

Extracting keyphrases from a text is more complex task than [extracting keywords](index.php?option=com_content&view=article&id=50:getting-keywords-of-the-text&catid=39:text-mining&Itemid=57) from a text. But, usually, keyphrases are more useful in text mining tasks than just keywords.

I have created Perl script that extracts keyphrases from the text.

Phrase is any sequence of 2 or more words that is repeated in text 2 or more times.

The script extracts words' sequences from the text that are most repeated. Then find phrases that are similar to another found phrases and remove them. Last step is to remove some common phrases that are used often but are not useful (like "he said" or "on the").

I have used data from Google News for testing.

It is possible see [how Perl scripts extracts keyphrases from the text](cgi-bin/science/textmining/3/keyphrases.cgi). There is list of latest news from Google News service. Click link against news message and see keyphrases. Then open message and see if keyphrases are correct.