Lab report: Content selection for single document summarization

|  |
| --- |
| Kelli Wiseth  San José State University  Linguistics 165 Coursework  Spring 2013  kelli.wiseth@gmail.com |

|  |  |
| --- | --- |
|  |  |

Abstract

This lab assignment uses two different approaches—(1) term-frequency and inverse document frequency (tf-idf); and (2) log-likelihood ratio (llr)—to identify the most informative words in a given document. The approaches have been implemented in Python code that processes a text

The assignment rests on an understanding of ‘informative’ that uses stats

concepts of document processing is selecting keywords from the document. This exercise and sample code steps-through a Python script that processes a text-file version of a roughly 2,000 word article from the New Yorker magazine and identifies the ‘important’ words using two different approaches:

1. Introduction

[introduction stub]

1. Section

[section 2 stub]

[section 2 stub]

* 1. Subsection

[subsection 2.1 stub]

1. Conclusion

[conclusion stub]

Reference

[reference stub]

[reference stub]