भारतीय वैज्ञानिक एवं औद्योगिक अनुसंधान पत्रिका
वर्ष 31 अंक (2) दिसम्बर 2023 पृ. 135-141
DOI: 10.56042/bvaap.v31i2.6388

हिंदी पाठ के लिए प्राकृतिक भाषा संसाधन के परिप्रेक्ष्य में पूर्व-संसाधन

अंजना किशनपुरी, धीरेंद्र यादव एवं हर्षलता पेटकर
सूचना एवं भाषा अभियांत्रिकी केंद्र, महात्मा गांधी अंतर्राष्ट्रीय हिंदी विश्वविद्यालय, वर्धा (महाराष्ट्र)
[ई-मेलः anjanaresearcher@gmail.com]

सारांश : डिजिटल प्रारूप में पाठ्य डेटा ऑनलाइन और ऑफलाइन दोनों मोड़ में आजकल तेजी से बढ़ रहा है। इन दस्तावेजों को प्रबंधित करना और पुनर्प्राप्त करना मुश्किल हो जाता है। प्राकृतिक भाषा संसाधन (NLP) जैसे अभिलेखीय, पुनर्प्राप्ति, प्रश्न प्रतिक्रिया, पाठ सारांश, मशीनी अनुवाद आदि पाठ्य दस्तावेज़ों के कुशल पूर्व-संसाधन पर अत्यधिक निर्भर है। प्राकृतिक भाषा संसाधन के इस विशेष क्षेत्र में शोधकर्ताओं को मशीन लर्निंग एल्गोरिद्म को भाषाओं के आधार पर स्वचालित रूप से पूर्व-संसाधनके लिए दस्तावेज़ों पर लागू करने के लिए प्रेरित किया है, जो इसके संदर्भ के आधार पर दस्तावेज़ों को संसाधित करने के तरीकों को विकसित करने के लिए प्रेरित करते हैं। प्रस्तुत शोधपत्र के अंतर्गत प्राकृतिक भाषा संसाधन के परिप्रेक्ष्य में पूर्व-संसाधन अनुप्रयोग प्रस्तावित किया गया है। पूर्व संसाधन पाठ्य खनन, प्राकृतिक भाषा संसाधन (NLP) और सूचना पुनर्प्राप्ति (IR) में एक महत्वपूर्ण कार्य है। हालांकि, पूर्व संसाधन के बिना कोई भी कच्चे पाठ्य डेटा पर कार्य नहीं किया जा सकता है। पाठ्य पूर्व संसाधन ठीक से निष्पादित होने पर सर्वोत्कृष्ट परिणाम सुनिश्चित करता है। प्राकृतिक भाषा संसाधन के क्षेत्र में, डेटा पूर्व संसाधन का उपयोग असंरचित पाठ्य डेटा से रोचक एवं ज्ञान निकालने के लिए किया जाता है। इस शोधपत्र के द्वारा हिंदी विधि डोमेन के लिए पूर्व संसाधन अनुप्रयोग प्रस्तावित किया गया है जिसके अंतर्गत सामान्यीकरण, टोकनाइजेशन, विराम शब्द निष्कासन एवं स्टेमिंग जैसी महत्वपूर्ण भाषिक प्रक्रियाओं की व्यापक और उपयोगी समझ प्रदान कराना है।

Pre-processing in the context of natural language resources for Hindi lessons

Anjana Kishanpuri, Dhirendra Yadav & Harshalata Petkar
Center for Information and Language Engineering, Mahatma Gandhi International Hindi University, Wardha (Maharashtra)

Abstract

Nowadays, text data in digital format of online and offline mode is increasing rapidly, it becomes difficult to manage and retrieve the text documents. Natural language processing (NLP) is highly dependent on efficient pre-processing of text documents such as archival, retrieval, query response, text summarization, machine translation, etc.This specialized area of natural language processing has led inspired researchers to do apply machine learning algorithms to automatically pre-process documents based on languages, developing methods to process documents based on their context.Under the present research paper, a pre-processing application has been proposed in the context of natural language processing. Pre-processing is an important function in text mining, natural language processing (NLP), and information retrieval (IR). However, no raw text data can be worked on without pre-processing. Text pre-processing ensures optimum results when executed properly. In the field of natural language processing, text pre-processing is used to extract interesting and knowledgeable information from unstructured textual data.This paper proposes a pre-processing application for the Hindi legal domain to provide a comprehensive and useful understanding of important linguistic processes such as normalization, tokenization, stop word removal and stemming.

प्रस्तावना

पूर्व संसाधन पाठ्य खनन, प्राकृतिक भाषा संसाधन (NLP) और सूचना पुनर्प्राप्ति (IR) में एक महत्वपूर्ण कार्य और महत्वपूर्ण कदम है। पाठ्य खनन के क्षेत्र में, डेटा पूर्व संसाधन का उपयोग असंरचित पाठ्य डेटा से रोचक एवं ज्ञान निकालने के लिए किया जाता है। सूचना पुनर्प्राप्ति अनिवार्य रूप से यह निर्णय लेने का विषय है कि एक संग्रह में कौन से प्रलेखों को पुनः प्राप्त किया जाना चाहिए ताकि प्रयोक्ता की सूचना की आवश्यकता को पूरा किया जा सके। प्रयोक्ता की सूचना की आवश्यकता को प्रश्न या प्रोफाइल द्वारा व्यक्त किया जाता है और इसमें एक या अधिक

