DYNAMICALLY GENERATING STOPWORD LISTS FROM SOFTWARE ENGINEERING ARTIFACTS

Shimon Johnson, Dr. Christian Newman, 2020.

INTRODUCTION & MOTIVATION

- Working with text data is challenging, Machine learning approaches like Text classification often require Preprocessed data.
- Preprocessing helps clean and prepare data before predictive models are developed.
- Predictions from Incorrect data models lead to inaccurate & misleading results, effectively impacting system performance & reliability.
- Most natural language Text data contains Fluff, Stopwords.
- Stopwords are words that are commonly used within text to improve its richness and basically add no semantic value.
- Dynamically generating Stopword lists could significantly improve our understanding of how words impact machine learning approaches by allowing us to study how different heuristics for determining Stopwords improve or degrade the results of ML & other data processing approaches. This will allow us to optimize data preprocessing for these approaches.

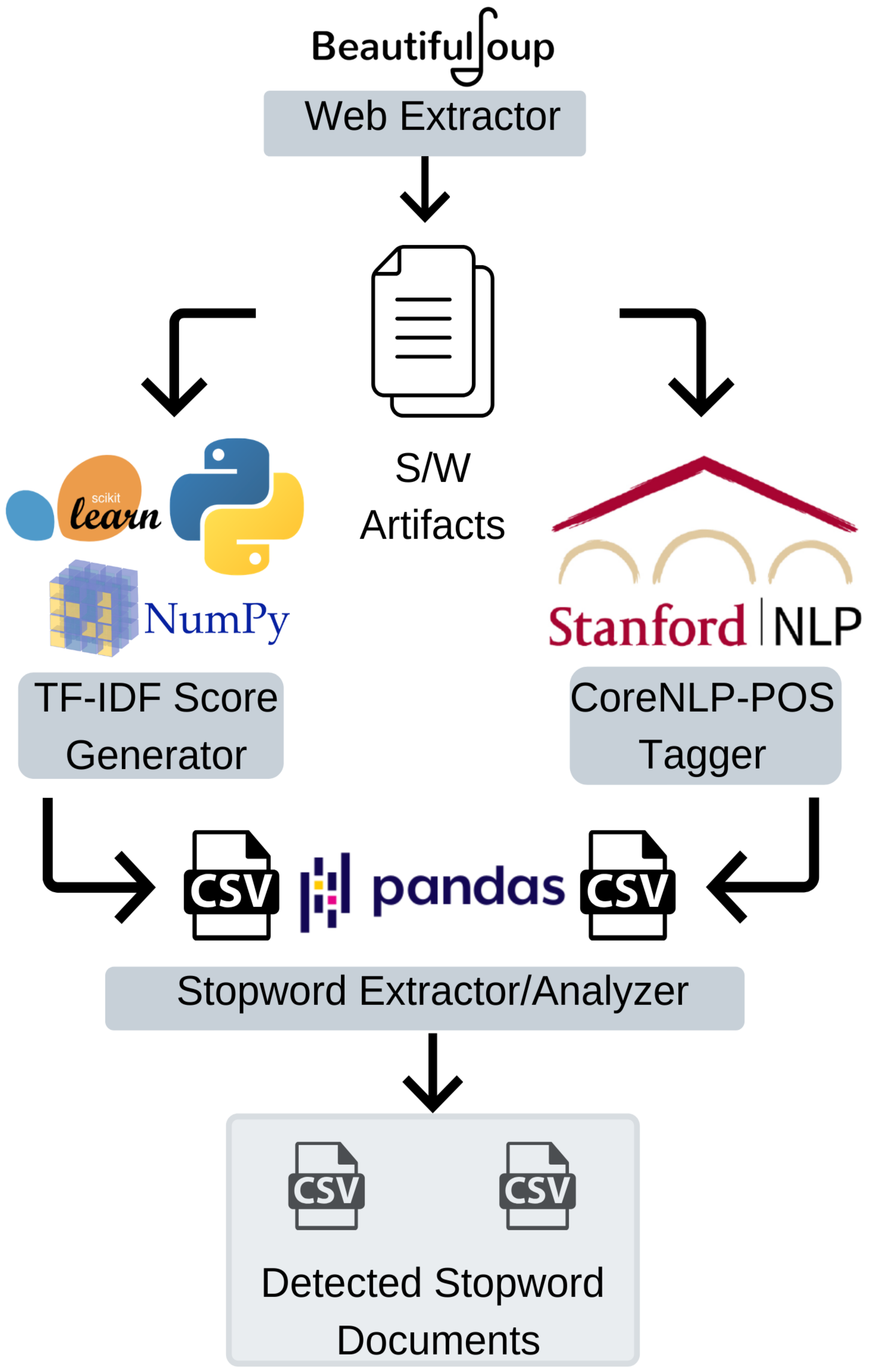
RESEARCH METHODOLOGY

- Synthesizing TF-IDF and POS Tagging into a Hybrid Approach.
- **TF-IDF-** is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.
- **POS Tagging-** is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech based on both its definition and its context
- Performed on S/W Artifacts (API Documentation produced by Javadoc)

