : http://www-scf.usc.edu/~venu/annotator.html
Annotator Source Code : https://github.com/swaroopv/annotation-helper

Introduction:

With the abundance of articles available on the internet, it is generally tedious to go through the entire article to find how relevant the article is for the user. With this, we want to implement an art of intrinsic extraction which highlights the most important sentences in a document. We want to specifically summarize articles related to the most important topics from Kannada newspapers.

The primary reason for this project is the non-existence of a working implementation of a summarizer for Kannada language, a corpus to act as a baseline for summarization and an efficient stemmer that aids in a better canonicalization of the tokens.

Creation of corpus itself is a challenging task. We were able to obtain corpuses of 5000, 5000 and 12,000 articles related to state news, cinema and sports topics from the Kannada e-newspaper “Udayavani” and “Prajavani”. Another reason that makes the task challenging is the requirement of a stemmer algorithm to find the root of a word. We have implemented a statistical stemmer algorithm[1]. The complexity of developing stemmer for Dravidian language like Kannada is comparatively higher than the other languages like English. Most of the words may change spelling when stems are inflected. In the purview of all the considerations, the stemmer was found to stem the most commonly used Kannada words with fair accuracy.

The existing work involves the proposal and theoretical explanation of TF-IDF approaches. But there has neither been a corpus nor has there been a working implementation of such approaches. Also, there has been no showcase of good preprocessing techniques such as stemming that aid in better summarization of the articles. But, our approach involves the creation of a corpus, incorporating TF-IDF, GSS coefficients and sentence weight and also in implementing good preprocessing techniques.

Method:

Materials:

1. NLTK - NLP toolkit
2. Scrapy - A web Crawler
3. Python
4. Javascript, HTML and Bootstrap framework
5. Web Archives of Kannada newspapers - www.prajavani.net www.udayavani.net
6. Rouge - tool link

Procedure:
The analysis contains the following steps:

Building the corpus :
Using Scrapy , the newspapers mentioned above are scraped and articles are parsed and stored in one file in a JSON format. The file has all articles in dictionary format having fields - URL, Content, Title for each article. The articles scraped are from State, Cinema and Sports categories. A corpus of around 22,000 articles has been created.

Annotation of corpus :
A simple and scalable tool is built to provide UI to user to annotate the corpus. Each article is presented in sentence by sentence format to user to choose the best sentences for summarization. Once annotated, the corpus can be downloaded and can be used for training the model. Around 600 articles have been annotated and used as test data.

Stemmer:
Method of affix removal to reduce a word to its stem form is called stemming. To stem a given word following steps are followed:

1. If the word length is less than or equal to 3, the word itself is the root. In kannada, words are not in inflected form.
2. The word is split into all possible prefix + suffix combinations. For each such prefix-suffix combination, frequency of prefix and suffix is found on the entire udayavani.com corpus.
3. For each prefix - suffix combination, the pair that maximizes P is chosen as the correct split and the prefix is taken as stem.
4. _P = ((string length of prefix word) * log(frequency of the prefix) + (string length of suffix word) * log(frequency of the suffix))_
5. For example, for the word ಹೋಗುತ್ತೇನೆ,

   ಹೋಗ + ುತ್ತೇನೆ = 70.28886341836466

   ಹೋಗು + ತ್ತೇನೆ = 67.53406051597022

   ಹೋಗುತ + ್ತೇನೆ = 64.7574719443816

   ಹೋಗುತ್ + ತೇನೆ = 64.56159434985277

   ಹೋಗುತ್ತ + ೇನೆ = 66.49605238613647

   ಹೋಗುತ್ತೇ + ನೆ = 51.945908102411316

   ಹೋಗುತ್ತೇನ + ೆ = 45.631472161108285

   stem chosen = ಹೋಗ

TF-IDF and GSS coefficients:
The entire corpus of all the articles of all categories is used as input to calculate the TF-IDF values. The two terms are calculated as follows :

TF= frequency of a term in a document / Number of terms in a given document .... (1) IDF= Log /O (N / n) .... (2)

Where 'N' is the total number of documents indexed across all categories and 'n' is the number of documents containing a particular term

GSS (Galavotti - Sebastiani - Simi) coefficient is a feature selection technique used as the relevance measure in our case.
Given a word w and category c it is defined as:

$$f(w, c) = p(w, c) * p(w' , c') - p(w '' c) * p(w, c') . \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

Where, p (w, c) is the probability that a document contains word 'w' and belongs to category 'c'. p (w' , c') is the probability that a document does not contain word 'w' and does not belong to category 'c'. p (w' , c) is the probability that a document does not contain word 'w' and belongs to category 'c'. p (w , c') is the probability that a document contains word 'w' and does not belong to category 'c'.

First we create an IDF, Naive Bayes model for all 22,000 articles. We take test data and find out the category of each article using *Naive Bayes' Algorithm*. Each word in the article is taken and tf is calculated. Once the TF is calculated, Term weightage is calculated using

*word weight = Term Frequency * Inverse Document Frequency + GSS * Term Frequency .... (4)*

The weight of a sentence is the sum of weights of all words. We calculate sentence rank Of each sentence using :

Rank of sentence = sum of values of words in the sentence / Total number of words in a sentence .... (5)

Finally we sort the sentences according to descending order of ranks and choose top “20%” of the sentences with max=6 and min=4 sentences and also preserve the relative positions of sentences in the result.

Evaluation:

The model we generated to display the summary of an article is evaluated using Inter-Annotator agreement. The inter-annotator agreement is an agreement on between humans on the output of an article. We have annotated around 600 articles for evaluation and 200 of them are unique.

Location of annotated corpus

A standard evaluation method known as Rouge Evaluation Method is used for the evaluation of Inter-Annotator results with each other and for the evaluation of Human summaries with the system generated summaries.
In rogue model, two folders : system folder and reference folder exist. In system folder all the results of machine generated are stored and human output are stored and in reference folder, all the human annotators summaries to be compared are kept and the Rouge script is run to compare the results.

Inter-Annotator agreement(IAA) : We keep one person’s annotations of each article in system and rest of the human annotations of same article are kept in reference folder and repeat the procedure for all annotators .
All the results which have a more than 50% IAA have been selected to test out the machine generated output.

Human Vs Machine Summary Evaluation : We keep the output of machine generated in Systems folder for an article whose IAA > 50% and in reference folder, the same article annotated by humans are kept . Finally the results generated by the rouge technique are put in a .CSV file which has Recall, Precision and F1 score for each article.