खोज शब्द (पद), साथ ही कुछ अतिरिक्त जानकारी जैसे शब्दों के भार को समाहित किया जाता है। अतः पुनर्प्राप्ति निर्णय प्रश्नों की शर्तों की तुलना सूचक शब्दों (महत्वपूर्ण शब्दों या वाक्यांशों) के साथ स्वयं प्रलेख में प्रस्तुत होने से की जाती है। यह निर्णय द्विआधारी (पुनः प्राप्त/अस्वीकार) हो सकता है या जिन प्रलेख में प्रश्न होते हैं उसमें प्रासंगिकता की उस डिग्री का प्राक्कलन समाहित हो सकता है। दुर्भाग्यवश, प्रलेखों तथा प्रश्नोत्तरों में दिखाई देने वाले शब्दों के संरचनात्मक प्रकार प्रायः अनेक प्रकार के होते हैं। अतः प्रलेखों से सूचना पुनर्प्राप्ति से पहले, डेटा पूर्व संसाधन तकनीकों को डेटा सेट के आकार को कम करने के लिए लक्ष्य डेटा सेट पर लागू किया जाता है जिससे डेटा सेट के आकार को कम किया जा सके और इस अध्ययन का उद्देश्य इनके पूर्व संसाधन विधियों जैसे- सामान्यीकरण, टोकेनाइजेशन, विराम शब्द निष्कासन एवं स्टेमिंग का विशलेषण करना है।

पूर्व संसाधन

पाठ पूर्व संसाधन किसी भी NLP प्रणाली का एक अनिवार्य भाग है, क्योंकि चरण पर पहचाने गए अक्षर, शब्द और वाक्य आगे की प्रक्रिया चरणों में पारित की जाने वाली मूल इकाईयां हैं, विश्लेषण और टैगिंग के घटकों से, जैसे कि रूपिम विश्लेषक (Morphological Analyzer) और शब्द भेद टैगर (Part-of-Speech Tagger) को सूचना पुनर्प्राप्ति और मशीन अनुवाद प्रणाली जैसे अनुप्रयोगों के माध्यम से किया जाता है। यह गतिविधियों का एक संग्रह है जिसमें पाठ प्रलेख पूर्व संसाधित होते हैं क्योंकि पाठ डेटा में प्रायः कुछ विशेष प्रारूप होते हैं जैसे संख्या प्रारूप, तिथि प्रारूप और सबसे सामान्य शब्द जो पाठ्य खनन में मदद करने की संभावना नहीं रखते हैं जैसे कि पूर्वसर्ग, लेख, और सर्वनामों को समाप्त किया जा सकता है।

NLP प्रणाली में पाठ पूर्व संसाधन की आवश्यकता हैः-

(Need of Text Pre-processing in NLP System)

1. मूल पाठ प्रलेख की सूची फाइल के आकार को कम करने के लिए,
   i. किसी विशेष पाठ प्रलेख में शब्दों के कुल शब्द गणना का 20-30% स्टॉप वर्डका होना।
   ii. Stemming द्वारा पाठ के आकार को 40-50% तक कम कर सकता है।
2. सूचना पुनर्प्राप्तिप्रणाली की दक्षता एवं प्रभावशीलता को सुधारने के लिए,
   i. Stop words ख़ोज या पाठ खनन के लिए उपयोगी नहीं होते है तथा वे पुनर्प्राप्ति प्रणाली को भ्रमित कर सकते हैं।
   ii. Stemming किसी पाठ प्रलेख में समान शब्दों के मिलान में प्रयुक्त होती है।

साहित्य पुनर्वलोकन

प्रस्तुत शोध पत्र में प्राकृतिक भाषा संसाधन, सूचना पुनर्प्राप्ति एवं सूचना निष्कर्षण में पूर्व संसाधन की प्रक्रिया को केंद्रित किया गया है। टर्पीस और Gurpreet (2011) ने बताया कि सर्वप्रथम प्राकृतिक भाषा संसाधन के अंतर्गत पंजाबी पाठ के लिए पूर्व संसाधन पर कार्य किया था। इन्होंने इस शोध पत्र में पंजाबी पाठ सारांशीकरण प्रणाली के लिए पूर्व संसाधन के विभिन्न चरणों का गहन विश्लेषण कर कार्यान्वयन किया था। पूर्व संसाधन में उपयोग किए जाने वाले अधिकांश शाब्दिक संसाधन जैसे पंजाबी स्टेमर, पंजाबी उचित नाम सूची, अंग्रेजी-पंजाबी संज्ञा सूची आदि को स्क्रैच प्रोग्रामिंग इंटरफ़ेस द्वारा विकसित किया। तथा इन संसाधनों को विकसित करने के लिए मानवीकृत एवं स्वचालित उपकरणों का प्रयोग करके पंजाबी शब्द संग्रह और पंजाबी शब्द रूपों का गहन विश्लेषण किया।

पॉल एवं अन्य (2013) द्वारा हिंदी के लिए एक लेमेटाइज़र के विकास पर चर्चा की गयी है। जोकि नियम आधारित दृष्टिकोण का उपयोग करता है जिसमें सभी हिंदी शब्द शामिल हैं जो आमतौर पर दैनिक जीवन में उपयोग किए जाते हैं। इस दृष्टिकोण ने स्थान के बजाय समय अनुकूलन समस्या पर भी जोर दिया गया है। चूंकि आजकल अंतरिक्ष कोई बड़ी समस्या नहीं है, इसलिए उनके दृष्टिकोण का उद्देश्य समय को अनुकूलित करना और बहुत ही कम समय में सटीक परिणाम उत्पन्न कराना था। प्रणाली द्वारा 91% सटीकता प्रदान की गई है।