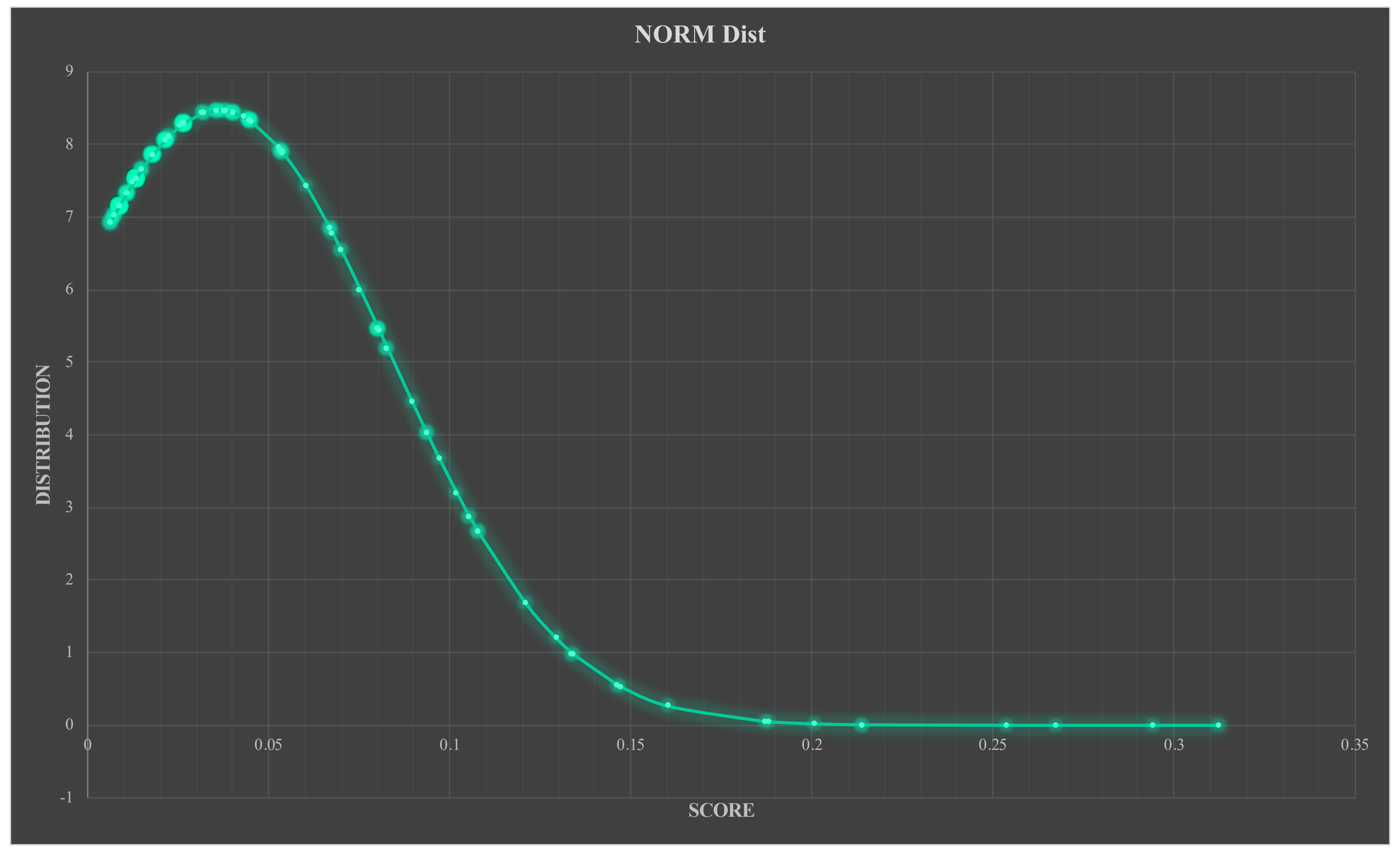
RESEARCH APPLICATIONS

- Information Retrieval
- Text Classification
- Spam Filtering
- Caption Generation