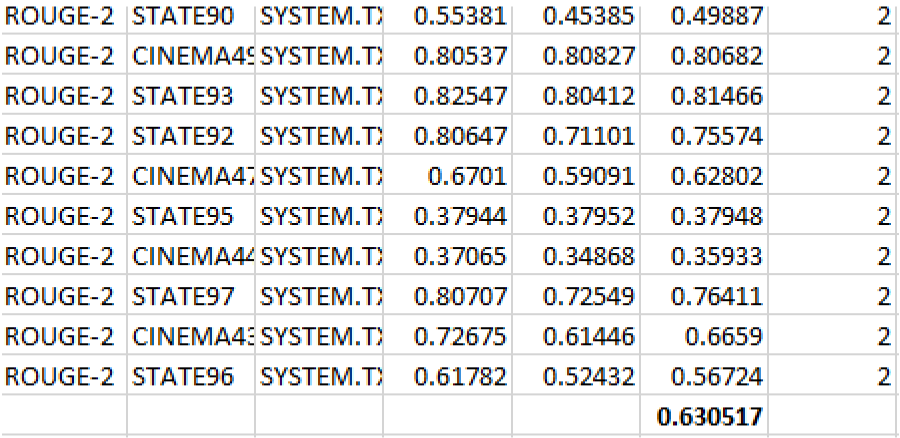
A cumulative performance of our model is calculated by taking the average of results of F1 score.

Results:

Our system performs with considerable accuracy level. We are using TF-IDF and GSS coefficient. TF-IDF makes our model to select those terms whose importance increases proportionally to the number of times the term appears in the document but is offset by the frequency of the term in the corpus. GSS coefficient aids in selecting terms relevant to a specific category. With these two scores, the term weightage and finally the sentence weightage is computed. Therefore, our model best summarizes those articles, which have terms that were previously seen during the training phase.

We supplied stop words list to Rouge tool.
We made two evaluation one without stemmer and one with stemmer.

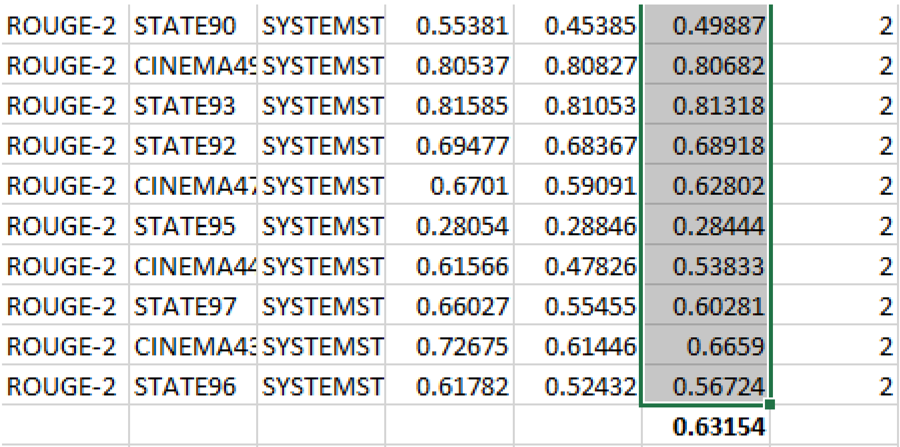
Result without stemmer



ROUGE-2	STATE90	SYSTEM.T	0.55381	0.45385	0.49887	2
ROUGE-2	CINEMA49	SYSTEM.T	0.80537	0.80827	0.80682	2
ROUGE-2	STATE93	SYSTEM.T	0.82547	0.80412	0.81466	2
ROUGE-2	STATE92	SYSTEM.T	0.80647	0.71101	0.75574	2
ROUGE-2	CINEMA47	SYSTEM.T	0.6701	0.59091	0.62802	2
ROUGE-2	STATE95	SYSTEM.T	0.37944	0.37952	0.37948	2
ROUGE-2	CINEMA44	SYSTEM.T	0.37065	0.34868	0.35933	2
ROUGE-2	STATE97	SYSTEM.T	0.80707	0.72549	0.76411	2
ROUGE-2	CINEMA43	SYSTEM.T	0.72675	0.61446	0.6659	2
ROUGE-2	STATE96	SYSTEM.T	0.61782	0.52432	0.56724	2
					0.630517	