फेनसन (2014) ने असंरचित ऑनलाइन वित्तीय ग्रंथों से अवांछित पाठ हटाने और ऐसे पाठों को एक उन्नत प्रारूप में व्यवस्थित करने के लिए एक प्राकृतिक भाषा संसाधन आध ाारित पूर्व-संसाधन दृष्टिकोण प्रस्तुत किया था। जो पाठ्य विशेषता निष्कर्षण के लिए अधिक उपयोगी है। ऑनलाइन अनौपचारिक पाठ में अवांछित पाठ को कम करने के लिए प्रस्तावित दृष्टिकोण छह प्राकृतिक भाषा संसाधन प्रसंस्करण चरणों को एकीकृत करता है, जिसमें एक विकसित वाक्य विन्यास संबंधी (Syntactic) एवं अर्थ-संबंधी (Semantics) संयुक्त निषेध संचालित कलनविधि (Negation handling Algorithm) शामिल है। साथ ही प्रत्येक प्रणाली कार्यान्वयन में त्रिवर्गीय भाववर्गीकरण भी प्रस्तुत किया। प्रायोगिक परिणाम

बताते हैं कि प्रस्तावित पूर्व-संसाधन दृष्टिकोण अन्य पूर्व-संसाधन विधियों से बेहतर प्रदर्शन करता है।

आशीष (2014) ने इस शोधपत्र में पाठ के आकार को कम करने के लिए पाठ सारांशीकरण के पूर्व-संसाधन के चरणों पर चर्चा की है। पूर्व-संसाधन चरण में हाइफ़न, विराम शब्द, समानार्थी शब्द और समानार्थी वाक्यों को निष्कासित कर दिया जाता है। साथ ही इस शोध पत्र में पाठ सारांशीकरण के संदिग्ध विषय परिभाषित हैं। इस शोध पत्र में परिभाषित पूर्व संसाधन चरण के सभी चरणों का उपयोग करके सभी संदिग्ध विषयों को हल किया जा सकता है।

जयदीप एवं जितेन्द्र (2016) संस्कृत भाषा के लिए विराम शब्द निष्कासन एल्गोरिद्म और इसके कार्यान्वयन को डिजाइन करने के लिए एक सरल दृष्टिकोण प्रस्तुत किया था। एल्गोरिद्म और कार्यान्वयन शब्दकोश आधारित दृष्टिकोण का उपयोग करता था। शब्दकोश आधारित दृष्टिकोण में विराम शब्दों की पूर्वनिर्धारित सूची की तुलना लक्ष्य पाठ से तब की जाती थी जब विराम शब्द निष्कासन की आवश्यकता होती थी।

अंजुशा (2016) ने प्रस्तावित शोध पत्र में हिंदी पाठ सारांश प्रणाली के लिए पूर्व-संसाधन चरण पर चर्चा की है। पूर्व संसाधन में उपयोग किए जाने वाले अधिकांश संसाधन जैसे हिंदी स्टेमर, हिंदी विराम शब्द सूची आदि को डोमेन विशेष अच्छे परिणाम उत्पन्न करने के लिए विकसित किया गया। पूर्व संसाधन का इनपुट कोई भी असंरचित हिंदी पाठ और आउटपुट संरचित पाठ होता है।

मुजम्मल एवं महमुदुल (2018) द्वारा यह शोध अध्ययन मुख्य रूप से बंगाली क्रिया के विश्लेषण पर केंद्रित किया गया था। इस शोधपत्र ने अनुकूलित बैग-ऑफ-वर्ड्स बनाने के लिए बंगाली क्रिया के बहुरूप से एक अद्वितीय और सार्थक शब्द खोजने के लिए एक एल्गोरिद्म और संबंधित नियमों का प्रस्ताव दिया गया है जिसके परिणामस्वरूप पाठ वर्गीकरण अधिक सटीक होगा।

रवि स्टेनिस्लस (2018) ने अपने सर्वेक्षण पत्र में सामान्य रूप से पाठ खनन और विशेष रूप से प्राकृतिक भाषा संसाधन के लिए पूर्व-संसाधन टूल प्रस्तुत किया था। साथ ही सर्वाधिक उपयोग की जाने वाली पाठ खनन तकनीकों की व्यापक श्रेणियां भी प्रस्तुत कीं। उनके शोध पत्र का उद्देश्य पाठ संसाधन तकनीकों और टूल्स की कई विशेषताओं का पता लगाना और उनका विश्लेषण करना था।

आनन्द (2019) यह शोधपत्र हिंदी में किसी शब्द के मूल रूप को खोजने के लिए एक एल्गोरिद्म प्रस्तुत करता है। प्रस्तावित एल्गोरिद्म word2vec का उपयोग करता है, जो एक अर्ध-पर्यवेक्षित शिक्षण एल्गोरिद्म है, जो एक कॉर्पस से 10 सबसे समान शब्दों को खोजने के लिए है फिर मूलरूप को खोजने के उपर्युक्त कार्य को प्राप्त करने के लिए एक गणितीय कार्य प्रस्तावित किया जाता है। प्रस्तावित एल्गोरिद्म को किसी एनोटेट कॉर्पस की आवश्यकता नहीं है और स्टेम को खोजने के लिए किसी जटिल कोडेड नियमों का उपयोग नहीं करता है। परिणामों को एक संग्रह से बेहतर ढंग से लिए गए 1000 हिंदी शब्दों के एक सेट का चयन करके और प्रस्तावित एल्गोरिद्म द्वारा दिए गए परिणामों और मैन्युअल रूप से बनाए गए वास्तविक परिणामों की तुलना करके सत्यापित किया जाता है।