RESEARCH IMPLEMENTATION & TECH STACK



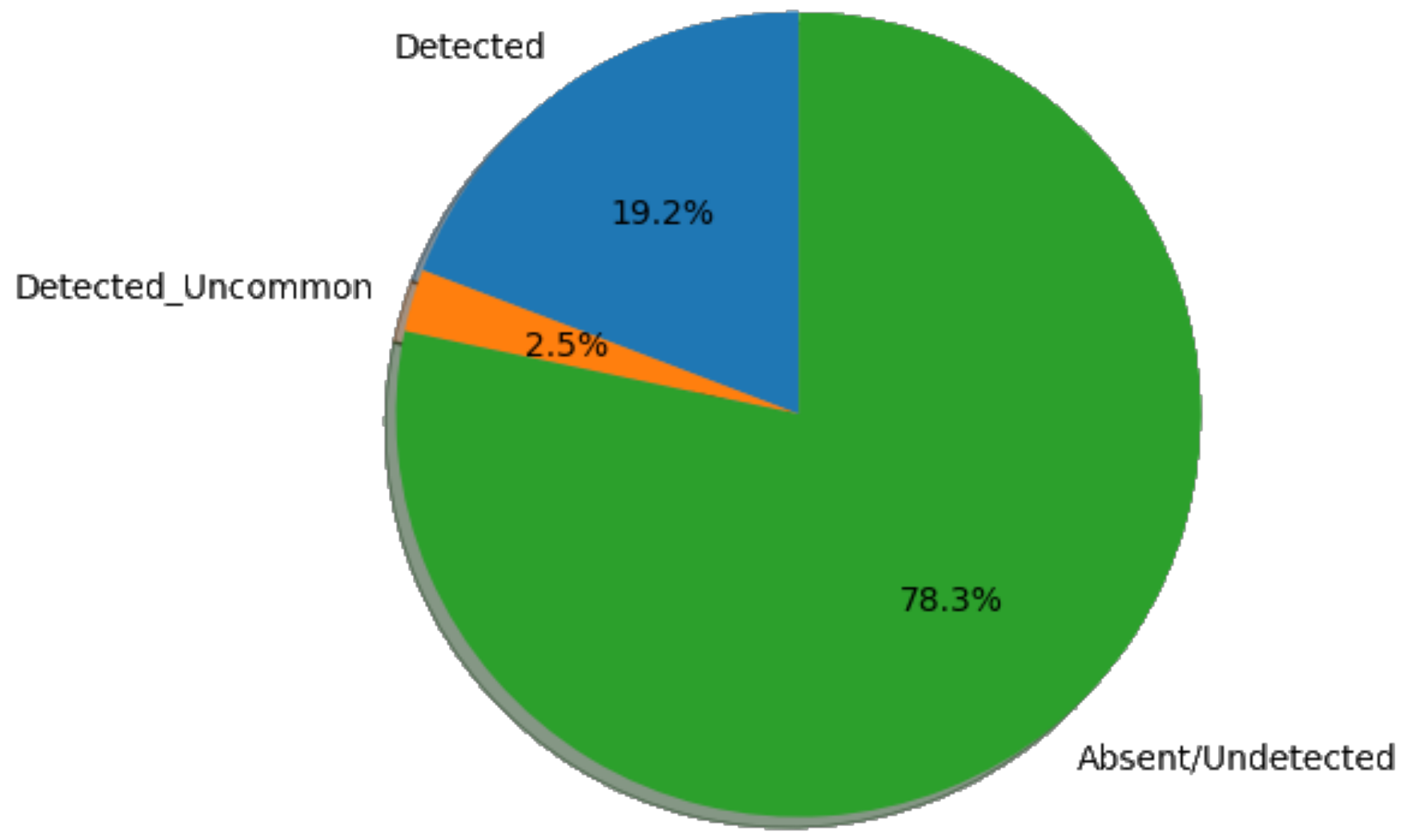
EVALUATION



NORMAL DISTRIBUTION PLOT FOR DOCUMENT 1

- Corpus consisted of five software engineering artifacts, which were Java API documentation obtained from Oracle's Website.
- Files obtained from score generator & POS tagger were analyzed together.
- A Normal Distribution was created for each file and then visualized.
- POS Tags such as '**IN**', '**CC**', '**DT**', '**PRP**', '**WDT**', '**PRP$**', '**WP**', '**WP$**', '**RP**', '**TO**', '**PDT**', '**WRB**' and '**CD**' were considered as Stopwords.
- The +1SD was chosen as the Threshold below which all words occurring were checked for their POS tags before being marked as Stopwords

RESULTS



- Obtained list was compared with Scikit's Stopword list.
- 78.3 % of the words present in Scikit's list were Absent in the corpus.