We got Rouge F1-score of 63.05%

Result with Stemmer

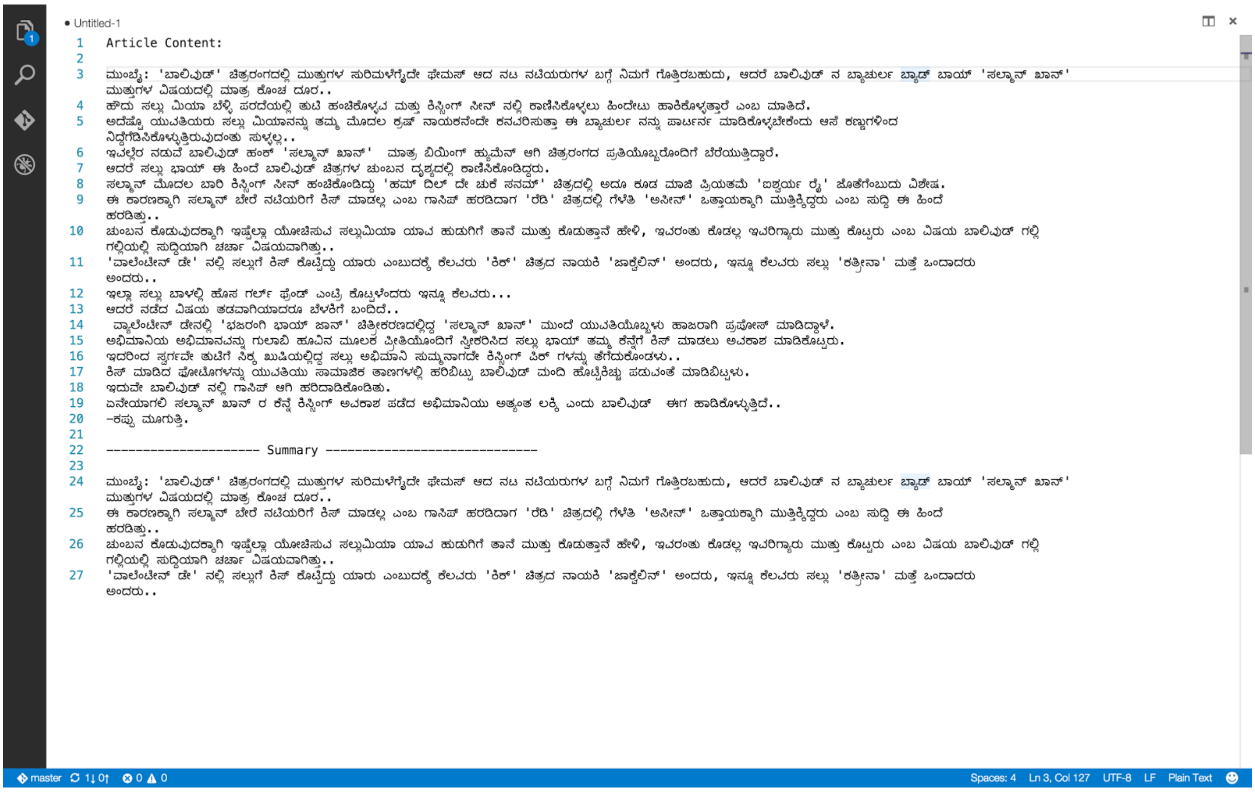


ROUGE-2	STATE90	SYSTEMST	0.55381	0.45385	0.49887	2
ROUGE-2	CINEMA49	SYSTEMST	0.80537	0.80827	0.80682	2
ROUGE-2	STATE93	SYSTEMST	0.81585	0.81053	0.81318	2
ROUGE-2	STATE92	SYSTEMST	0.69477	0.68367	0.68918	2
ROUGE-2	CINEMA47	SYSTEMST	0.6701	0.59091	0.62802	2
ROUGE-2	STATE95	SYSTEMST	0.28054	0.28846	0.28444	2
ROUGE-2	CINEMA44	SYSTEMST	0.61566	0.47826	0.53833	2
ROUGE-2	STATE97	SYSTEMST	0.66027	0.55455	0.60281	2
ROUGE-2	CINEMA43	SYSTEMST	0.72675	0.61446	0.6659	2
ROUGE-2	STATE96	SYSTEMST	0.61782	0.52432	0.56724	2
					0.63154	

With stemmer we got minor improvement 63.15%

There was not a significant improvement after stemming. This might be because of the technique we employed for stem creation i.,e it is completely statistical based stem creation and solely depends on the corpus. We could have used stemming using rule based or wordnet based mechanism but however this language is extremely complex with inflectional words and stemming itself will take a large part which would deviate from the goal of our project given the scenario .