अपूर्वा (2020) ने मराठी ई-न्यूज़ आलेखों (शैक्षिक, राजनीतिक और खेल समाचारों जैसे डोमेन विशेष) पर पूर्व संसाधन तकनीकों का ध्यान केंद्रित कर अन्वेषण किया है। वे एक ऐसी पूर्व संसाधन प्रणाली विकसित करने की कोशिश कर रही हैं जो मराठी ई-न्यूज को सारांशित करने के लिए तुलनात्मक रूप से अधिक सक्षम और कुशल होगी।

आइशा एवं राजेन्द्र (2020) ने अपने सर्वेक्षण पत्र में प्राकृतिक भाषा संसाधनके पाठ्य विशेषता निष्कर्षण तकनीकों का उपयोग करके सूचना पुनर्प्राप्ति के साथ-साथ पाठ पूर्व संसाधन तकनीकों के उपयोग के महत्व पर जोर दिया है। जिसके अंतर्गत पाठ्य विशेषता निष्कर्षण, बैग ऑफ़ वर्ड्स, टी एफ. आई डी एफ जैसी तकनीकों को सविस्तार से समझाया है।

पूर्व संसाधन प्रक्रिया

I. **सामान्यीकरण (Normalization)-** इस प्रक्रिया में पाठ डेटा में प्रायः कुछ विशेष प्रारूप होते हैं जैसे संख्या प्रारूप, तिथि प्रारूप और सबसे सामान्य शब्द जो पाठ्य खनन में मदद करने की संभावना नहीं रखते हैं जैसे कि रिक्त स्थान, अन्य विशेष प्रतीकों, पूर्वसर्ग और संक्षेपाक्षर को समाप्त किया जाता है साथ ही वाक्यों को विघटित किया जाता है और फिर प्रत्येक वाक्य में शब्दों की संख्या गिनी जाती है। हिंदी में, वाक्य की सीमा 'पूर्ण विराम' की पहचान करके जिससे वाक्य समाप्त होता है को खंडित किया जाता है।

II. **टोकेनाइजेशन (Tokenization)-** एक ऐसी प्रक्रिया है जिसमें वाक्य से शब्दों, वाक्यांशों, प्रतीकों को तोड़ने का कार्य किया जाता है और अन्य पृथक अर्थपूर्ण पद को token कहा जाता है। टोकेनाइजेशन का उद्देश्य एक वाक्य में शब्दों की खोज करना है। इन Tokens की सूची आगे संसाधन जैसे Parsing या पाठ खनन के लिए इनपुट बन जाता है। भाषा

विज्ञान (जहाँ यह पाठ विभाजन का एक रूप है) और कंप्यूटर विज्ञान में, जहाँ यह शाब्दिक विश्लेषण का एक भाग है, दोनों में ही टोकेनाइजेशन उपयोगी होता है। शाब्दिक डेटा, प्रारंभ में वर्गों का केवल एक खंड होता है। सूचना पुनर्प्राप्ति की सभी प्रक्रियाओं में डेटा सेट के शब्दों की आवश्यकता होती है। इसलिए, parser के लिए प्रलेखों का टोकेनाइजेशन होना आवश्यक होता है। टोकेनाइजेशन का मुख्य उपयोग सार्थक कीवर्ड की पहचान करना होता है, भिन्न संख्या और समय प्रारूपों में विसंगति हो सकती है। दूसरी समस्या संक्षिप्ताक्षर (abbreviations) और acronyms शब्द हैं, जिन्हें एक मानक रूप में रूपांतरित करना होता है।

टोकेनाइजेशन की समस्याएं - टोकेनाइजेशन में चुनौतियां भाषा के प्रकार पर निर्भर करती हैं। अंग्रेज़ी एवं फ्रांसीसी भाषाओं को अंतराल-सीमांकित माना जाता है क्योंकि अधिकांश शब्द सफ़ेद स्थानों से अलग-अलग होते हैं। चीनी और थाई जैसी अखंडित भाषाओं को जिसमें शब्दों को स्पष्ट सीमाओं में नहीं रखा गया है। अखंडित भाषा वाक्यों को tokenize करने के लिए अतिरिक्त शाब्दिक एवं रुपवाचक जानकारी की आवश्यकता होती है। शब्दों के लेखन-प्रणाली तथा शब्द-मुद्रण की संरचना से टोकेनाइजेशन प्रभावित होता है। भाषाओं की संरचना को तीन श्रेणियों में बांटा जा सकता है :

Isolating (पृथक): शब्द छोटे इकाइयों में विभाजित नहीं करते हैं। उदाहरणः Mandarin Chinese.

