Compute all the words by use of natural language processing of full text articles

Chang Che Wei
N26041678

4A02C014@stust.edu.tw

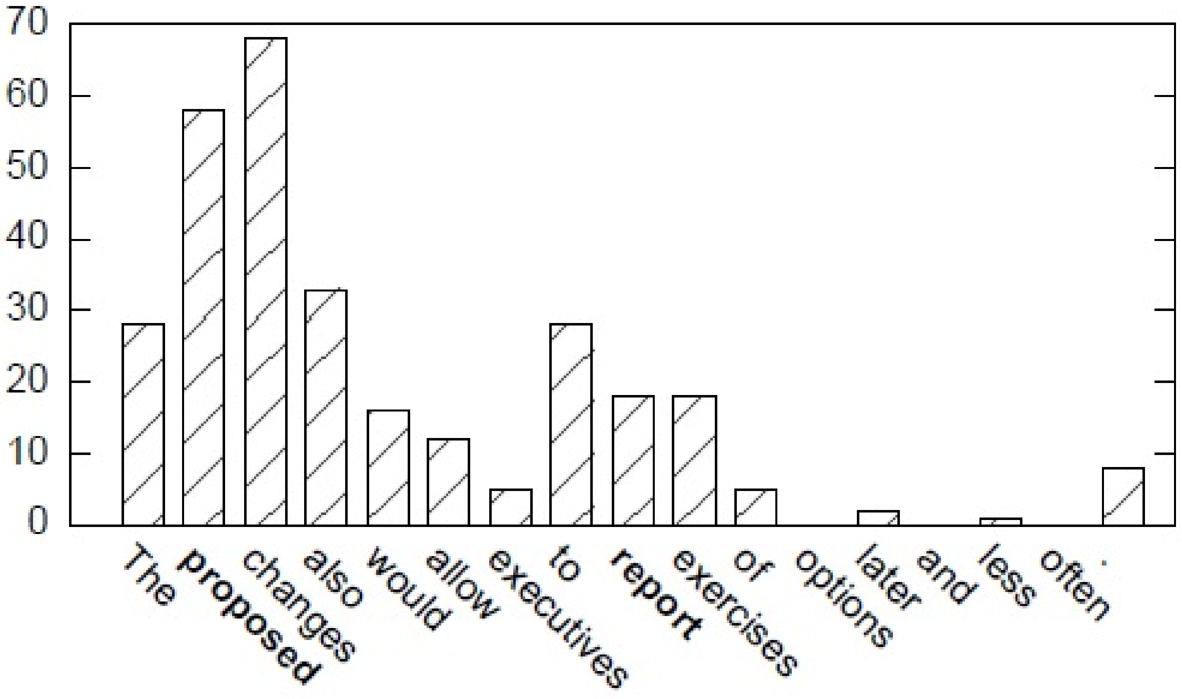
March 23, 2016

Reason

In order to highlight the focus of the article ,the special and important words in the article need to be found mostly. The article will promote these words to make the article's mainpoint prominent. Compute all the words in the article and show the number of occurrences of most rankings. By these words ,the analyzer know which words in the article should be analyzed.

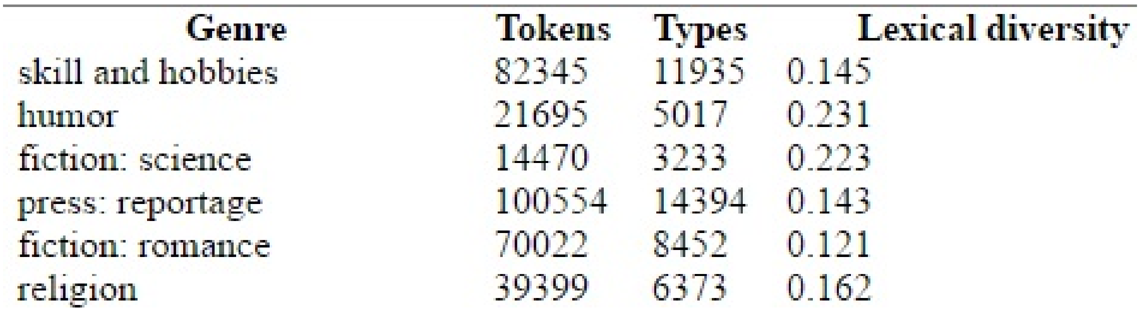
Operate Process

1. distinguish : distinguish all the words in the article.



2. classification : Divided into four categories by word count.
a. Word Count ≥ 15
b.$14 \geq WordCount \geq 10$
c. $9 \geq WordCount \geq 5$
d. $4 \geq WordCount$

3.$search$: $SearchWordrepetitionrateinthearticle.$



Genre	Tokens	Types	Lexical diversity
skill and hobbies	82345	11935	0.145
humor	21695	5017	0.231
fiction: science	14470	3233	0.223
press: reportage	100554	14394	0.143
fiction: romance	70022	8452	0.121
religion	39399	6373	0.162

4.$statistics$: $counttherepetitionrateofwordsinthearticle.$

5.$sort$: $Showthenumberofoccurrencesofmostrankingsineverycategorie.$

Reference

[1]. Collins2011 Natural Language Processing Machine Learning
[2]. Weiscbedel2006 White Paper on Natural Language Processing