## Overview

This document outlines the development and functionality of the text processing program designed for efficient keyword-based searches within a document. This is program that filters out noise words, builds a search index using a trie data structure, and enables various search queries.

## Data Structures

The central data structure is the **TrieNode**, which consists of an array representing children nodes, a boolean to denote the end of a word, and a set to store paragraph numbers. The trie allows for efficient insertion and search operations for words within the text.  In addition to the trie, a HashSet is used for storing noise words, and an ArrayList holds the processed paragraphs.

## Text Parsing and Noise Word Elimination

The program reads the documents line-by-line, identifying and omitting noise words using a predefined list. It splits paragraphs into words, converting them to lowercase and stripping non-alphanumeric characters, before inserting them into the trie. Paragraphs are indexed numerically as they are processed.

## Search Functionality

The program supports five types of search queries (which must me uncommented for testing purposes):

1. **Single Keyword Search:** Finds paragraphs containing a single keyword.
2. **Conjunctive Search:** Retrieves paragraphs containing all specified keywords.
3. **Disjunctive Search:** Identifies paragraphs containing at least one of the specified keywords.
4. **Conjunctive with 'Or':** Searches for paragraphs containing the first keyword and any one of the subsequent keywords.
5. **Disjunctive with 'And':** Finds paragraphs containing the first keyword or all of the subsequent keywords together.

## Memory Management

All data, including the trie and sets for paragraph indexing, are stored in memory. This approach ensures quick access and manipulation of data during search operations.

## Input/Output Specification

The program accepts text files for the main document and noise words list. The output consists of paragraph numbers and the corresponding text that match the search criteria.

Query 1 (Single Keyword: urban)

Paragraph 4:

We are currently experiencing an extraordinary acceleration in the growth rate of digital data One of the reasons for this increase is the digitization of virtually all communications and records This exponential growth is evidenced by the fact that the sum total of data produced in the last year or two would exceed all data that existed in digital form prior to that time Just as the Industrial Revolution created an entirely new urban way of life for what is now more than half of the population of the planet the Big Data revolution will dramatically alter the ways in which we interact not only with each other but with all of the institutions which mediate our lives Our age has come to be known as the era of Big Data and we are only beginning to understand its opportunities and challenges This position paper underlines some of these emerging promises and threats as well as their implications for researchers and scientists across the database community

Query 2 (Conjunctive Search: urban, data)

Paragraph 4:

We are currently experiencing an extraordinary acceleration in the growth rate of digital data One of the reasons for this increase is the digitization of virtually all communications and records This exponential growth is evidenced by the fact that the sum total of data produced in the last year or two would exceed all data that existed in digital form prior to that time Just as the Industrial Revolution created an entirely new urban way of life for what is now more than half of the population of the planet the Big Data revolution will dramatically alter the ways in which we interact not only with each other but with all of the institutions which mediate our lives Our age has come to be known as the era of Big Data and we are only beginning to understand its opportunities and challenges This position paper underlines some of these emerging promises and threats as well as their implications for researchers and scientists across the database community

Query 3 (Disjunctive Search: growth, rate)

Paragraph 4:

We are currently experiencing an extraordinary acceleration in the growth rate of digital data One of the reasons for this increase is the digitization of virtually all communications and records This exponential growth is evidenced by the fact that the sum total of data produced in the last year or two would exceed all data that existed in digital form prior to that time Just as

the Industrial Revolution created an entirely new urban way of life for what is now more than half of the population of the planet the Big Data revolution will dramatically alter the ways in which we interact not only with each other but with all of the institutions which mediate our lives Our age has come to be known as the era of Big Data and we are only beginning to understand its opportunities and challenges This position paper underlines some of these emerging promises and threats as well as their implications for researchers and scientists across the database community

Paragraph 69:

