**Text Pre-processing and Parsing for Information Retrieval Applications**

**Abstract:**

This study describes a text parsing and pre-processing system intended for use in information retrieval applications. A tokenizer that handles numerals, punctuation marks, and letter case is included, as well as a stop word eliminator and a stemming algorithm. The token stream that results is used to construct a term dictionary that maps unique terms to unique numerical IDs, as well as a file dictionary that maps document names to numerical IDs. Using the TREC data set, the system was proven to be successful in parsing and pre-processing texts for information retrieval applications.

**Introduction:**

In several disciplines, including academia, business, and government, the capacity to efficiently retrieve useful information from massive document collections is crucial. Text parsing and pre-processing are required operations that assist to clean and convert unprocessed text input into a format that information retrieval systems can utilize. In this work, we offer a system for text parsing and pre-processing that is intended for use in information retrieval applications.

**Methods:**

Our system contains a tokenizer that applies fundamental tokenization principles to divide words, eliminates numbers, disregards words containing numbers, splits on non-alphanumeric letters, and transforms all words to lowercase. This tokenizer is followed by a stop word remover, which eliminates frequent words that do not significantly contribute to the sense of the text, and a stemming algorithm, which reduces words to their root form. The resultant token stream is used to construct a term dictionary, which maps unique terms to unique number IDs, and a file dictionary, which maps document names to numerical IDs.

**Results:**

For the purpose of testing our system, we used the TREC data set, which comprises several documents in a single file with distinct tags. We used our system to parse the documents and then printed the document IDs and token streams to ensure that they were successfully processed. We also built a dictionary file including the words dictionary and dictionary file. It has been determined that our system is capable of parsing and preparing texts for information retrieval applications.

**Conclusion:**

In conclusion, our text parsing and pre-processing system is a great resource for applications that do information retrieval. By properly cleansing and converting unprocessed text data, our method allows more precise and efficient retrieval of pertinent information from massive document collections. Future studies may include greater system optimization and testing with bigger data sets.