Sample Result:



```
Article Content:

ಮುಂಬೈ: 'ಬಾಲಿವುಡ್' ಚಿತ್ರರಂಗದಲ್ಲಿ ಮುತ್ತುಗಳ ಸುರಿಮಳೆಗೈದೇ ಫೇಮಸ್ ಆದ ನಟ ನಟಿಯರುಗಳ ಬಗ್ಗೆ ನಿಮಗೆ ಗೊತ್ತಿರಬಹುದು, ಆದರೆ ಬಾಲಿವುಡ್ ನ ಬ್ಯಾಚುರ್ಲ ಬ್ಯಾಡ್ ಬಾಯ್ 'ಸಲ್ಮಾನ್ ಖಾನ್'
ಮುತ್ತುಗಳ ವಿಷಯದಲ್ಲಿ ಮಾತ್ರ ಕೊಂಚ ದೂರ..
ಹೌದು ಸಲ್ಲು ಮಿಯಾ ಬೆಳ್ಳಿ ಪರದೆಯಲ್ಲಿ ತುಟಿ ಹಂಚಿಕೊಳ್ಳವ ಮತ್ತು ಕಿಸ್ಸಿಂಗ್ ಸೀನ್ ನಲ್ಲಿ ಕಾಣಿಸಿಕೊಳ್ಳಲು ಹಿಂದೇಟು ಹಾಕಿಕೊಳ್ಳತ್ತಾರೆ ಎಂಬ ಮಾತಿದೆ.
ಅದೆಷ್ಟೊ ಯುವತಿಯರು ಸಲ್ಲು ಮಿಯಾನನ್ನು ತಮ್ಮ ಮೊದಲ ಕ್ರಷ್ ನಾಯಕನೆಂದೇ ಕನವರಿಸುತ್ತಾ ಈ ಬ್ಯಾಚುರ್ಲ ನನ್ನು ಪಾರ್ಟರ್ನ ಮಾಡಿಕೊಳ್ಳಬೇಕೆಂದು ಆಸೆ ಕಣ್ಣುಗಳಿಂದ
ನಿದ್ದೆಗೆಡಿಸಿಕೊಳ್ಳುತ್ತಿರುವುದಂತು ಸುಳ್ಳಲ್ಲ..
ಇವಲ್ಲೆರ ನಡುವೆ ಬಾಲಿವುಡ್ ಹಂಕ್ 'ಸಲ್ಮಾನ್ ಖಾನ್'  ಮಾತ್ರ ಬಿಯಿಂಗ್ ಹ್ಯುಮೆನ್ ಆಗಿ ಚಿತ್ರರಂಗದ ಪ್ರತಿಯೊಬ್ಬರೊಂದಿಗೆ ಬೆರೆಯುತ್ತಿದ್ದಾರೆ.
ಆದರೆ ಸಲ್ಲು ಭಾಯ್ ಈ ಹಿಂದೆ ಬಾಲಿವುಡ್ ಚಿತ್ರಗಳ ಚುಂಬನ ದೃಶ್ಯದಲ್ಲಿ ಕಾಣಿಸಿಕೊಂಡಿದ್ದರು.
ಸಲ್ಮಾನ್ ಮೊದಲ ಬಾರಿ ಕಿಸ್ಸಿಂಗ್ ಸೀನ್ ಹಂಚಿಕೊಂಡಿದ್ದು 'ಹಮ್ ದಿಲ್ ದೇ ಚುಕೆ ಸನಮ್' ಚಿತ್ರದಲ್ಲಿ ಅದೂ ಕೂಡ ಮಾಜಿ ಪ್ರಿಯತಮೆ 'ಐಶ್ವರ್ಯ ರೈ' ಜೊತೆಗೆಂಬುದು ವಿಶೇಷ.
ಈ ಕಾರಣಕ್ಕಾಗಿ ಸಲ್ಮಾನ್ ಬೇರೆ ನಟಿಯರಿಗೆ ಕಿಸ್ ಮಾಡಲ್ಲ ಎಂಬ ಗಾಸಿಪ್ ಹರಡಿದಾಗ 'ರೆಡಿ' ಚಿತ್ರದಲ್ಲಿ ಗೆಳೆತಿ 'ಅಸೀನ್' ಒತ್ತಾಯಕ್ಕಾಗಿ ಮುತ್ತಿಕ್ಕಿದ್ದರು ಎಂಬ ಸುದ್ದಿ ಈ ಹಿಂದೆ
ಹರಡಿತ್ತು..
ಚುಂಬನ ಕೊಡುವುದಕ್ಕಾಗಿ ಇಷ್ಟೆಲ್ಲಾ ಯೋಚಿಸುವ ಸಲ್ಲುಮಿಯಾ ಯಾವ ಹುಡುಗಿಗೆ ತಾನೆ ಮುತ್ತು ಕೊಡುತ್ತಾನೆ ಹೇಳಿ, ಇವರಂತು ಕೊಡಲ್ಲ ಇವರಿಗ್ಯಾರು ಮುತ್ತು ಕೊಟ್ಟರು ಎಂಬ ವಿಷಯ ಬಾಲಿವುಡ್ ಗಲ್ಲಿ
ಗಲ್ಲಿಯಲ್ಲಿ ಸುದ್ದಿಯಾಗಿ ಚರ್ಚಾ ವಿಷಯವಾಗಿತ್ತು..
'ವಾಲೆಂಟೀನ್ ಡೇ' ನಲ್ಲಿ ಸಲ್ಲುಗೆ ಕಿಸ್ ಕೊಟ್ಟಿದ್ದು ಯಾರು ಎಂಬುದಕ್ಕೆ ಕೆಲವರು 'ಕಿಕ್' ಚಿತ್ರದ ನಾಯಕಿ 'ಜಾಕ್ವೆಲಿನ್' ಅಂದರು, ಇನ್ನೂ ಕೆಲವರು ಸಲ್ಲು 'ಕತ್ರೀನಾ' ಮತ್ತೆ ಒಂದಾದರು
ಅಂದರು..
ಇಲ್ಲಾ ಸಲ್ಲು ಬಾಳಲ್ಲಿ ಹೊಸ ಗರ್ಲ್ ಫ್ರೆಂಡ್ ಎಂಟ್ರಿ ಕೊಟ್ಟಳೆಂದರು ಇನ್ನೂ ಕೆಲವರು...
ಆದರೆ ನಡೆದ ವಿಷಯ ತಡವಾಗಿಯಾದರೂ ಬೆಳಕಿಗೆ ಬಂದಿದೆ..
 ವ್ಯಾಲೆಂಟೇನ್ ಡೇನಲ್ಲಿ 'ಭಜರಂಗಿ ಭಾಯ್ ಜಾನ್' ಚಿತ್ರೀಕರಣದಲ್ಲಿದ್ದ 'ಸಲ್ಮಾನ್ ಖಾನ್' ಮುಂದೆ ಯುವತಿಯೊಬ್ಬಳು ಹಾಜರಾಗಿ ಪ್ರಪೋಸ್ ಮಾಡಿದ್ದಾಳೆ.
ಅಭಿಮಾನಿಯ ಅಭಿಮಾನವನ್ನು ಗುಲಾಬಿ ಹೂವಿನ ಮೂಲಕ ಪ್ರೀತಿಯೊಂದಿಗೆ ಸ್ವೀಕರಿಸಿದ ಸಲ್ಲು ಭಾಯ್ ತಮ್ಮ ಕೆನ್ನೆಗೆ ಕಿಸ್ ಮಾಡಲು ಅವಕಾಶ ಮಾಡಿಕೊಟ್ಟರು.
ಇದರಿಂದ ಸ್ವರ್ಗವೇ ತುಟಿಗೆ ಸಿಕ್ಕ ಖುಷಿಯಲ್ಲಿದ್ದ ಸಲ್ಲು ಅಭಿಮಾನಿ ಸುಮ್ಮನಾಗದೇ ಕಿಸ್ಸಿಂಗ್ ಪಿಕ್ ಗಳನ್ನು ತೆಗೆದುಕೊಂಡಳು..
ಕಿಸ್ ಮಾಡಿದ ಫೋಟೊಗಳನ್ನು ಯುವತಿಯು ಸಾಮಾಜಿಕ ತಾಣಗಳಲ್ಲಿ ಹರಿಬಿಟ್ಟು ಬಾಲಿವುಡ್ ಮಂದಿ ಹೊಟ್ಟೆಕಿಚ್ಚು ಪಡುವಂತೆ ಮಾಡಿಬಿಟ್ಟಳು.
ಇದುವೇ ಬಾಲಿವುಡ್ ನಲ್ಲಿ ಗಾಸಿಪ್ ಆಗಿ ಹರಿದಾಡಿಕೊಂಡಿತು.
ಏನೇಯಾಗಲಿ ಸಲ್ಮಾನ್ ಖಾನ್ ರ ಕೆನ್ನೆ ಕಿಸ್ಸಿಂಗ್ ಅವಕಾಶ ಪಡೆದ ಅಭಿಮಾನಿಯು ಅತ್ಯಂತ ಲಕ್ಕಿ ಎಂದು ಬಾಲಿವುಡ್  ಈಗ ಹಾಡಿಕೊಳ್ಳುತ್ತಿದೆ..
-ಕಪ್ಪು ಮೂಗುತ್ತಿ.

-------------------- Summary ----------------------------

ಮುಂಬೈ: 'ಬಾಲಿವುಡ್' ಚಿತ್ರರಂಗದಲ್ಲಿ ಮುತ್ತುಗಳ ಸುರಿಮಳೆಗೈದೇ ಫೇಮಸ್ ಆದ ನಟ ನಟಿಯರುಗಳ ಬಗ್ಗೆ ನಿಮಗೆ ಗೊತ್ತಿರಬಹುದು, ಆದರೆ ಬಾಲಿವುಡ್ ನ ಬ್ಯಾಚುರ್ಲ ಬ್ಯಾಡ್ ಬಾಯ್ 'ಸಲ್ಮಾನ್ ಖಾನ್'
ಮುತ್ತುಗಳ ವಿಷಯದಲ್ಲಿ ಮಾತ್ರ ಕೊಂಚ ದೂರ..
ಈ ಕಾರಣಕ್ಕಾಗಿ ಸಲ್ಮಾನ್ ಬೇರೆ ನಟಿಯರಿಗೆ ಕಿಸ್ ಮಾಡಲ್ಲ ಎಂಬ ಗಾಸಿಪ್ ಹರಡಿದಾಗ 'ರೆಡಿ' ಚಿತ್ರದಲ್ಲಿ ಗೆಳೆತಿ 'ಅಸೀನ್' ಒತ್ತಾಯಕ್ಕಾಗಿ ಮುತ್ತಿಕ್ಕಿದ್ದರು ಎಂಬ ಸುದ್ದಿ ಈ ಹಿಂದೆ
ಹರಡಿತ್ತು..
ಚುಂಬನ ಕೊಡುವುದಕ್ಕಾಗಿ ಇಷ್ಟೆಲ್ಲಾ ಯೋಚಿಸುವ ಸಲ್ಲುಮಿಯಾ ಯಾವ ಹುಡುಗಿಗೆ ತಾನೆ ಮುತ್ತು ಕೊಡುತ್ತಾನೆ ಹೇಳಿ, ಇವರಂತು ಕೊಡಲ್ಲ ಇವರಿಗ್ಯಾರು ಮುತ್ತು ಕೊಟ್ಟರು ಎಂಬ ವಿಷಯ ಬಾಲಿವುಡ್ ಗಲ್ಲಿ
ಗಲ್ಲಿಯಲ್ಲಿ ಸುದ್ದಿಯಾಗಿ ಚರ್ಚಾ ವಿಷಯವಾಗಿತ್ತು..
'ವಾಲೆಂಟೀನ್ ಡೇ' ನಲ್ಲಿ ಸಲ್ಲುಗೆ ಕಿಸ್ ಕೊಟ್ಟಿದ್ದು ಯಾರು ಎಂಬುದಕ್ಕೆ ಕೆಲವರು 'ಕಿಕ್' ಚಿತ್ರದ ನಾಯಕಿ 'ಜಾಕ್ವೆಲಿನ್' ಅಂದರು, ಇನ್ನೂ ಕೆಲವರು ಸಲ್ಲು 'ಕತ್ರೀನಾ' ಮತ್ತೆ ಒಂದಾದರು
ಅಂದರು..
```