A British Columbia BC woman in Vancouver is instrumental in setting in motion a class action suit 51 against an OSN which used her name and or image without permission in a sponsored ads to her on line friends The judge in allowing the suit to go ahead in BC ruled as follows with the creation and growth of the internet the potential implications for a loss of privacy are greater than ever The difficulty in proving quantifiable damage remains great for an individual whose privacy is lost but the social harm can be monumental if the loss of privacy includes publicity over the internet with its almost infinite reach and timelessness I conclude that the legislative conferral of exclusive jurisdiction on this Court for claims under the Privacy Act evidences both a legislative intention to override any forum selection clause to the contrary and a strong public policy reason for not enforcing the Forum Selection Clause I conclude that the plaintiff has shown strong cause why the Forum Selection Clause should not cause this Court to decline jurisdiction

Query 4 (Conjunctive with 'Or': bullying, growth, harassment)

Paragraph 43:

Many cases of cyberbullying and harassment have been reported in the popular press While bullying and harassment obviously pre existed the web and OSNs information technologies such as text messaging recording and instant uploading features always readily accessible by smartphones have provided easy to use and easy to exploit media As with older forms of bullying and harassment anyone could be a victim of cyberbullying however children and youth are the most common perpetrators and targets 29

Paragraph 49:

In Canada the current government has proposed an omnibus criminal bill C 13 popularly called the Cyber bullying Bill 36 It includes provisions for monitoring and recording internet traffic by ISPs and web sites While it is evident the society and its law enforcement machine needs tools to fight online crime and cyber bullying and harassment this must be done without trampling on people s fundamental rights to privacy C 13 proposes new investigative powers and its use by a

large number of public officials is not subject to any requirements of accountability or reporting mechanism From a privacy standpoint the bill is further appalling since it gives the ISP and telecoms full immunity to disclose their customers information without a warrant

Query 5 (Disjunctive with 'And': proliferation, urban, way)

Paragraph 4:

We are currently experiencing an extraordinary acceleration in the growth rate of digital data One of the reasons for this increase is the digitization of virtually all communications and records This exponential growth is evidenced by the fact that the sum total of data produced in the last year or two would exceed all data that existed in digital form prior to that time Just as the Industrial Revolution created an entirely new urban way of life for what is now more than half of the population of the planet the Big Data revolution will dramatically alter the ways in which we interact not only with each other but with all of the institutions which mediate our lives Our age has come to be known as the era of Big Data and we are only beginning to understand its opportunities and challenges This position paper underlines some of these emerging promises and threats as well as their implications for researchers and scientists across the database community

Paragraph 8:

Therefore in order to understand this new age in terms of some of its origins it is useful to briefly recall the history of data storage From here on this paper will examine more closely some specific consequences of this new age which bear most directly on the theory and practice of the data engineer namely the proliferation of data and the novel possibilities it affords

Example of unsuccessful query:

Query 5 (Disjunctive with 'And': data, urban, hazard)

No results found.

## Generalization Capability

By design, the program can work with any text file, making it a versatile tool for different document searches.

## Error Handling

The program includes try-catch blocks to handle I/O exceptions and checks for the validity of paragraph numbers during retrieval, ensuring robustness against file reading errors and invalid references.

## Testing

The program's functionality was verified against a variety of text files, including "The State of Data Final.txt" and "Privacy in the Age of Information.txt," to ensure adequate performance across both expected and unforeseen scenarios. Notably, for the execution of test cases, specific search queries within the code must be uncommented. This allows for selective testing and a focused examination of each query's operational integrity.

Test results, exemplified in the report, showcase the program's successful retrieval of paragraph numbers and corresponding text for each type of search query, as well as instances where no results are found, thereby validating the program's comprehensive error handling and search capabilities.

## Assumptions

The current implementation assumes that the document is in English and contains only alphanumeric characters. Words are considered in their lowercase form, and noise words are defined in a separate file.

## Conclusion

The developed program successfully meets the objectives of the assignment, providing a reliable and efficient search mechanism for large text files. Future improvements could include support for Unicode characters, phrase searching, and performance optimization for even larger datasets.