Agglutinative: शब्द छोटे इकाईयों में विभाजित होते हैं। उदाहरणः जापानी, तमिल।

Inflectional: रूपिम के बीच की सीमाएं व्याकरणिक अर्थ के सन्दर्भ में अस्पष्ट और संदिग्धार्थक हैं। उदाहरणः लैटिन।

III. विराम शब्द निष्कासन (Stop Word Removal)- प्रलेख में प्रायः अनेक शब्द पुनः प्रयोग होते हैं परंतु अनिवार्य रूप से निरर्थक होते हैं क्योंकि इसका प्रयोग वाक्यों में शब्दों को एक साथ जोड़ने में किया जाता है। यह सामान्यतः समझा जाता है कि पाठ प्रलेखों के संदर्भ या विषय वस्तु में विराम शब्द योगदान नहीं देते। विराम शब्दों की उच्च आवृत्ति आने के कारणए पाठ्य खनन में उनकी उपस्थिति से प्रलेखों के विषय-वस्तु को समझने में बाधा आती है। विराम शब्द प्रायः सामान्य शब्दों जैसे 'और', हैं', 'यह' आदि का उपयोग किया जाता है। वे प्रलेखों के वर्गीकरण में उपयोगी नहीं होते हैं इसलिए उन्हें हटाना ही चाहिए। फिर भी, ऐसी विराम शब्दों की सूची का विकास पाठ स्रोतों के बीच कठिन और असंगत है। यह प्रक्रिया पाठ डेटा को भी कम करती है और प्रणाली के निष्पादन में सुधार करती है। प्रत्येक पाठ प्रलेख इन शब्दों से संबंधित होते हैं जो पाठ खनन अनुप्रयोगों के लिए आवश्यक नहीं हैं। उदाहरणः- के, का, एक, में, है, यह, और, इस, जो, कर, अपने, ने, गया, किया, जबकि, हालांकि इत्यादि ।

IV. स्टेमिंग (Stemming)-स्टेमिंग, एक शब्द के भिन्न रूपों के मिलान के एक सामान्य प्रतिनिधित्व की प्रक्रिया है। उदाहरण के लिए, शब्दः 'राष्ट्रीय', 'राष्ट्रीयता', 'राष्ट्रीयकरण', 'राष्ट्रवाद' सभी को 'राष्ट्र' रूप में कम करके प्रदर्शित किया जा सकता है। यह पाठ संसाधन में सूचना पुनर्प्राप्ति के लिए एक व्यापक रूप से प्रयुक्त की जाने वाली प्रक्रिया है जो कि इस धारणा पर आधारित है। स्टेमिंग में त्रुटियां (Error in Stemming)- स्टेमिंग में मुख्य रूप से दो त्रुटियां होती हैंः.

1. ओवर स्टेमिंग- जब दो शब्द अलग-अलग stems के साथ एक ही root पर stemmed होते हैं, वह प्रक्रिया Over Stemming कहलाती है। इसे false positive के नाम से भी जाना जाता है।

2. अंडर स्टेमिंग-जब दो शब्द एक ही root पर stemmed नहीं होते हैं, वह प्रक्रिया Under Stemming कहलाती है। इसे false negative के नाम से भी जाना जाता है।

प्रस्तावित कार्यविधि

इस शोध पत्र में हिंदी पाठ के लिए पूर्व संसाधन के प्रत्येक चरण का गहन विश्लेषण कर प्राकृतिक भाषा संसाधन के अनुप्रयोगों के लिए एक मॉडल विकसित करना है। हिंदी पाठ के रूप में विशेष विधि डोमेन का चुनाव कर विश्लेषण कर टूल निर्माण किया जाना है।

पूर्व संसाधन प्रक्रिया के कार्यान्वयन चरण

- **सामान्यीकरण (Normalization)-** चरण में रिक्त स्थान, अन्य विशेष प्रतीकों, पूर्वसर्ग और संक्षेपाक्षर को समाप्त किया जाता है साथ ही वाक्यों को विघटित किया जाता है और फिर प्रत्येक वाक्य में शब्दों की संख्या गिनी जाती है। हिंदी में, वाक्य की सीमा की पहचान करके वाक्य को खंडित किया जाता है जो अंत चिन्हक 'पूर्ण विराम =।' के साथ समाप्त होता है।
- **टोकेनाइजेशन (Tokenization)-** रिक्त स्थान और अन्य विशेष प्रतीकों की पहचान करके वाक्यों को शब्दों में विभाजित करने की प्रक्रिया है। उदाहरणः- 'अभिलेख' 'से'

'यह' 'प्रतीत' 'होता' 'है' 'कि' 'आवेदकगण' 'म' 'प्र' 'उत्पाद' 'अधिनियम' 'की' 'धारा' '34' '2' 'के' 'अधीन' 'दण्डनीय' 'अपराध' 'हेतु' 'न्यायिक' 'मजिस्ट्रेट' 'प्रथम' 'श्रेणी' 'जिला' 'पन्ना' 'के' 'समक्ष' 'लम्बित' 'दाण्डिक' 'वाद' 'संख्या' '268' '2019' 'में' 'विचारण' 'का' 'सामना' 'कर' 'रहे' 'हैं'।

- **विराम शब्द निष्कासन (Stop Word Removal)-** में विराम शब्द सबसे अधिक बार आने वाले शब्द हैं। इस चरण में बिना शब्दार्थ वाले सामान्य शब्द और जो पाठ के बारे में प्रासंगिक जानकारी एकत्र नहीं करते हैं, उन्हें समाप्त कर दिया जाता है। हमने हिंदी के लिए विराम शब्दों की सूची बनाई है जिसमें लगभग '390' विराम शब्द प्राप्त हुए। हिंदी कार्पस का विश्लेषण विधि डोमेन पर किया गया है। हमने मैन्युअल रूप से इन अद्वितीय शब्दों का विश्लेषण किया और विराम शब्दों की पहचान की। कार्पस के विश्लेषण में, हमनें देखा कि विराम शब्द पाठ के '40'% को कवर करते हैं। उदाहरणः- यह, इसके, उन्होंने, अपने, क्या, जो, किसे, किसको, कि, ये, हूँ, में, के, की, कोई, जिसे, यद्यपि, हालांकि, इसलिए, जबकि, केवल, वह, द्वारा, करते, होते इत्यादि।

