# CALIFORNIA STATE SCIENCE FAIR
## 2013 PROJECT SUMMARY

| | |
|---|---|
| **Name(s)**<br>**Tanay Tandon** | **Project Number**<br>**S1426** |

**Project Title**

## Clipped: Automated Text Summarization through Semantic Natural Language Processing and Clustered Machine Learning

### Abstract

**Objectives/Goals**

Individuals and companies in today#s world are inundated with a flood of data. The issue of information overload has made it exceedingly difficult to function efficiently as part of modern day society. Between web pages, social networks, news articles, and email; everyone parses through hundreds if not thousands of documents on a daily basis. People need a way to quickly determine what these documents mean, and extract the most important information from them. The purpose of this research is to develop a text summarization algorithm that makes use of novel grammatical models and linguistics to statistically extract the most relevant information from a text sample. The algorithm generates a bullet point summary of the text by discovering information concentration through subject-predicate relationships and keyword ranking based structures; and makes the process of reading the content more concise and efficient.

**Methods/Materials**

All development occurred in the server side language of PHP, and the entire algorithm was written independently. The program made use of the lexicon database known as the Brown Corpus to tag sentences through POS (Parts of Speech). The algorithm code was tested and deployed on a server with a user-testing beta to ensure scalability. The algorithm was developed in 3 major stages # the ICF (Initial Content Filtration), GPM (Grammatical Pattern Matching), and the CCR (Contextual Content Review). The algorithm was trained to identify 23 patterns of information concentration, involving subject-predicate positioning and keyphrase concentrations.

**Results**

The results of the study indicate that the algorithm holds high precision in summarization of content. The ROUGE metric generated an average F1 score of .5868, and an average precision score of .640. The ROUGE metric evaluates a summary through a similarity index run between a gold summary and the test summary. The summaries generated are concise and comprehensive, and the algorithm was released as a beta product on the iOS and Android Market # reaching the top of the app store within 2 months.

**Conclusions/Discussion**

The research indicates that the Clipped algorithm has the potential to change the way we consume content. The novel grammatical approach identifies information concentration more accurately, and the bullet-point summaries make the process of reading more efficient than ever before.

**Summary Statement**

I developed a text summarization algorithm that analyzes information grammatically, and generates a bullet-point summary of the text sample; making the process of consuming information efficient.

**Help Received**

Parents helped make display board. Gary Griffiths of Trapit provided advice on application