To aid the interpretation of the language to the English speakers an example is provided as shown below:

Original Text: ಮುಂಬೈ: 'ಬಾಲಿವುಡ್ ಚಿತ್ರರಂಗದಲ್ಲಿ ಮುತ್ತುಗಳ ಸುರಿಮಳೆಗೈದೇ ಫೇಮಸ್ ಆದ ನಟ ನಟಿಯರುಗಳ ಬಗ್ಗೆ ನಿಮಗೆ ಗೊತ್ತಿರಬಹುದು, ಆದರೆ ಬಾಲಿವುಡ್ ನ ಬ್ಯಾಚುಲರ್ ಬ್ಯಾಡ್ ಬಾಯ್ 'ಸಲ್ಮಾನ್ ಖಾನ್ ಮುತ್ತುಗಳ ವಿಷಯದಲ್ಲಿ ಮಾತ್ರ ಕೊಂಚ ದೂರ..
ಈ ಕಾರಣಕ್ಕಾಗಿ ಸಲ್ಮಾನ್ ಬೇರೆ ನಟಿಯರಿಗೆ ಕಿಸ್ ಮಾಡಲ್ಲ ಎಂಬ ಗಾಸಿಪ್ ಹರಡಿದಾಗ 'ರೆಡಿ' ಚಿತ್ರದಲ್ಲಿ ಗೆಳೆತಿ 'ಅಸೀನ್ ಒತ್ತಾಯಕ್ಕಾಗಿ ಮುತ್ತಿಕ್ಕಿದ್ದರು ಎಂಬ ಸುದ್ದಿ ಈ ಹಿಂದೆ ಹರಡಿತ್ತು..

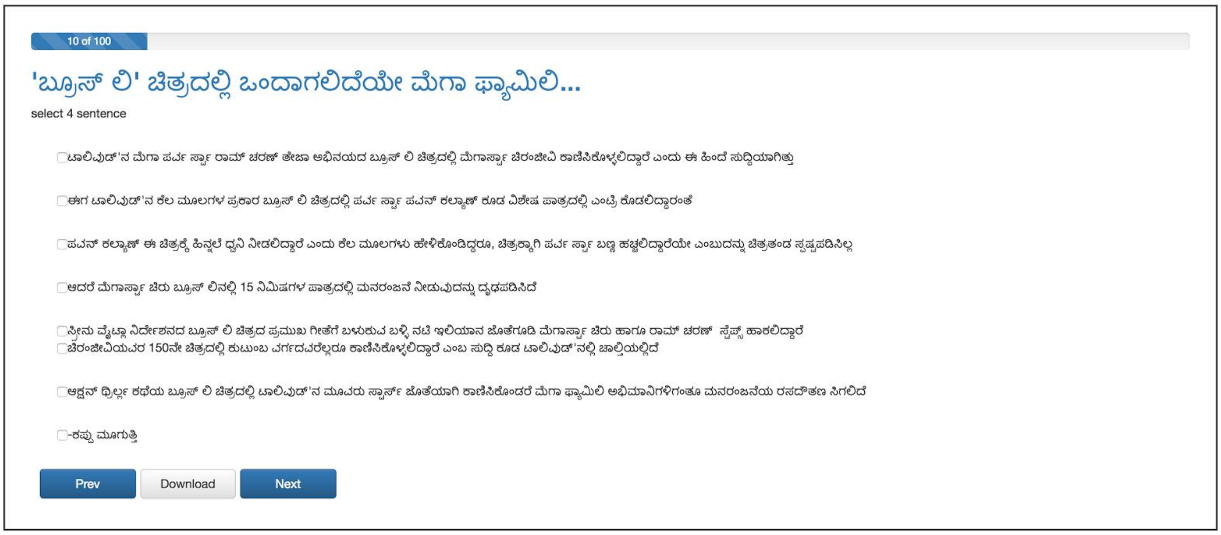
Transcription to Latin: Bollywood chitra rangadalli muttugala surimalegendhefamous aada nata natiyarugala bagge nimage gothirabahudu, aadhare bollywood na bachelor bad boy Salman Khan muttugala vishayadhalli maatra koncha doora.
ee kaaranakkagi Salman beere natiyarige kiss maadalla emba gossip haradidaaga ready chitradalli gelathi Asin othayakkagi muttikidharu emba suddhi ee hinde haradittu.

Word-by-word gloss: Bollywood Movie Field kiss famous waterfalls became actor and actresses related you aware, however bollywood bachelor boy Salman Khan kiss matter only little far.
this reason Salman other actresses kiss not doing gossip spread ready movie friend Asin insisting kiss giving like news previous spread.

Sentence meaning:You would be aware of the actors and actresses who became famous for kisses in Bollywood. However, Bollywood's bachelor boy Salman Khan stays away far from all these.
For this reason, Salman will not kiss other actress on his friend from Ready movie's insistence - this gossip was widespread.

However, the model fails in few scenarios. Since, the sentence weight is computed as the sum of the term weights, sentences of higher length would be given more weightage and hence larger sentences would be provided in the summary even if smaller sentences exist with higher relevance. **We have tried to reduce this anomaly with the help of stemming and GSS co-efficients**. Another scenario would be the existence of terms that rank a higher relevance in test data, but was not seen during the training phase. Under such a circumstance our model selects sentences that might not have unseen and higher relevancy terms.

Annotator tool:



'ಬ್ರೂಸ್ ಲಿ' ಚಿತ್ರದಲ್ಲಿ ಒಂದಾಗಲಿದೆಯೇ ಮೆಗಾ ಫ್ಯಾಮಿಲಿ...

select 4 sentence

□ಟಾಲಿವುಡ್'ನ ಮೆಗಾ ಪರ್ವ ಸ್ಟಾರ್ ರಾಮ್ ಚರಣ್ ತೇಜಾ ಅಭಿನಯದ ಬ್ರೂಸ್ ಲಿ ಚಿತ್ರದಲ್ಲಿ ಮೆಗಾಸ್ಟಾರ್ ಚಿರಂಜೀವಿ ಕಾಣಿಸಿಕೊಳ್ಳಲಿದ್ದಾರೆ ಎಂದು ಈ ಹಿಂದೆ ಸುದ್ದಿಯಾಗಿತ್ತು

□ಈಗ ಟಾಲಿವುಡ್'ನ ಕೆಲ ಮೂಲಗಳ ಪ್ರಕಾರ ಬ್ರೂಸ್ ಲಿ ಚಿತ್ರದಲ್ಲಿ ಪರ್ವ ಸ್ಟಾರ್ ಪವನ್ ಕಲ್ಯಾಣ್ ಕೂಡ ವಿಶೇಷ ಪಾತ್ರದಲ್ಲಿ ಎಂಟ್ರಿ ಕೊಡಲಿದ್ದಾರಂತೆ