स्टेमिंग (Stemming)- स्टेमिंग प्रक्रिया में, विभक्त शब्दों से प्रत्यय और उपसर्ग को हटा कर मूल शब्द प्राप्त किए जाते हैं। हमने हिंदी के लिए मूल शब्द, प्रत्यय और उपसर्ग की सूची तैयार कर कार्यान्वयन किया गया है। उदाहरणः-

कलनविधि

चरण 1 – प्रोग्राम शुरू किया गया ।

चरण 2 – टेक्स्ट बॉक्स 1 में हिंदी वाक्यों के संग्रह को इनपुट किया गया।

चरण 3 – इनपुट में दिए गए वाक्य संग्रह पर सामान्यीकरण की प्रक्रिया की जाएगी। फिर आउटपुट के रूप में टेक्स्ट बॉक्स 2 में दिखा दिया जाएगा।

चरण 4 – तत्पश्चात इनपुट में दिए गए वाक्य संग्रह पर टोकेनाइजेशन की प्रक्रिया की जाएगी। फिर आउटपुट के रूप में टेक्स्ट बॉक्स 2 में दिखा दिया जाएगा।

चरण 5 – तत्पश्चात् इनपुट में दिए गए वाक्य संग्रह पर विराम शब्द निष्कासन की प्रक्रिया की जाएगी। फिर आउटपुट के रूप में टेक्स्ट बॉक्स 2 में दिखा दिया जाएगा।

सारणी 3 – हिंदी मूल शब्द, विभक्त शब्द, प्रत्यय एवं उपसर्ग के उदाहरण

क्र. सं.	विभक्त शब्द	मूल शब्द	प्रत्यय/उपसर्ग
1.	लेखा	लेख	आ = 'ा'
2.	योग्यता	योग्य	ता
3.	समाप्ति	समाप्त	इ = 'ि'
4.	आदर्शवाद	आदर्श	वाद
5.	परिस्थितियां	स्थिति	परि, यां

सारणी 1 – सामान्यीकरण एवं टोकेनाइजेशन के परिणाम

क्र. सं.	वाक्य	विराम शब्दों को हटाने से पहले शब्दों की संख्या
01.	अभिलेख से यह प्रतीत होता है कि आवेदकगण मप्र उत्पाद अधिनियम की धारा 34 2 के अधीन दण्डनीय अपराध हेतु न्यायिक मजिस्ट्रेट प्रथम श्रेणी जिला पन्ना के समक्ष लम्बित दाण्डिकवाद संख्या 2682019 में विचारण का सामना कर रहे हैं	42
02.	इस न्यायालय ने अभिलेख तथा दोनों पक्षकारों के विद्वान अधिवक्तागण द्वारा प्रस्तुत तर्कों का परिशीलन किया है	17

सारणी 2 – विराम शब्द निष्कासन के परिणाम

क्र.सं.	वाक्य	विराम शब्दों को हटाने के बाद शब्दों की संख्या
01.	अभिलेख प्रतीत आवेदकगण मप्र उत्पाद अधिनियम धारा 34 2 अधीन दण्डनीय अपराध न्यायिक मजिस्ट्रेट प्रथम श्रेणी जिला पन्ना समक्ष लम्बित दाण्डिकवाद संख्या 2682019 विचारण सामना	28
02.	न्यायालय अभिलेख दोनों पक्षकारों विद्वान अधिवक्तागण प्रस्तुत तर्कों परिशीलन	9

चरण 6 तत्पश्चात् इनपुट में दिए गए वाक्य संग्रह पर स्टेमिंग की प्रक्रिया की जाएगी। इसमें पहले मूल शब्द डेटाबेस में मिल गए तो आउटपुट के रूप में टेक्स्ट बॉक्स 2 में दिखा दिया जाएगा अन्यथा विभक्त शब्दों को डेटाबेस में मिलान करके मूल शब्द और प्रत्यय, उपसर्ग को आउटपुट के रूप में टेक्स्ट बॉक्स 2 में दिखा दिया जाएगा।