□ಪವನ್ ಕಲ್ಯಾಣ್ ಈ ಚಿತ್ರಕ್ಕೆ ಹಿನ್ನಲೆ ಧ್ವನಿ ನೀಡಲಿದ್ದಾರೆ ಎಂದು ಕೆಲ ಮೂಲಗಳು ಹೇಳಿಕೊಂಡಿದ್ದರೂ, ಚಿತ್ರಕ್ಕಾಗಿ ಪರ್ವ ಸ್ಟಾರ್ ಬಣ್ಣ ಹಚ್ಚಲಿದ್ದಾರೆಯೇ ಎಂಬುದನ್ನು ಚಿತ್ರತಂಡ ಸ್ಪಷ್ಟಪಡಿಸಿಲ್ಲ

□ಆದರೆ ಮೆಗಾಸ್ಟಾರ್ ಚಿರು ಬ್ರೂಸ್ ಲಿನಲ್ಲಿ 15 ನಿಮಿಷಗಳ ಪಾತ್ರದಲ್ಲಿ ಮನರಂಜನೆ ನೀಡುವುದನ್ನು ದೃಢಪಡಿಸಿದೆ

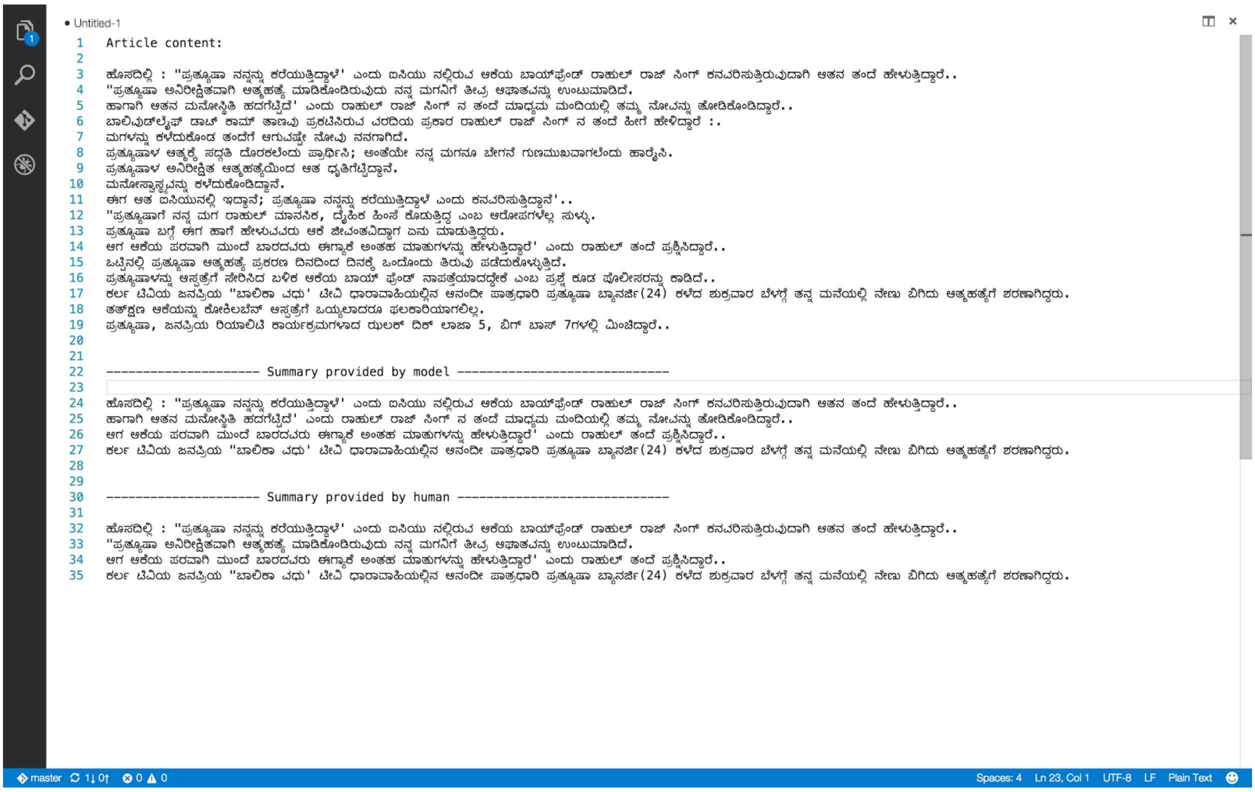
□ಸ್ರೀನು ವೈಟ್ಲ ನಿರ್ದೇಶನದ ಬ್ರೂಸ್ ಲಿ ಚಿತ್ರದ ಪ್ರಮುಖ ಗೀತೆಗೆ ಬಳುಕುವ ಬಳ್ಳಿ ನಟಿ ಇಲಿಯಾನ ಜೊತೆಗೂಡಿ ಮೆಗಾಸ್ಟಾರ್ ಚಿರು ಹಾಗೂ ರಾಮ್ ಚರಣ್ ಸ್ಟೆಪ್ಸ್ ಹಾಕಲಿದ್ದಾರೆ
□ಚಿರಂಜೀವಿಯವರ 150ನೇ ಚಿತ್ರದಲ್ಲಿ ಕುಟುಂಬ ವರ್ಗದವರೆಲ್ಲರೂ ಕಾಣಿಸಿಕೊಳ್ಳಲಿದ್ದಾರೆ ಎಂಬ ಸುದ್ದಿ ಕೂಡ ಟಾಲಿವುಡ್'ನಲ್ಲಿ ಚಾಲ್ತಿಯಲ್ಲಿದೆ

□ಆಕ್ಷನ್ ಥ್ರಿಲ್ಲರ್ ಕಥೆಯ ಬ್ರೂಸ್ ಲಿ ಚಿತ್ರದಲ್ಲಿ ಟಾಲಿವುಡ್'ನ ಮೂವರು ಸ್ಟಾರ್ಸ್ ಜೊತೆಯಾಗಿ ಕಾಣಿಸಿಕೊಂಡರೆ ಮೆಗಾ ಫ್ಯಾಮಿಲಿ ಅಭಿಮಾನಿಗಳಿಗಂತೂ ಮನರಂಜನೆಯ ರಸದೌತಣ ಸಿಗಲಿದೆ

□-ಕಪ್ಪು ಮೂಗುತ್ತಿ

A simple annotator tool was built to facilitate the summarization of articles. All the articles under a category are stored on the server. Each individual can annotate and download the annotated summary which is later used in inter-annotator agreement and ROUGE evaluation.

Comparison of summaries by model and human:



```
Article content:

ಹೊಸದಿಲ್ಲಿ : "ಪ್ರತ್ಯೂಷಾ ನನ್ನನ್ನು ಕರೆಯುತ್ತಿದ್ದಾಳೆ' ಎಂದು ಐಸಿಯು ನಲ್ಲಿರುವ ಆಕೆಯ ಬಾಯ್‌ಫ್ರೆಂಡ್ ರಾಹುಲ್ ರಾಜ್ ಸಿಂಗ್ ಕನವರಿಸುತ್ತಿರುವುದಾಗಿ ಆತನ ತಂದೆ ಹೇಳುತ್ತಿದ್ದಾರೆ..
"ಪ್ರತ್ಯೂಷಾ ಅನಿರೀಕ್ಷಿತವಾಗಿ ಆತ್ಮಹತ್ಯೆ ಮಾಡಿಕೊಂಡಿರುವುದು ನನ್ನ ಮಗನಿಗೆ ತೀವ್ರ ಆಘಾತವನ್ನು ಉಂಟುಮಾಡಿದೆ.
ಹಾಗಾಗಿ ಆತನ ಮನೋಸ್ಥಿತಿ ಹದಗೆಟ್ಟಿದೆ' ಎಂದು ರಾಹುಲ್ ರಾಜ್ ಸಿಂಗ್ ನ ತಂದೆ ಮಾಧ್ಯಮ ಮಂದಿಯಲ್ಲಿ ತಮ್ಮ ನೋವನ್ನು ತೋಡಿಕೊಂಡಿದ್ದಾರೆ..
ಬಾಲಿವುಡ್‌ಲೈಫ್ ಡಾಟ್ ಕಾಮ್ ತಾಣವು ಪ್ರಕಟಿಸಿರುವ ವರದಿಯ ಪ್ರಕಾರ ರಾಹುಲ್ ರಾಜ್ ಸಿಂಗ್ ನ ತಂದೆ ಹೀಗೆ ಹೇಳಿದ್ದಾರೆ :.
ಮಗಳನ್ನು ಕಳೆದುಕೊಂಡ ತಂದೆಗೆ ಆಗುವಷ್ಟೇ ನೋವು ನನಗಾಗಿದೆ.
ಪ್ರತ್ಯೂಷಾಳ ಆತ್ಮಕ್ಕೆ ಸದ್ಗತಿ ದೊರಕಲೆಂದು ಪ್ರಾರ್ಥಿಸಿ; ಅಂತೆಯೇ ನನ್ನ ಮಗನೂ ಬೇಗನೆ ಗುಣಮುಖವಾಗಲೆಂದು ಹಾರೈಸಿ.
ಪ್ರತ್ಯೂಷಾಳ ಅನಿರೀಕ್ಷಿತ ಆತ್ಮಹತ್ಯೆಯಿಂದ ಆತ ಧೃತಿಗೆಟ್ಟಿದ್ದಾನೆ.
ಮನೋಸ್ವಾಸ್ಥ್ಯವನ್ನು ಕಳೆದುಕೊಂಡಿದ್ದಾನೆ.
ಈಗ ಆತ ಐಸಿಯುನಲ್ಲಿ ಇದ್ದಾನೆ; ಪ್ರತ್ಯೂಷಾ ನನ್ನನ್ನು ಕರೆಯುತ್ತಿದ್ದಾಳೆ ಎಂದು ಕನವರಿಸುತ್ತಿದ್ದಾನೆ'..
"ಪ್ರತ್ಯೂಷಾಗೆ ನನ್ನ ಮಗ ರಾಹುಲ್ ಮಾನಸಿಕ, ದೈಹಿಕ ಹಿಂಸೆ ಕೊಡುತ್ತಿದ್ದ ಎಂಬ ಆರೋಪಗಳೆಲ್ಲ ಸುಳ್ಳು.
ಪ್ರತ್ಯೂಷಾ ಬಗ್ಗೆ ಈಗ ಹಾಗೆ ಹೇಳುವವರು ಆಕೆ ಜೀವಂತವಿದ್ದಾಗ ಏನು ಮಾಡುತ್ತಿದ್ದರು.
ಆಗ ಆಕೆಯ ಪರವಾಗಿ ಮುಂದೆ ಬಾರದವರು ಈಗ್ಯಾಕೆ ಅಂತಹ ಮಾತುಗಳನ್ನು ಹೇಳುತ್ತಿದ್ದಾರೆ' ಎಂದು ರಾಹುಲ್ ತಂದೆ ಪ್ರಶ್ನಿಸಿದ್ದಾರೆ..
ಒಟ್ಟಿನಲ್ಲಿ ಪ್ರತ್ಯೂಷಾ ಆತ್ಮಹತ್ಯೆ ಪ್ರಕರಣ ದಿನದಿಂದ ದಿನಕ್ಕೆ ಒಂದೊಂದು ತಿರುವು ಪಡೆದುಕೊಳ್ಳುತ್ತಿದೆ.
ಪ್ರತ್ಯೂಷಾಳನ್ನು ಆಸ್ಪತ್ರೆಗೆ ಸೇರಿಸಿದ ಬಳಿಕ ಆಕೆಯ ಬಾಯ್ ಫ್ರೆಂಡ್ ನಾಪತ್ತೆಯಾದದ್ದೇಕೆ ಎಂಬ ಪ್ರಶ್ನೆ ಕೂಡ ಪೊಲೀಸರನ್ನು ಕಾಡಿದೆ..
ಕರ್ಲ ಟಿವಿಯ ಜನಪ್ರಿಯ "ಬಾಲಿಕಾ ವಧು' ಟೀವಿ ಧಾರಾವಾಹಿಯಲ್ಲಿನ ಆನಂದೀ ಪಾತ್ರಧಾರಿ ಪ್ರತ್ಯೂಷಾ ಬ್ಯಾನರ್ಜಿ(24) ಕಳೆದ ಶುಕ್ರವಾರ ಬೆಳಗ್ಗೆ ತನ್ನ ಮನೆಯಲ್ಲಿ ನೇಣು ಬಿಗಿದು ಆತ್ಮಹತ್ಯೆಗೆ ಶರಣಾಗಿದ್ದರು.
ತತ್‌ಕ್ಷಣ ಆಕೆಯನ್ನು ಕೋಕಿಲಬೆನ್ ಆಸ್ಪತ್ರೆಗೆ ಒಯ್ಯಲಾದರೂ ಫಲಕಾರಿಯಾಗಲಿಲ್ಲ.
ಪ್ರತ್ಯೂಷಾ, ಜನಪ್ರಿಯ ರಿಯಾಲಿಟಿ ಕಾರ್ಯಕ್ರಮಗಳಾದ ಝಲಕ್ ದಿಕ್ ಲಾಜಾ 5, ಬಿಗ್ ಬಾಸ್ 7ಗಳಲ್ಲಿ ಮಿಂಚಿದ್ದಾರೆ..


--------------------- Summary provided by model -----------------------------

ಹೊಸದಿಲ್ಲಿ : "ಪ್ರತ್ಯೂಷಾ ನನ್ನನ್ನು ಕರೆಯುತ್ತಿದ್ದಾಳೆ' ಎಂದು ಐಸಿಯು ನಲ್ಲಿರುವ ಆಕೆಯ ಬಾಯ್‌ಫ್ರೆಂಡ್ ರಾಹುಲ್ ರಾಜ್ ಸಿಂಗ್ ಕನವರಿಸುತ್ತಿರುವುದಾಗಿ ಆತನ ತಂದೆ ಹೇಳುತ್ತಿದ್ದಾರೆ..
ಹಾಗಾಗಿ ಆತನ ಮನೋಸ್ಥಿತಿ ಹದಗೆಟ್ಟಿದೆ' ಎಂದು ರಾಹುಲ್ ರಾಜ್ ಸಿಂಗ್ ನ ತಂದೆ ಮಾಧ್ಯಮ ಮಂದಿಯಲ್ಲಿ ತಮ್ಮ ನೋವನ್ನು ತೋಡಿಕೊಂಡಿದ್ದಾರೆ..
ಆಗ ಆಕೆಯ ಪರವಾಗಿ ಮುಂದೆ ಬಾರದವರು ಈಗ್ಯಾಕೆ ಅಂತಹ ಮಾತುಗಳನ್ನು ಹೇಳುತ್ತಿದ್ದಾರೆ' ಎಂದು ರಾಹುಲ್ ತಂದೆ ಪ್ರಶ್ನಿಸಿದ್ದಾರೆ..
ಕರ್ಲ ಟಿವಿಯ ಜನಪ್ರಿಯ "ಬಾಲಿಕಾ ವಧು' ಟೀವಿ ಧಾರಾವಾಹಿಯಲ್ಲಿನ ಆನಂದೀ ಪಾತ್ರಧಾರಿ ಪ್ರತ್ಯೂಷಾ ಬ್ಯಾನರ್ಜಿ(24) ಕಳೆದ ಶುಕ್ರವಾರ ಬೆಳಗ್ಗೆ ತನ್ನ ಮನೆಯಲ್ಲಿ ನೇಣು ಬಿಗಿದು ಆತ್ಮಹತ್ಯೆಗೆ ಶರಣಾಗಿದ್ದರು.


--------------------- Summary provided by human -----------------------------

ಹೊಸದಿಲ್ಲಿ : "ಪ್ರತ್ಯೂಷಾ ನನ್ನನ್ನು ಕರೆಯುತ್ತಿದ್ದಾಳೆ' ಎಂದು ಐಸಿಯು ನಲ್ಲಿರುವ ಆಕೆಯ ಬಾಯ್‌ಫ್ರೆಂಡ್ ರಾಹುಲ್ ರಾಜ್ ಸಿಂಗ್ ಕನವರಿಸುತ್ತಿರುವುದಾಗಿ ಆತನ ತಂದೆ ಹೇಳುತ್ತಿದ್ದಾರೆ..
"ಪ್ರತ್ಯೂಷಾ ಅನಿರೀಕ್ಷಿತವಾಗಿ ಆತ್ಮಹತ್ಯೆ ಮಾಡಿಕೊಂಡಿರುವುದು ನನ್ನ ಮಗನಿಗೆ ತೀವ್ರ ಆಘಾತವನ್ನು ಉಂಟುಮಾಡಿದೆ.
ಆಗ ಆಕೆಯ ಪರವಾಗಿ ಮುಂದೆ ಬಾರದವರು ಈಗ್ಯಾಕೆ ಅಂತಹ ಮಾತುಗಳನ್ನು ಹೇಳುತ್ತಿದ್ದಾರೆ' ಎಂದು ರಾಹುಲ್ ತಂದೆ ಪ್ರಶ್ನಿಸಿದ್ದಾರೆ..
ಕರ್ಲ ಟಿವಿಯ ಜನಪ್ರಿಯ "ಬಾಲಿಕಾ ವಧು' ಟೀವಿ ಧಾರಾವಾಹಿಯಲ್ಲಿನ ಆನಂದೀ ಪಾತ್ರಧಾರಿ ಪ್ರತ್ಯೂಷಾ ಬ್ಯಾನರ್ಜಿ(24) ಕಳೆದ ಶುಕ್ರವಾರ ಬೆಳಗ್ಗೆ ತನ್ನ ಮನೆಯಲ್ಲಿ ನೇಣು ಬಿಗಿದು ಆತ್ಮಹತ್ಯೆಗೆ ಶರಣಾಗಿದ್ದರು.
```