चरण 7 – प्रोग्राम बंद किया गया।

फ्लो चार्ट

स्क्रीनशॉट

1) प्रथम स्क्रीनशॉट- सामान्यीकरण

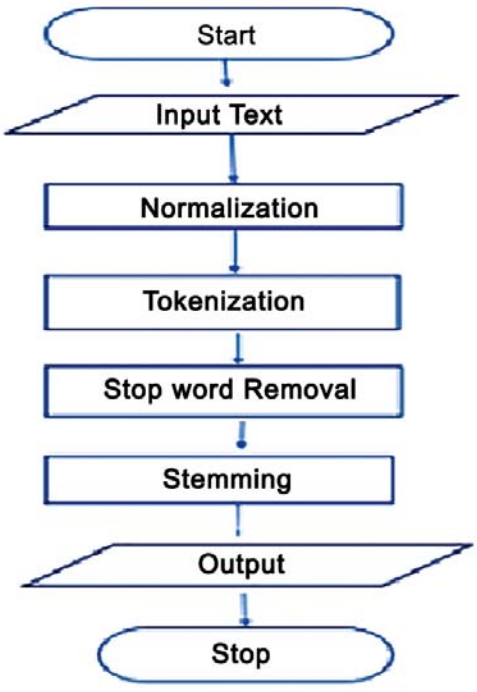
2) द्वितीय स्क्रीनशॉट- टोकेनाइजेसन

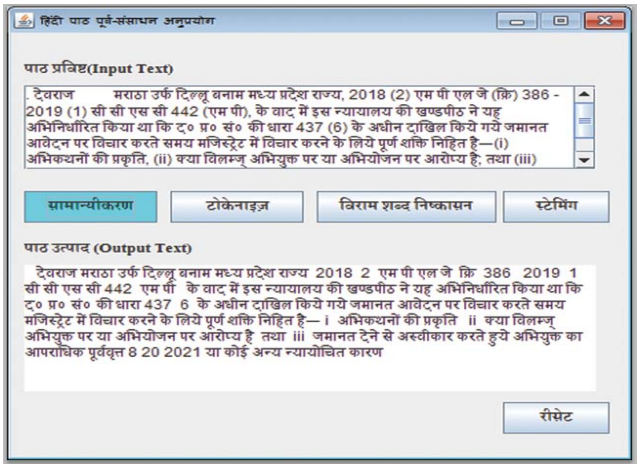
3) तृतीय स्क्रीनशॉट- विराम शब्द निष्कासन

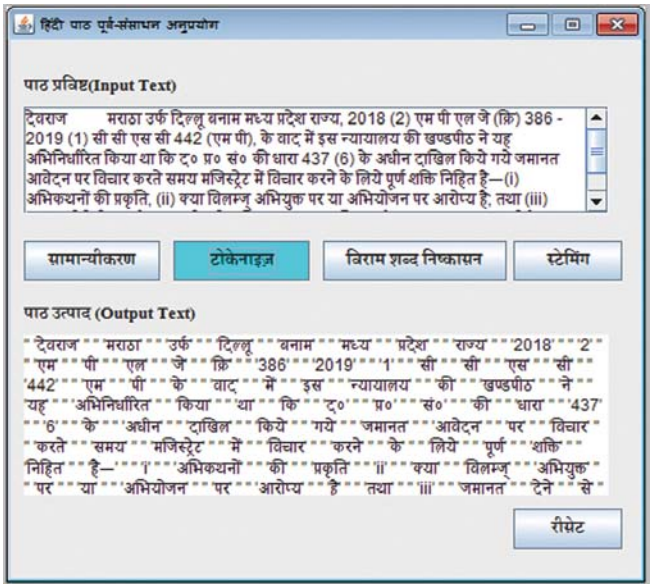
4) चतुर्थ स्क्रीनशॉट - स्टेमिंग

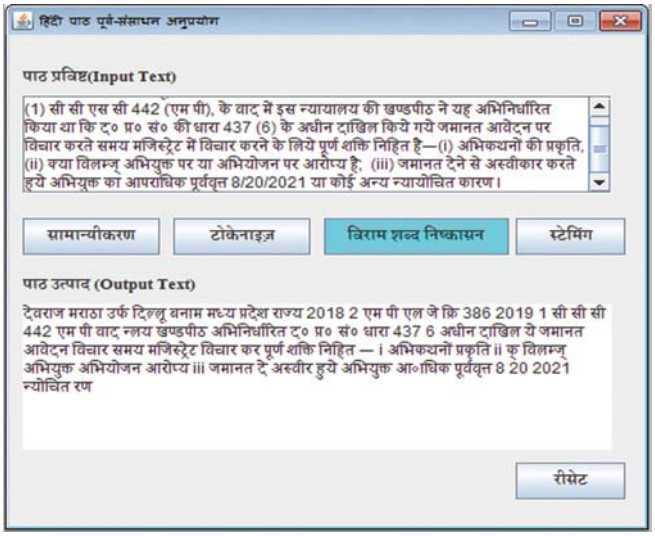
परिणाम एवं विवेचना

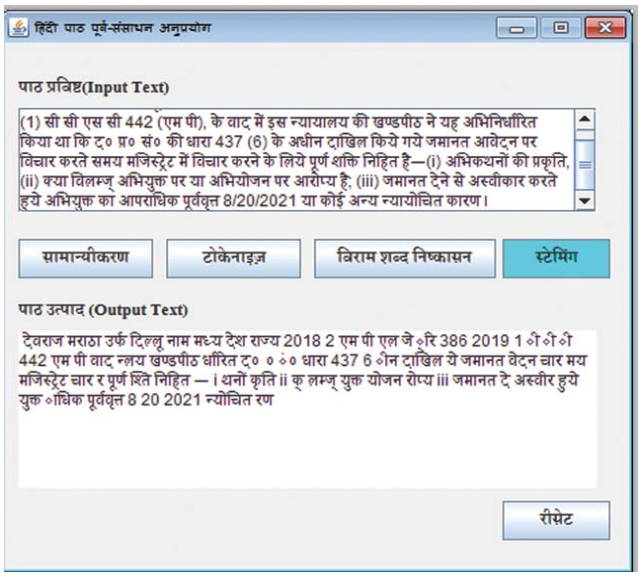
इस शोध पत्र में हिंदी पाठ के लिए पूर्ण पूर्व संसाधन चरण पर चर्चा की है। पूर्व संसाधन में उपयोग किए जाने वाले अधिकांश संसाधन जैसे हिंदी सामान्यीकरण, टोकेनाइजेसन, विराम शब्द निष्कासन एवं हिंदी स्टेमर आदि को डोमेन विशिष्ट अच्छे परिणाम उत्पन्न करने के लिए विकसित किया जाना था। इन संसाधनों को











विकसित करने के लिए मैन्युअल और स्वचालित टूल का उपयोग करके हिंदी कार्पस, हिंदी रूप का गहन विश्लेषण किया जाना था। इस चरण के परिणाम हिंदी के लिए प्राकृतिक भाषा संसाधन (NLP) अनुप्रयोगों को विकसित करने के लिए फायदेमंद हो सकते हैं।

संदर्भ सूची

1. धावले अपूर्वा, कुलकर्णी सोनाली और वैशाली कुंभकर्ण, 'आटोमेटिक प्री-प्रोसेसिंग ऑफ मराठी टेक्स्ट फॉर समराइजेसन', इंटरनेशनल जर्नल ऑफ इंजीनियरिंग एण्ड एडवांस्ड टेक्नोलॉजी (IJEAT) ISSN: 2249-8958 (Online), वॉल्यूम-10 (2020).
2. गुप्ता विशाल और लेहल गुरप्रीत, 'प्री-प्रोसेसिंग फेज़ ऑफ पंजाबी लैंग्वेज टेक्स्ट समराइजेसन', कम्युनिकेशन्स इन कंप्यूटर एण्ड इनफार्मेशन साइंस, ISBN: 978-3-642-19402-3(2011).
3. हॉक मोज़म्मेल और हसन मह्मुदल, 'प्री-प्रोसेसिंग दी बंगाली टेक्स्टः एन एनालिसिस ऑफ एप्रोप्रियेट व 'र्स', बंगाली कंप्यूटिंग (एन.एल.पी.)(2018).
4. लार्डसामी रवि और अब्राहम स्टनिस्लौस, 'ए सर्वे ऑन टेक्स्ट प्री-प्रोसेसिंग टेक्निक्स एण्ड टूल्स', इंटरनेशनल जर्नल ऑफ कंप्यूटर साइंसेज एण्ड इंजीनियरिंग, E-ISSN: 2347-2693 वॉल्यूम-6 (2018).
5. पॉल स्निग्धा, निशीथ जोशी और माथुर इति, 'डेवलपमेंट ऑफ ए हिंदी लेम्माटाईजर', इंटरनेशनल जर्नल ऑफ कम्प्यूटेशनल लिंग्विस्टिक्स एण्ड नेचुरल लैंग्वेज प्रोसेसिंग, वॉल्यूम -2, ISSN 2279 - 0756(2013).
6. पिम्पलशेंदे अन्जुषा और महाजन ए.आर., 'प्री-प्रोसेसिंग फेज़ ऑफ हिंदी लैंग्वेज टेक्स्ट समराईजेसन सिस्टम', इंटरनेशनल जर्नल ऑफ कंप्यूटर साइंस एण्ड इनफार्मेशन सिक्यूरिटी (IJCSIS), वॉल्यूम न. 14, (2016).
7. पाण्डेय बी.पी. ए तामता पवन और धामी एच.एस., 'ए देवनागरी स्क्रिप्ट बेस्ड स्टेमर', इंटरनेशनल जर्नल ऑफ कम्प्यूटेशनल लिंग्विस्टिक्स रिसर्च, वॉल्यूम-5 (2014).
8. रौलजी जयदीप सिंह और सैनी जतिंदर कुमार, 'स्टॉप वर्ड रिमूवल एण्ड इट्स इम्प्लीमेंटेशन फॉर संस्कृत लैंग्वेज', इंटरनेशनल जर्नल ऑफ कंप्यूटर एप्लीकेशन, (0975-8887) वॉल्यूम-150, (2016).
9. राजकुमार एन., सुबाशिनी टी.एस., राजन के. और रामलिंगम वी., 'प्री-प्रोसेसिंग : दी फर्स्ट स्टेप टुवर्ड्स टेक्स्ट डॉक्यूमेंट क्लासिफिकेशन-ए रिव्यु', 'इंटरनेशनल जर्नल ऑफ़ रिसर्च इन एड्वेंट टेक्नोलॉजी', (IJRAT) विशेष अंक- E-ISSN: 2321-9637(2018).
10. सुन फेन, बेलात्रेचे अम्मार, 'प्री-प्रोसेसिंग ऑनलाइन फाइनेंसियल टेक्स्ट फॉर सेंटिमेंट क्लासिफिकेशन : ए नेचुरल लैंग्वेज प्रोसेसिंग एप्रोच', आई. ई. ई. ई. कम्प्यूटेशनल इंटेलिजेंस फॉर फाइनेंसियल इंजीनियरिंग एण्ड इकोनॉमिक्स (2014).
11. सिंह सोनित, 'नेचुरल लैंग्वेज प्रोसेसिंग फॉर इनफार्मेशन एक्सट्रैक्शन' arXiv:1807.02383v1 [cs.CL](2018).
12. तिकार्य आशीष बी., कोठारी मयूर और पटेल पिंकेह एच. , 'प्री-प्रोसेसिंग फेज़ ऑफ टेक्स्टसमराइजेसन बेस्ड ऑन गुजराती लैंग्वेज', इंटरनेशनल जर्नल ऑफ इनोवेशन रिसर्च इन कंप्यूटर साइंस एण्ड टेक्नोलॉजी (IJIRCST) ISSN: 2347-5552, वॉल्यूम.2 (2014).
13. तबस्सुम आईशा और पाटिल डॉ. राजेंद्र, 'ए सर्वे ऑन टेक्स्ट प्री-प्रोसेसिंग एण्ड फीचर एक्सट्रैक्शन टेक्निक्स इन नेचुरल लैंग्वेज प्रोसेसिंग', इंटरनेशनल रिसर्च जर्नल ऑफ इंजीनियरिंग एण्ड टेक्नोलॉजी (IRJET) e-ISSN: 2395-0056(2020).