The above screenshot shows the actual article and the output of the model or the system built and also shows a summary of the same by Human on the basis of Inter-Annotator agreement.

Discussion:

We believe that our Kannada article summarizer will effectively summarize the Kannada articles with a fair accuracy level. Our belief is reinforced by a good level of concordance reached between the machine provided summaries and the summaries provided by the humans on the same set of articles. The significant contributions to the NLP community include a reliable working model of a Kannada summarizer, Kannada corpus, an efficient statistical-based Kannada stemmer and a user friendly annotation tool.

During the course of our research, we have found that there are more than 10000 articles in Kannada across the web and a significant number of people access these articles. Therefore, our summarizer could be a reliable tool for these users in determining the summaries of such articles. Our future research would include the incorporation of text ranking and in creation of a Wordnet for Kannada language that may boost the accuracy of summarization. We were not able to incorporate and evaluate the use of text ranking algorithm due to time constraints.

However, we are considering to incorporporate this algorithm in our future research. We would also explore other inter annotator strategies such as Kappa coefficient and in the comparison of various inter-annotator strategies to reach a reliable consensus between the summaries.

References:

[1]http://www.ijcaonline.org/volume27/number10/pxc3874583.pdfhttps://web.eecs.umich.edu/~mihalcea/papers/mihalcea.emnlp04.pdfhttp://nlp.cic.ipn.mx/Publications/2008/Text%20Summarization%20by%20Sentence%20Extraction%20Using.pdfhttp://research.ijcaonline.org/volume75/number6/pxc3890449.pdf

[2]http://airccj.org/CSCP/vol1/csit1311.pdfhttp://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6416635&tag=1
[3] http://www.mirlabs.org/jnic/secured/Volume1-Issue1/Paper14/JNIC_Paper14.pdf
[4] http://airccse.org/journal/ijsc/papers/2411ijsc08.pdf
[5] https://www.youtube.com/watch?v=Vw-7XkP9H1o
[6] http://www.berouge.com/Pages/default.aspx
[7] https://www.youtube.com/watch?v=IQo5dfMt8Cc
[8] https://staff.fnwi.uva.nl/r.fernandezrovira/teaching/MoLProject2011/annotation-reliability.pdf
[9] http://www.mirlabs.org/jnic/secured/Volume1-Issue1/Paper14/JNIC_Paper14.pdf
[10] http://kavita-ganesan.com/rouge-howto

Division of Labor:

Building the corpus (All 4)
Evaluation of Inter-annotator agreement (Achyut)
Building TF-IDF (Nitish)
Building GSS coefficients (Achyut)
Stemming(Virupaksha swamy)
Rouge Model Scripts(Swaroop)
Annotator Tool (Swaroop)
Scrapping(Swaroop)
Annotation tool(Swaroop)
Integration of All Scripts (Swaroop)
Evaluation of algorithmic approaches to build the model (All 4)
Comparing and reporting the performance of the model against an established baseline (All 4)
Documentation (All 4)

Word Count: 2040